Department of Mathematics
University of Toledo

Master of Science Degree
Comprehensive Examination
Applied Statistics

May 18, 1996

Instructions:

Do problems 1-4. These are 35 points each. These are *required.*
Then choose *two* from 5, 6, and 7. These are 20 points each.
The total is 180, equal to the number of minutes for the exam.
Show all of your computations in your Blue Book.
Prove all of your assertions or quote appropriate theorems.
Books, notes, and calculators *may be used.*
This is a three hour test.

1. Below are printed data and statistics for a study on two different systems for measuring acidity. The MINITAB program was used for the display and analyses. Thirty experimenters used the two systems on samples from the same solution. Below we give the measured concentrations and the difference for each experimenter. The overall goal is to determine if the two systems possess a systematic difference in the reported concentrations.

a. Based on a casual observation of the results shown below, state your opinion as to whether or not a systematic difference exists.

b. Do an appropriate parametric test on this data to answer the question above. Use α = .10. State your assumptions.

c. Do an appropriate nonparametric test on this data to answer the question above. Use α = .10. State your assumptions.

d. Without doing any long calculations, state which test(s) you would tell the researcher were more reliable in this situation (if any). Give your reasoning.

e. For this situation, discuss the advantages and disadvantages of this experimental design. If you had been asked to propose a design, what are the important differences between your design and this one?

```
MTB > Sort 'ACID1'-'DIFF' 'ACID1'-'DIFF';
SUBC>   By 'DIFF'.
MTB > Print 'ACID1'-'DIFF'.

 ROW    ACID1    ACID2     DIFF

   1    0.106    0.111   -0.005
   2    0.104    0.109   -0.005
   3    0.109    0.113   -0.004
   4    0.109    0.111   -0.002
   5    0.110    0.112   -0.002
   6    0.110    0.112   -0.002
   7    0.107    0.109   -0.002
   8    0.109    0.110   -0.001
   9    0.109    0.110   -0.001
  10    0.110    0.111   -0.001
  11    0.111    0.112   -0.001
  12    0.110    0.111   -0.001
  13    0.110    0.110    0.000
  14    0.110    0.110    0.000
  15    0.110    0.110    0.000
  16    0.110    0.110    0.000
  17    0.110    0.109    0.001
  18    0.110    0.109    0.001
  19    0.110    0.109    0.001
  20    0.111    0.110    0.001
  21    0.111    0.110    0.001
  22    0.113    0.109    0.004
  23    0.109    0.105    0.004
  24    0.115    0.110    0.005
  25    0.114    0.109    0.005
  26    0.112    0.105    0.007
  27    0.123    0.109    0.014
  28    0.101    0.011    0.090
  29    0.110    0.010    0.100
  30    0.108    0.003    0.105

MTB > Stem-and-Leaf 'ACID1'-'DIFF'.

Stem-and-leaf of ACID1      N  = 30
Leaf Unit = 0.0010

    1    10 1
    1    10
    2    10 4
    4    10 67
   10    10 899999
  (15)   11 000000000000111
    5    11 23
    3    11 45
    1    11
    1    11
    1    12
    1    12 3

Stem-and-leaf of ACID2      N  = 30
Leaf Unit = 0.0010

    1     0 3
    3     1 01
    3     2
    3     3
    3     4
    3     5
    3     6
    3     7
    3     8
    3     9
   13    10 5599999999
  (17)   11 00000000011112223

Stem-and-leaf of DIFF       N  = 30
Leaf Unit = 0.0010

   14    -0 55422211111100
  (12)    0 001111144557
    4     1 4
    3     2
    3     3
    3     4
    3     5
    3     6
    3     7
    3     8
    3     9 0
    2    10 05

MTB > Describe 'ACID1'-'DIFF'.

                 N      MEAN    MEDIAN    TRMEAN     STDEV
ACID1           30   0.11003   0.11000   0.10992   0.00361
ACID2           30   0.09963   0.11000   0.10581   0.03113
DIFF            30   0.01040   0.00000   0.00450   0.03011

               MIN       MAX        Q1        Q3
ACID1      0.10100   0.12300   0.10900   0.11100
ACID2      0.00300   0.11300   0.10900   0.11100
DIFF      -0.00500   0.10500  -0.00125   0.00425
```

2. Recovery Times.

Three hospitals participated in an experiment to study the effect of full-time in-room assistance on recovery time after a certain surgery. Each hospital selected two nurses. Each nurse assisted three patients; one patient for one hour, one patient for two hours and one patient for four hours. Thus, there were 18 patients in all. The time from the end of surgery until the attainment of a certain recovery objective was recorded for each patient.

Let Y_{ijk} be the recovery time for the patient from the i^{th} hospital ($i = 1, 2, 3$) with the j^{th} nurse ($j = 1, 2$) receiving the k^{th} treatment ($k = 1, 2, 3$, for the one, two and four hour treatments). If needed, let T_k be the treatment times, i.e., $T_1 = 1, T_2 = 2$, and $T_3 = 4$.

For each of the situations A, B, and C below, indicate a model for Y_{ijk}. Define your parameters and include any relevant distributional assumptions.

A. Write down the model for the simple linear regression of recovery time on treatment time (considered as a quantitative variable).

B. Write down the model for the two-way analysis of variance model (no interactions) for the recovery time which considers the hospitals to be blocks and the treatment time to be a qualitative (unordered) factor.

C. Write down the model for recovery time which treats hospitals as fixed effects, nurses as fixed effects nested within hospitals, and treatment as a quantitative covariate. How would the model need to be modified in order to have the nurses be a random effect?

D. The model $Y_{ijk} = \mu + \alpha_{ij} + \gamma_{ij} T_k + \epsilon_{ijk}$ with $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ is the analysis of covariance model for a combined hospital-nurse effect (α_{ij}) and allowing the slope (γ_{ij}) for the length of treatment covariate to vary depending on the hospital-nurse combination. Write down the design matrix X for the linear model $Y = X\beta$ in this case.

3. Twenty students were tracked through their Ohio proficiency exams in Math and Reading in 4th, 7th and 10th grades. The data and some statistics (from MINITAB) are given below. The basic question we want answered is whether the math scores or the reading scores change over time, and, if they do, which ones change and how. With this goal in mind, do the following:

a. Give the complete formula for the test statistic and the critical region for testing the null hypothesis that the mean math score is constant and the mean reading score is constant versus the alternative that at least one of the two varies over the given time period. Use $\alpha = .10$. Do not attempt to evaluate the test statistic.

b. There are six logical linear combinations of the mean vector which one would want simultaneous confidence intervals for if the null hypothesis is rejected. For example, one would compare the 4th and 7th grade math scores. Do the following:

* Identify the six linear combinations.

* Then, for 90% simultaneous intervals, find out whether intervals of the Scheffe or Bonferroni type are shorter.

* Finally, evaluate the shorter interval for the 4th and 7th grade math score comparison.

```
MTB > Print 'Math 4'-'Read 10'.

 ROW  Math 4  Read 4  Math 7  Read 7  Math 10  Read 10

   1      83      90      71      90       65       97
   2      66      54      71      67       71       68
   3      51      56      39      61       33       70
   4      98     100      99      98       89      100
   5      76      56      64      67       49       68
   6      78      75      80      84       74       90
   7      73      90      69      95       69       99
   8      88      70      85      64       76       71
   9      70      60      66      72       68       69
  10      52      45      39      43       29       45
  11      91      61      86      59       90       57
  12      89      92      92      89       91       86
  13      91      71      79      79       76       83
  14      98      90      95      83       95       81
  15      97      96     100      89      100       87
  16      55      78      59      75       61       72
  17      67      69      63      76       60       77
  18      92      80      80      73       76       75
  19      63      63      65      63       57       72
  20      77      95      64      97       62       95

MTB > Describe 'Math 4'-'Read 10'.

                N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
Math 4         20    77.75    77.50    78.11    15.32     3.43
Read 4         20    74.55    73.00    74.78    16.49     3.69
Math 7         20    73.30    71.00    73.72    17.17     3.84
Read 7         20    76.20    75.50    76.83    14.59     3.26
Math 10        20    69.55    70.00    70.11    18.89     4.22
Read 10        20    78.10    76.00    78.72    14.33     3.21

              MIN      MAX       Q1       Q3
Math 4      51.00    98.00    66.25    91.00
Read 4      45.00   100.00    60.25    90.00
Math 7      39.00   100.00    64.00    85.75
Read 7      43.00    98.00    64.75    89.00
Math 10     29.00   100.00    60.25    85.75
Read 10     45.00   100.00    69.25    89.25

MTB > Correlation 'Math 4'-'Read 10'.

          Math 4   Read 4   Math 7   Read 7  Math 10
Read 4     0.631
Math 7     0.924    0.645
Read 7     0.523    0.916    0.545
Math 10    0.868    0.644    0.967    0.553
Read 10    0.464    0.867    0.471    0.961    0.444

NTB > Covariance 'Math 4'-'Read 10'.

            Math 4   Read 4   Math 7   Read 7  Math 10  Read 10
Math 4     234.618
Read 4     159.461  271.839
Math 7     243.132  182.616  294.853
Read 7     116.842  220.253  136.568  212.905
Math 10    251.145  200.471  313.826  152.358  356.997
Read 10    101.974  204.837  115.916  200.926  120.363  205.463
```

4. Rice Yields.

The data below is from a randomized block design. Seven spacings were compared in three randomized blocks. The spacings were the spacing between rice plants. The interest was in the effect of spacing on rice yields.

	Block		
Spacing	I	II	III
30cm x 30cm	5.95	5.30	6.50
30cm x 24cm	7.10	6.45	6.60
30cm x 20cm	7.00	6.50	6.35
30cm x 15cm	8.10	5.50	6.60
24cm x 24cm	8.85	7.65	7.00
24cm x 20cm	7.65	6.90	8.25
24cm x 15cm	7.80	6.75	8.20

A. Produce a row PLUS column fit. Describe briefly your interpretation of the results.

B. Produce a stem-and-leaf plot of the residuals from the row PLUS column fit above. (If you do not have residuals from part *A* above, then produce a stem-and-leaf plot of the raw data instead.)

C. Fill in the blanks in the analysis of variance table below. There are 3 degrees of freedom entries missing, 1 sum of squares entry missing, 1 mean square, 1 F value missing, and 1 p-value missing.

```
> summary(aov(Yield~Block+Spacing,rice))
           Df Sum of Sq  Mean Sq  F Value      Pr(F)
Block            3.965 1.982500
Spacing          8.600 1.433333 3.887006 0.02170428
Residuals
```

5. Statistical Computing and Software.

A. Below is C code for a function which performs a common statistical operation. If the x defined below is a vector of real numbers of length n, what statistic does the aa function compute and return?

```
#include <math.h>
double aa(double *x, int n)
{
 double bb;
 double cc;

 int i;

 bb = 0.0;
 for(i=0; i < n; i++)
   bb += x[i];
 cc = bb/n;

 bb = 0.0;
 for(i=0; i < n; i++)
   bb += (x[i]-cc)*(x[i]-cc);

 return(cc/sqrt(bb/n));
}
```

B. Using either Splus, Minitab or SAS commands, write some code to test the coverage probabilities for 95% confidence intervals for samples drawn from a normal distribution with an unknown mean and an unknown standard deviation. In particular, your code should draw 100 samples, each of size 50, from a normal distribution with mean 5 and standard deviation 2. The code should compute the confidence intervals, and then count the number of intervals which fail to cover the true mean.

C. In part B above, explain what effect each of the following would have on your simulation. [This is not asking how the code would change, but how it would affect the coverage probabilities relative to the results from the simulation above.]

i. Increasing the mean of the normal distribution from which the sets of samples were taken (to some number bigger than 5).

ii. Increasing the standard deviation of the normal distribution from which the sets of samples were taken (to some number bigger than 2).

iii. Increasing the number of sets of samples (to some number bigger than 100).

iv. Increasing the number of units in each sample (to a number bigger than 50).

6. We desire to compare the mean home prices for two local communities. To do so, we will use a 90% confidence interval. Our goal is that the estimate will possess an error of no more than $10,000. We have a general sense that home prices in each of the two communities have a minimum of about $80,000 and almost all are below $350,000. On the other hand, a random sample of size 5 from one of the communities gave us the following (ordered) data: { $97,500 , $121,000 , $148,900 , $167,000 , $200,000 }. Assume that we will sample the same number of houses in each community. How many houses should we sample? Tell why you do what you do.

7. We have two urns, each with ten tags, numbered 1-5 with some duplications. There is at least one tag with each number in both urns. From Urn #1 we obtained two random samples of size n=50 with replacement and from Urn #2 we obtained one sample of size n=50 with replacement. Below are the tallies from the three samples. Note that we have not identified which sample comes from which urn.

Tallies = Number of times in the sample of n=50 that tags with this number appeared

	Tag Number					
	1	2	3	4	5	Total
Sample 1	7	14	3	13	13	50
Sample 2	18	9	8	7	8	50
Sample 3	8	10	7	10	15	50

a. Do a test to see if the pattern of tags in the two urns (and thus for the three samples) is the same. Use $\alpha = .05$. State your conclusion in words about the two urns.

b. If we were to reject the null hypothesis in part (a), we would typically want to investigate a total of 15 differences of probabilities of drawing a particular tag number between pairs of the three samples. State how you would use the Bonferroni technique to find 95% simultaneous confidence intervals for the 15 differences.

c. Find the interval for the difference in probabilities of drawing tag #1 for samples 1 and 2 ONLY.