Department of Mathematics
University of Toledo

Master of Science Degree
Comprehensive Examination
Applied Statistics

April 7, 2001

Instructions:
Do all five problems
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
Each problem is worth 20 points
This is a three hour test

1. An aptitude test consists of 3 components: Verbal, Quantitative, and Analytical. The scores on the Verbal component are normally distributed with mean 70 and standard deviation 5. The scores on the Quantitative component are normally distributed with mean 67 and standard deviation 6. The scores on the Analytical component are normally distributed with mean 76 and standard deviation 4. The correlation between Verbal and Quantitative scores is .51. The correlation between Verbal and Analytical scores is .64. The correlation between Quantitative and Analytical scores is .56. A student's Total score is the sum of that student's Verbal, Quantitative, and Analytical scores.

a. What is the simple correlation between the Total and Verbal scores?

b. What is the partial correlation between the Total and Verbal scores, conditional on fixed levels of the Quantitative and Analytical scores?

A student named Kenny took the aptitude test and obtained a Verbal score of 78, a Quantitative score of 72, and an Analytical score of 81.

c. In what percentile did Kenny's Verbal score fall?

d. In what percentile did Kenny's Quantitative score fall?

e. In what percentile did Kenny's Analytical score fall?

f. In what percentile did Kenny's Total score fall?

g. Is it possible for a student's Total score to fall in a percentile that is higher than the percentile obtained for each of the three individual components? (Hint: What if Kenny had gotten a Quantitative score of 76?)

2. The University of X has 2 professional schools, a Law School and an Engineering School. Over a period of years the following data set, pertaining to admissions broken down by school and gender, was compiled:

School	Gender	Admission Status	Frequency
Law	Male	Admitted	80
Law	Male	Not admitted	289
Law	Female	Admitted	72
Law	Female	Not admitted	174
Engineering	Male	Admitted	151
Engineering	Male	Not admitted	35
Engineering	Female	Admitted	46
Engineering	Female	Not admitted	8

Initially it was argued that there was a bias against female applicants, because the admission rates for females was less than that for males (in the data set for the two schools combined).

a. Compute the admission rates for males and females (in the data set for the two schools combined).

b. Obtain a point estimate and 95% confidence interval for the odds ratio (odds of admission for males vs. females).

c. At this point in the data analysis, does the argument of bias against females with regard to admissions seem to be supported by the data?

After the initial accusations of gender bias were made, University administrators argued that the data should be analyzed taking into account which school a particular student had applied to.

d. Compute the admission rates for males and females (broken down by school).

e. Compute odds ratios (odds of admission for males vs. females) broken down by school, and then conduct a hypothesis test to assess whether it is plausible that the underlying school-specific odds ratios are equal.

f. Under the assumption of equal underlying school-specific odds ratios, obtain a point estimate and 95% confidence interval for the assumed common odds ratio.

g. Does the argument of bias against females with regard to admissions seem to be supported by the data after this analysis?

3. To assess whether there is a difference in the mean birth weight between the first (i.e. μ_1) and second (i.e. μ_2) born of twins, the following data (birth weight measured in lbs) for a simple random sample of 8 twin pairs was obtained:

Twin Pair No.	1	2	3	4	5	6	7	8
First born	7.5	8.6	6.8	7.3	9.3	7.2	7.6	7.7
Second born	7.2	9.4	6.3	7.2	8.5	6.7	8.1	6.8

a. Obtain a point estimate of $\mu_1 - \mu_2$.

b. Obtain a 90% confidence interval for $\mu_1 - \mu_2$.

c. What is the P-value for the hypothesis test that, on the average, the first born of twins outweigh the second born of twins?

d. What assumptions are necessary in order for the above analysis to be valid?

4. Consider the following data on the amount of water applied (in inches) and the yield of alfalfa (in tons per acre) on an experimental farm.

Water	12	18	24	30	36	42	48
Yield	5.27	5.68	6.25	7.21	8.02	8.71	8.42

The following descriptive statistics have been obtained:

$$\overline{X} = 30, \quad \overline{Y} = 7.08, \quad S_x^2 = 168, \quad S_y^2 = 1.879666668, \quad r^2 = .945577989$$

We will assume that all 3 assumptions of the simple linear regression model (linear regression function, constant variance, and normality) are applicable.

a. Obtain a point estimate of the regression slope coefficient. Be sure to include a statement of the associated units of measurement.

b. Obtain a point estimate of the regression intercept coefficient. Be sure to include a statement of the associated units of measurement.

c. Draw a scatterplot of the data, and on the scatterplot draw in the fitted regression line.

d. Determine the fitted and residual values associated with the data point (48, 8.42).

e. Obtain a point estimate of the assumed constant standard deviation. Be sure to include a statement of the associated units of measurement.

f. Determine the P-value for a test of the null hypothesis that the regression slope coefficient is 0.8, against the alternative that the slope is greater than 0.8.

g. Obtain an 80% confidence interval for the mean alfalfa yield (in tons per acre) when the water level is set at 40 inches.

5. Over the decade running from January, 1990 through December, 1999, the earnings for a certain department (say Department A) in Company Y can be computed on a monthly basis. Since this computation is relatively expensive, we wish to sample months in order to estimate the total earnings for Department A over the period. We have selected the method of repeated systematic sampling.

a. Describe the potential advantages and disadvantages of this method as compared to SRS-WOR.

Say that we have chosen to take five one-out-of-twenty systematic samples. That is, we have selected five numbers at random without replacement between 1 and 20, and then sampled every 20th month starting at each of these five. The computed earnings for the sampled months, in thousands of dollars and rounded to the nearest thousand, are listed as follows:

Starting Month	Remaining Sampled Months	Computed Earnings
3 - March, 1990	23, 43, 63, 83, 103	40, 30, 51, 80, 61, 86
5 – May, 1990	25, 45, 65, 85, 105	39, 57, 63, 81, 73, 71
9 – September, 1990	29, 49, 69, 89, 109	46, 57, 47, 97, 79, 96
13 – January, 1991	33, 53, 73, 93, 113	51, 34, 52, 73, 90, 89
16 – April, 1991	36, 56, 76, 96, 116	26, 42, 69, 78, 88, 90

b. Use this data to provide an estimate of the total earnings over the ten years and give a bound on the error of estimation.

c. If we assume that the above data all came from a simple random sample without replacement (which it did not), then give the estimate of the total and the associated error bound.

d. Comment on whether the sample design used (repeated systematic) is good or bad considering that there may be both seasonal variation (patterns that have a period of one year) and an increasing trend over time.