TRENT UNIVERSITY
Faculty of Arts and Science
Final Examinations – 1999/2000

MATHEMATICS 150

PART B: Time: 2 hours and 30 minutes
Books, notes, calculators and 'laptop' computers (with battery power supply) may be used.
Solutions to Part A must be submitted before Part B is commenced.
Each question is worth 16 marks
The **four** questions that you answer best will be counted.

1. a) An operator of a bingo centre has a mailing list that indicates that there are about 3500 'regular' participants who play bingo at the centre on a fairly frequent basis. The operator believes that over half, and 'probably' at least 60% of these 'regulars' are smokers. The operator wants to estimate the percentage of smokers more precisely and wants a 95% chance of estimating the percentage to within 5 percentage points with the percentage in a random sample of 'regulars'. How many 'regulars' should be included in the sample?

b) In a very small survey of 20 bingo players (from a 'general population' of bingo players, not just the centre in part a), there were 15 smokers.

i) Does this sample provide sufficient evidence at the 5% level of significance that over 60% of bingo players are smokers?

ii) What is the *P*-value for part i)?

c) Consider the following Minitab output (below and on the next page) as part of an analysis of a survey of 250 individuals classified in four income categories (low, lower middle, upper middle, high) and four gambling categories (none; lotteries only; lotteries and bingo and casinos; bingo and casinos only.) The survey is to be used to determine whether gambling behaviour differs across income levels. As noted, the analyst asked for all percentages and Minitab organized the categories alphabetically.

i) Produce a revised cross-tabulation display with the categories ordered as listed above and with only percentages applicable to the purpose of the survey.

ii) State the null and alternative hypotheses for a test to determine whether gambling behaviour differs across income levels and state the resulting *P*-value.

ii) Do these data indicate that gambling behaviour differs across income levels? State what level of significance you used.

```
MTB > Read 'gamble' c1 c2;
SUBC>   Format (3x,a6,4x,a8).
Entering data from file: gamble.dat
    250 rows read.
MTB > name c1 'income' c2 'gamble'
```

1. c) (Continued)

```
MTB > #   CROSS-TABULATION & CHI-SQUARE
MTB > Table 'income' 'gamble';
SUBC>   RowPercents;
SUBC>   ColPercents;
SUBC>   TotPercents;
SUBC>   ChiSquare 2.

Tabulated Statistics

 Rows: income     Columns: gamble

       cas_bing     lott  lottand     none      All

 high     50.00    11.11     5.56    33.33   100.00
          17.65     3.85     1.39     8.00     7.20
           3.60     0.80     0.40     2.40     7.20
              9        2        1        6       18
           3.67     3.74     5.18     5.40    18.00

 low      10.00    46.67    23.33    20.00   100.00
           5.88    26.92     9.72     8.00    12.00
           1.20     5.60     2.80     2.40    12.00
              3       14        7        6       30
           6.12     6.24     8.64     9.00    30.00

 lowmid   13.39    18.75    39.29    28.57   100.00
          29.41    40.38    61.11    42.67    44.80
           6.00     8.40    17.60    12.80    44.80
             15       21       44       32      112
          22.85    23.30    32.26    33.60   112.00

 upmid    26.67    16.67    22.22    34.44   100.00
          47.06    28.85    27.78    41.33    36.00
           9.60     6.00     8.00    12.40    36.00
             24       15       20       31       90
          18.36    18.72    25.92    27.00    90.00

 All      20.40    20.80    28.80    30.00   100.00
         100.00   100.00   100.00   100.00   100.00
          20.40    20.80    28.80    30.00   100.00
             51       52       72       75      250
          51.00    52.00    72.00    75.00   250.00

Chi-Square = 36.229, DF = 9, P-Value = 0.000

  Cell Contents --
                  % of Row
                  % of Col
                  % of Tbl
                  Count
                  Exp Freq
```

2. The following data represent a four-year record of quarterly sales figures

t	1	2	3	4	5	6	7	8
sales	1602	2496	1497	3190	1990	2526	1530	3427
t	9	10	11	12	13	14	15	16
sales	966	2683	1784	3360	2065	2720	1857	3349

a) Sketch a time series plot of these data.

b) The quarterly seasonal factors are 0.8, 1.1, 0.7 and 1.4. Determine seasonally adjusted values for each quarter of the third and fourth years.

c) The sales figures are assumed to be following a trend line $y = 2300 + 44\,t$. What are the sales forecasts for all four quarters of year 5?

d) Determine the total sales figure for each of the four years and determine the actual average percentage increase in total sales figures.

3. a) Water quality standards indicate that the concentration of nitrates in drinking water should not exceed 10 mg/L. Agricultural treatments, rainfalls, seepage into aquifers, etc cause fluctuation in the concentration of nitrates present in water in springs and wells. What is the probability that water drawn from a given source at one point in time would have over 10 mg/L if the nitrate concentration fluctuates randomly according to a normal distribution.

i) with a mean of 8.5 mg/L and a standard deviation of 1.5 mg/L?

ii) with a mean of 6.0 mg/L and a standard deviation of 2.0 mg/L?

iii with a mean of 7.0 mg/L and a standard deviation of 1.0 mg/L?

b) Three bottles of water were drawn at various times from a source with random concentrations as in a) i) and put into storage in a lab. As well two bottles were drawn from a source as in a) ii) and five from a source as in a) iii). If one of these ten bottles is selected at random, what is the probability that it will have a nitrate concentration in excess of 10 mg/L?

c) If the bottle selected in part b) did have a nitrate concentration in excess of 10 mg/L, what is the probability that it came from a source as in a) i)?

d) If four of the ten bottles in part b) were selected at random, what is the probability that two of them will be from the source as in a) iii)?

4. The Minitab output on the next **three** pages represents part of an analysis of a sample of drying time (nearest five minutes). Temperature (º C) and relative humidity (%) were recorded as well.

a) Plot a scatter diagram of drying time *vs* temperature and plot a scatter diagram of drying time *vs* humidity. Does one of temperature and humidity appear to be a better predictor of time? If so, which one?

b) What is the correlation between drying time and temperature and between drying time and humidity. Which of temperature and humidity appears to be the better predictor of time? What is the prediction equation with the better predictor?

c) How much is the model improved by adding the second predictor to the model with just the better of the two individual predictors?

d) What are the predicted drying times for a temperature of 15 º C and relative humidity of 60% and for a temperature of 20 º C and relative humidity of 50%?

4. (Continued)

```
MTB > Read 'temps' c1 c2 c3;
Entering data from file: temps.dat
     12 rows read.
MTB > name c1 'temp' c2 'humid' c3 'drytime'
MTB > read c11 c12 c13
DATA> 15 50 *
DATA> 15 60 *
DATA> 20 50 *
DATA> 20 60 *
DATA> end
      4 rows read.
MTB > stack (c1 c2 c3) (c11 c12 c13) (c1 c2 c3)
MTB > sort c1 c2 c3 c1 c2 c3;
SUBC> by c1 c2.
MTB > print c1 c2 c3
```

Data Display

Row	temp	humid	drytime
1	10	60	305
2	10	72	345
3	12	50	270
4	12	77	325
5	13	47	240
6	15	48	240
7	15	50	*
8	15	55	255
9	15	60	*
10	17	63	275
11	17	71	305
12	20	50	*
13	20	60	*
14	20	63	270
15	21	72	280
16	22	75	250

```
MTB > corr c1 c2 c3
```

Correlations (Pearson)

	temp	humid
humid	0.173	
	0.521	
drytime	-0.453	0.637
	0.139	0.026

Cell Contents: Correlation
P-Value

4. (Continued)

```
MTB > regress c3 1 c1

Regression Analysis

The regression equation is
drytime = 336 - 3.68 temp

12 cases used 4 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       336.48       36.26       9.28    0.000
temp           -3.684       2.290      -1.61    0.139

S = 31.45       R-Sq = 20.6%     R-Sq(adj) = 12.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2560.2      2560.2      2.59    0.139
Residual Error    10      9889.8       989.0
Total             11     12450.0


MTB > regr c3 1 c2

Regression Analysis

The regression equation is
drytime = 156 + 1.98 humid

12 cases used 4 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       155.79       48.21       3.23    0.009
humid          1.9794      0.7580       2.61    0.026

S = 27.21       R-Sq = 40.5%     R-Sq(adj) = 34.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      5047.5      5047.5      6.82    0.026
Residual Error    10      7402.5       740.2
Total             11     12450.0

Unusual Observations
Obs      humid    drytime         Fit   StDev Fit    Residual    St Resid
 16       75.0     250.00      304.25       12.16      -54.25      -2.23R

R denotes an observation with a large standardized residual
```

4. (Continued)

```
MTB > regress c3 2 c1 c2;
SUBC> fits c4.

Regression Analysis

The regression equation is
drytime = 202 - 5.87 temp + 2.68 humid

12 cases used 4 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       201.78       24.40       8.27    0.000
temp          -5.8725      0.9996      -5.88    0.000
humid          2.6814      0.3825       7.01    0.000

S = 13.04       R-Sq = 87.7%     R-Sq(adj) = 85.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     10919.0      5459.5     32.09    0.000
Residual Error     9      1531.0       170.1
Total             11     12450.0

Source       DF      Seq SS
temp          1      2560.2
humid         1      8358.8

Unusual Observations
Obs       temp    drytime         Fit   StDev Fit    Residual    St Resid
 16       22.0     250.00      273.70        7.81      -23.70      -2.27R

R denotes an observation with a large standardized residual

MTB > name c4 'fit-time'
MTB > print c1-c4

Data Display

 Row   temp  humid  drytime  fit-time
   1     10     60      305   303.946
   2     10     72      345   336.123
   3     12     50      270   265.387
   4     12     77      325   337.785
   5     13     47      240   251.470
   6     15     48      240   242.406
   7     15     50        *   247.769
   8     15     55      255   261.176
   9     15     60        *   274.584
  10     17     63      275   270.883
  11     17     71      305   292.334
  12     20     50        *   218.407
  13     20     60        *   245.221
  14     20     63      270   253.265
  15     21     72      280   271.526
  16     22     75      250   273.698
```

5. In a comparison of searching times for two different database structures, eight random searches were conducted. Each search was conducted with each of the two structures. The resulting search times are as listed below

Database Search Times

Search	1	2	3	4	5	6	7	8
Search time with structure I	27	16	30	20	24	23	28	20
Search time with structure II	36	14	32	25	33	30	32	21

a) Compare the two sets of search times with a combination box-and-whisker plot.
b) Determine an appropriate one-sided 95% confidence limit to indicate by at least how much search times for the second structure exceed those for the first "on the average" for similar searches.
c) Do the data provide sufficient evidence at the 5% level that search times for the second structure exceed those for the first "on the average" for similar searches?
d) Do the data provide sufficient evidence at the 5% level that search times for the second structure exceed those for the first by at least 3.0 "on the average" for similar searches?

6. a) In a study on percentage loss in performance due to sleep deprivation, it was anticipated that loss (in percent) would be normally distributed with a mean of 30 and a standard deviation of 2.5.
i) If losses were distributed as anticipated, in about how many of 1000 cases would there be a loss of over 35?
i) If losses were distributed as anticipated, what would be the maximum loss in the lowest 25% of all cases?

b) If the standard deviation was as anticipated but the mean was unknown, how many sample losses would be required to have a 95% chance of having an estimation error of at most 1.0 when estimating the unknown population mean with the sample mean?
c) Suppose that neither the population mean nor standard deviation were known and that a sample of 25 losses produced a sample mean of 36 and a sample standard deviation of 2.8. Determine a 95% confidence limit to indicate at least how much of a loss would occur "on the average."

7. In an analysis of field equipment battery lifetimes under one set of environmental operating conditions, six batteries of each of three types (one type from group A and three types from group B) were monitored for useful life. The resulting lifetimes were entered into Minitab in C1 through C4 and then analysed as indicated in the session on the next page.
a) Compare the four battery type lifetimes with a mean and standard deviation display.
b) Do the data provide evidence of different mean lifetimes? What is the *P*-value?
c) Which battery types differ from which?
d) What distribution assumptions are used in b) and c)?

7. (Continued)

```
MTB > desc c1-c4

Descriptive Statistics

Variable             N       Mean     Median     TrMean      StDev    SE Mean
A                    6      486.3      479.5      486.3       36.5       14.9
B1                   6      616.3      589.5      616.3       66.8       27.3
B2                   6      571.8      561.5      571.8       38.9       15.9
B3                   6      613.8      607.0      613.8       47.0       19.2

Variable       Minimum    Maximum         Q1         Q3
A                443.0      542.0      455.0      521.7
B1               556.0      718.0      562.8      689.5
B2               528.0      623.0      538.5      617.0
B3               562.0      695.0      575.5      647.8

MTB > stack c1-c4 c11;
SUBC> subs c10.
MTB > name c10 'type' c11 'time'
MTB > Oneway 'time' 'type';
SUBC>   Fisher .0083333.

One-way Analysis of Variance

Analysis of Variance for time
Source     DF        SS        MS        F        P
type        3     66326     22109     9.30    0.000
Error      20     47566      2378
Total      23    113892
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
1           6    486.33     36.47  (------*------)
2           6    616.33     66.82                       (------*------)
3           6    571.83     38.85               (------*------)
4           6    613.83     47.00                      (------*------)
                                   ------+---------+---------+---------+
Pooled StDev =    48.77                480       540       600       660

Fisher's pairwise comparisons

    Family error rate = 0.0384
Individual error rate = 0.00833
Critical value = 2.927

Intervals for (column level mean) - (row level mean)

                   1           2           3

       2      -212.4
               -47.6

       3      -167.9       -37.9
                -3.1       126.9

       4      -209.9       -79.9      -124.4
               -45.1        84.9        40.4
```