UNIVERSITI SAINS MALAYSIA

First Semester Examination
Academic Session 2008/2009

November 2008

MGM 561 – Statistical Methods for Research
[Kaedah Statistik untuk Penyelidikan]

Duration : 3 hours
[Masa : 3 jam]

Please check that this examination paper consists of FIFTEEN pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi LIMA BELAS muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]

Instructions: Answer **all four** [4] questions.

*[**Arahan:** Jawab **semua empat** [4] soalan.]*

1. (a) Improperly filled orders are costly problem for mail-order houses. To estimate the mean loss per incorrectly filled order, a large firms plan to sample n incorrectly filled orders and to determine the added cost associated with each one. The firm estimates that the added cost is between \$53 and \$430. How many incorrectly filled orders must be sampled to estimate the mean additional cost using a 95% confidence interval of width \$25?

(b) A standardized math test is given to 30 students in the tenth grade. Their scores with descriptive statistics and a stem-and-leaf plot are listed here:

44	49	62	45	51	59	57	55	70	64	54
58	65	75	43	42	67	63	71	54	60	53
40	49	52	54	61	42	38				

```
Descriptive Statistics:
N              Mean      Median      TrMean      StDev      SE Mean
30             54.90     54.00       54.73       9.75       1.78

Minimum        Maximum   Q1          Q2
38.00          75.00     48.00       62.25

Stem-and-leaf of math N = 30
Leaf Unit = 1.0

  1   3   8
  6   4   02234
  9   4   599
 (7)  5   0123444
 14   5   5789
 10   6   01234
  5   6   57
  3   7   01
  1   7   5
```

(i) Do you think that the data are bell shaped? If so, interpret the amount of variability in the data using the empirical rule.

(ii) If 10 is added to every observation, what will be the new mean, median, and trimmed mean?

(iii) Generally, what effect does adding every observation by a constant have on measures of center? What effect does it have on measures of variability?

(c) (i) A campground has 5 rustic campsites not accessible to campers on wheels. Some nights, some of these campsites are unoccupied because of the small number of campers with equipment for such campsites. The ranger keeps track of the number of unoccupied sites for 50 nights.

1. (a) *Borang pesanan yang salah diisi menyebabkan masalah kos pada syarikat pesanan pos. Untuk menganggar purata kerugian bagi setiap borang pesanan yang salah diisi ini, sebuah syarikat besar merancang untuk membuat pensampelan n borang pesanan yang salah diisi untuk menganggarkan kos tambahan bagi setiap borang tersebut. Syarikat menganggarkan bahawa kos tambahan ialah antara $53 dan $430. Berapakah bilangan borang pesanan yang salah diisi diperlukan untuk persampelan bagi menganggar selang keyakinan 95% kos tambahan dengan panjang selang sebanyak $25.*

(b) *Satu ujian matematik yang piawai diberikan kepada 30 pelajar dalam gred sepuluh. Markah mereka dengan statistik deskriptif dan rajah stem-dan leaf seperti berikut diperolehi;*

44	49	62	45	51	59	57	55	70	64	54
58	65	75	43	42	67	63	71	54	60	53
40	49	52	54	61	42	38				

```
Descriptive Statistics:
N            Mean      Median     TrMean     StDev     SE Mean
30           54.90     54.00      54.73      9.75      1.78

Minimum      Maximum   Q1         Q2
38.00        75.00     48.00      62.25

Stem-and-leaf of math N = 30
Leaf Unit = 1.0

  1   3  8
  6   4  02234
  9   4  599
 (7)  5  0123444
 14   5  5789
 10   6  01234
  5   6  57
  3   7  01
  1   7  5
```

(i) *Pada fikiran anda, adakah data berbentuk loceng? Jika ya, berikan tafsiran jumlah variability dalam data tersebut menggunakan 'empirical rule'.*

(ii) *Jika 10 ditambah pada setiap cerapan, apakah nilai baru bagi min, median dan min terpangkas?*

(iii) *Secara umum, apabila setiap cerapan ditambah dengan suatu pemalar terhadap, apakah kesannya terhadap ukuran memusatnya? Apakah pula kesannya terhadap ukuran variabiliti?*

(c) (i) *Satu kawasan perkhemahan mempunyai 5 tapak khemah yang ringkas yang tidak dapat dimasuki oleh pekhemah yang membawa kenderaan. Tapak khemah ini tidak berpenghuni beberapa malam disebabkan bilangan yang kecil dalam kalangan pekhemah yang mempunyai alatan untuk tapak khemah seperti itu. Polis hutan menyimpan rekod bilangan tapak khemah yang tidak diduduki untuk 50 malam seperti berikut:*

Number unoccupied	0	1	2	3	4	5
Frequency	22	20	7	1	0	0

Do these data fit a binomial distribution?

(ii) It is reported that offspring of users of a certain drug may have a higher incidence of birth defects than the general population. To obtain information about a possible relationship between this drug and birth defects, 100 offspring of female rats are fed the drug and 100 offspring from untreated female rats are examined. The results are given below:

	Progeny	
Females	Birth Defects	Normal
Treated	30	70
Untreated	20	80

Analyze these data at $\alpha = 0.05$. Is this a test of homogeneity or independence? State your conclusion from this study?

[25 marks]

2. (a) The following values were computed from the length of life of two brands of light bulbs (in hour):

	Brand A	*Brand B*
n	9	25
$\bar{y}$	1560	1573
$\sum(y-\bar{y})^2$	440	1860

(i) Is there a difference in the variability of lifetimes for the two brands of bulbs? (Use $\alpha = 0.02$).

(ii) Find a 98% confidence interval on the ratio of the two variability.

(b) An oncologist wants to evaluate the usefulness of the CAT scan for uterine tumor diagnosis. For 12 women with fibroid tumors, certain measurements are taken by CAT scan techniques prior to surgery and then compared with other measurements taken on the tumors in the pathology laboratory after they had been surgically removed. Suppose the paired measurements on tumor mass are:

Patient	**A**	**B**	**C**	**D**	**E**	**F**
CAT scan, x	18	17	28	20	11	24
Pathology, y	20	4	25	16	19	21
Patient	**G**	**H**	**I**	**J**	**K**	**L**
CAT scan, x	16	15	19	24	23	13
Pathology, y	22	10	23	27	18	11

Bilangan tapak khemah tak diduduki	*0*	*1*	*2*	*3*	*4*	*5*
Kekerapan	*22*	*20*	*7*	*1*	*0*	*0*

Adakah data ini bersesuaian dengan taburan Binomial?

(ii) *Dilaporkan bahawa pengguna suatu dadah berkemungkinan tinggi mendapat anak dengan kecacatan semasa lahir berbanding populasi yang umum. Untuk mendapat maklumat bagi kemungkinan hubungan antara penggunaan dadah dengan kecacatan semasa lahir, 100 anak daripada tikus betina yang diberikan dadah dan 100 anak daripada tikus betina yang tidak diberikan dadah, diperiksa. Hasilnya seperti ditunjukkan di bawah:*

	Anak	
Betina	*Cacat*	*Normal*
Diberi dadah	*30*	*70*
Tidak diberi dadah	*20*	*80*

Analisis data tersebut dengan $\alpha = 0.05$. *Adakah ini merupakan Ujian Homogenus atau Ujian Ketaksandaran? Nyatakan kesimpulan dari kajian ini.*

[25 markah]

2. (a) *Nilai berikut, diperolehi daripada panjang hayat mentol lampu bagi dua jenama lampu (dalam jam):*

	Brand A	*Brand B*
n	*9*	*25*
$\bar{y}$	*1560*	*1573*
$\sum(y-\bar{y})^2$	*440*	*1860*

(i) *Adakah terdapat perbezaan dalam variabiliti terhadap hayat lampu bagi kedua-dua jenama tersebut?*

(ii) *Carikan selang keyakinan 98% terhadap nisbah kedua variability itu.*

(b) *Seorang ahli onkologi mahu mentaksir kepentingan imejan CAT bagi diagnosis ketumbuhan uterin. Untuk 12 wanita yang mempunyai ketumbuhan fibroid, pengukuran diambil menggunakan imejan CAT sebelum pembedahan dan dibandingkan dengan ukuran ketumbuhan dalam makmal patologi selepas pembedahan. Andaikan ukuran bagi ketumbuhan adalah seperti berikut:*

Pesakit	***A***	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***	***K***	***L***
Imejan CAT, x	*18*	*17*	*28*	*20*	*11*	*24*	*16*	*15*	*19*	*24*	*23*	*13*
Patologi, y	*20*	*4*	*25*	*16*	*19*	*21*	*22*	*10*	*23*	*27*	*18*	*11*

and the statistics computed are

$$\sum(x-\bar{x})^2 = 278\,;\ \sum(y-\bar{y})^2 = 498\,;$$

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = 0.723\,; \qquad \frac{[\sum(x-\bar{x})(y-\bar{y})]^2}{\sum(x-\bar{x})^2} = 108.58$$

(i) Find the sample correlation coefficient.

(ii) State the most logical hypotheses about the correlation between the CAT scan measurement of tumor mass and that obtained at pathology.

(iii) Test your hypothesis in (ii) for an $\alpha = 0.05$

(iv) Do you think the relationship would be useful in being able to use the CAT scan information to predict fibroid tumor mass prior to surgery?

[25 marks]

3. (a) The data given here are from analyses of the magnesium concentration in stream samples that were collected along a river. Sampling locations were identified on an aerial photograph, and later the distances between samples were measured. The data and the regression output are listed here:

Distance(m)	0	1820	2542	2889	3460	4586	6020	6841	7323	10903
Magnesium(mg)	6.44	8.61	5.24	5.73	3.81	4.05	2.95	2.57	3.37	3.84

Distance (m)	1109	11922	12530	14065	14937	16244	17632	19002	20860	22471
Magnesium(mg)	2.86	1.22	1.09	2.36	2.24	2.05	2.23	0.42	0.87	1.26

```
The regression equation is
magnesiu = 5.77 - 0.000252 distance


Predictor         Coef      SE Coef      T      P
Constant        5.7678       0.4952  11.65  0.000
distance   -0.00025184   0.00004027  -6.25  0.000


S = 1.19526   R-Sq = 68.5%   R-Sq(adj) = 66.7%
```

Dan statistik berikut diperolehi;

$$\sum(x-\bar{x})^2 = 278\,; \ \sum(y-\bar{y})^2 = 498\,;$$

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = 0.723\,; \qquad \frac{[\sum(x-\bar{x})(y-\bar{y})]^2}{\sum(x-\bar{x})^2} = 108.58$$

(i) *Cari pekali korelasi sampel tersebut.*

(ii) *Nyatakan hipotesis yang logik tentang korelasi antara ukuran oleh imejan CAT dengan ukuran secara patologi.*

(iii) *Uji hipotesis anda dalam (ii) pada* $\alpha = 0.05$.

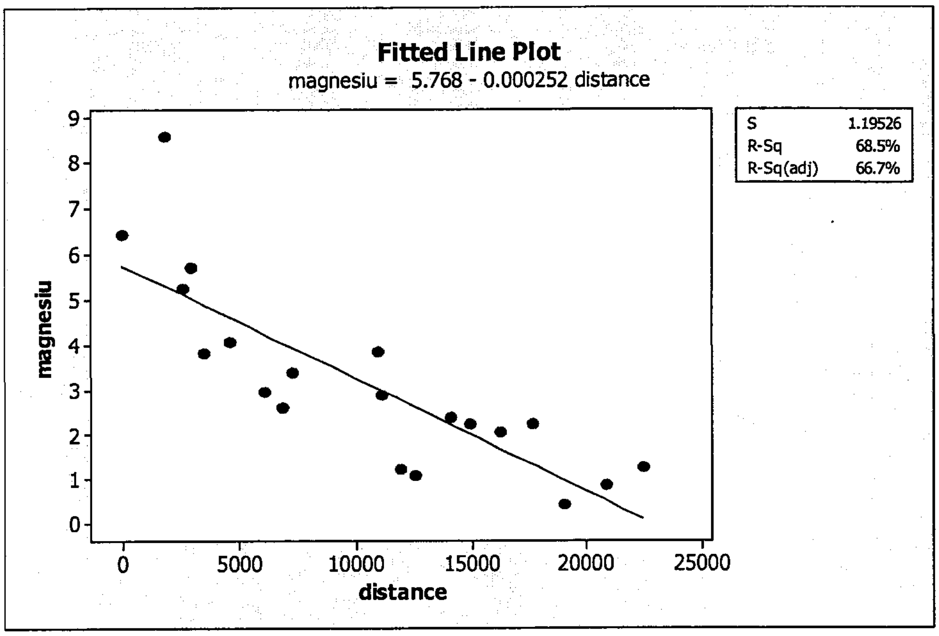
(iv) *Pada pendapat anda adakah hubungan ini berguna untuk membolehkan penggunaan imejan CAT untuk meramal saiz ketumbuhan fibroid sebelum pembedahan?*

[25 markah]

3. (a) *Diberikan data daripada analisis kepekatan magnesium dalam sampel aliran air sungai yang dikumpulkan sepanjang sebuah sungai. Lokasi persampelan ditentukan menggunakan satu gambaran udara (areal photograph)l, dan kemudian jarak di antara sample diukur. Data dan output analisis regresi adalah seperti berikut:*

Jarak (m)	*0*	*1820*	*2542*	*2889*	*3460*	*4586*	*6020*	*6841*	*7323*	*10903*
Magnesium(mg)	*6.44*	*8.61*	*5.24*	*5.73*	*3.81*	*4.05*	*2.95*	*2.57*	*3.37*	*3.84*
Jarak (m)	*11098*	*11922*	*12530*	*14065*	*14937*	*16244*	*17632*	*19002*	*20860*	*22471*
Magnesium (mg)	*2.86*	*1.22*	*1.09*	*2.36*	*2.24*	*2.05*	*2.23*	*0.42*	*0.87*	*1.26*

```
The regression equation is
magnesiu = 5.77 - 0.000252 distance


Predictor         Coef      SE Coef        T       P
Constant        5.7678       0.4952    11.65   0.000
distance   -0.00025184   0.00004027    -6.25   0.000


S = 1.19526    R-Sq = 68.5%    R-Sq(adj) = 66.7%
```



(i) From the regression output, determine the independent variable and the dependent variable.

(ii) What is the estimated mean for magnesium when the distance is 30 000m?

(iii) What is the regression equation. Does it fit the data reasonably well?

(iv) What would you suggest to describe relationship between the magnesium and the distance (based on the Fitted Line Plot)?

(b) A study to relate the number of years of experience of traffic police with the average number of tickets they write per week found these data (with the regression output):

Model : $y = \beta_0 + \beta_1 x + \varepsilon$

Police	1	2	3	4	5	6	7	8	9	10
Years	3	8	2	15	5	20	1	10	7	12
Tickets	42	30	54	12	32	8	75	28	20	15

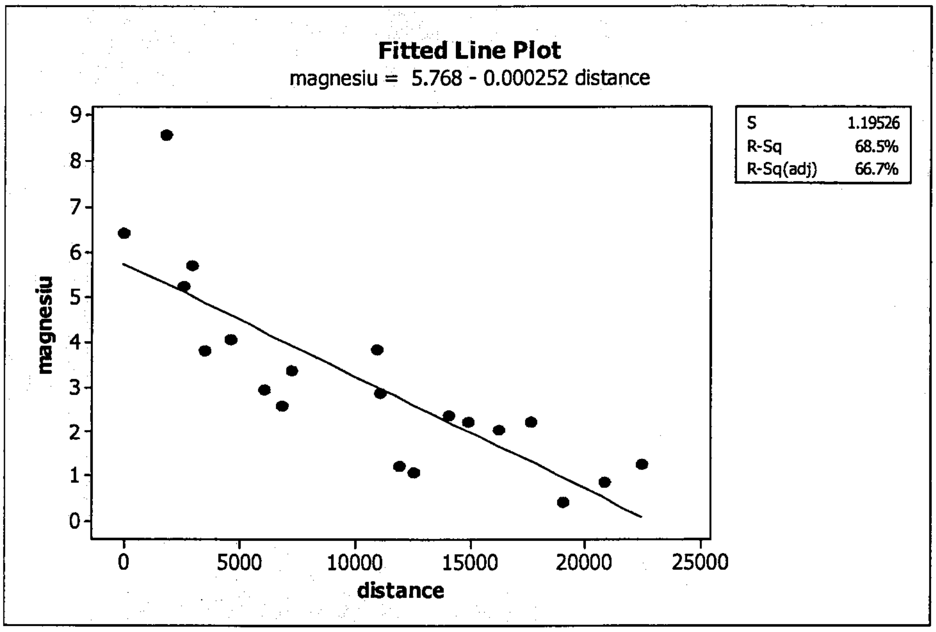
Regression output:

```
The regression equation is
tickets = 55.9 - 2.93 years


Predictor       Coef   SE Coef       T       P
Constant      55.938     6.215    9.00   0.000
years        -2.9322    0.6150   -4.77   0.001


S = 11.2081    R-Sq = 74.0%    R-Sq(adj) = 70.7%
```

(i) Find a 98% confidence interval for β_1.

(ii) Test the hypothesis that $\beta_1 = 0$ $(\text{use } \alpha = 0.05)$. State your conclusion.

[25 marks]



(i) *Daripada output regresi tersebut, nyatakan pemboleh ubah bersandar dan pemboleh ubah tak bersandar.*

(ii) *Apakah anggaran bagi min magnesium pada jarak 30 000m?*

(iii) *Nyatakan persamaan regresi. Adakah ianya dapat memberikan anggaran yang baik terhadap data tersebut?*

(iv) *Apakah yang anda cadangkan untuk menghurai hubungan antara magnesium dengan jarak tersebut (berdasarkan rajah garis yang dipadankan).*

(b) *Satu kajian untuk menghubungkan bilangan tahun pengalaman seorang polis trafik dengan purata bilangan saman yang mereka keluarkan dalam seminggu mendapati data seperti berikut (dengan output regresi):*

Model: $y = \beta_0 + \beta_1 x_1 + \varepsilon$

Police	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Years	*3*	*8*	*2*	*15*	*5*	*20*	*1*	*10*	*7*	*12*
Tickets	*42*	*30*	*54*	*12*	*32*	*8*	*75*	*28*	*20*	*15*

Regression output:

```
The regression equation is
tickets = 55.9 - 2.93 years

Predictor      Coef   SE Coef      T      P
Constant     55.938     6.215   9.00  0.000
years       -2.9322    0.6150  -4.77  0.001

S = 11.2081   R-Sq = 74.0%   R-Sq(adj) = 70.7%
```

(i) *Tentukan selang keyakinan 98% untuk* β_1.

(ii) *Lakukan ujian hipotesis bahawa* $\beta_1 = 0$ $\left(\text{guna } \alpha = 0.05\right)$. *Nyatakan kesimpulan anda.*

[25 markah]

4. To evaluate cigarettes, the Federal Trade Commision (FTC) uses "smoking machines" that measure the tar, nicotine, and carbon monoxide in each cigarette. Carbon monoxide has been linked to heart disease, tar has been linked to cancer, and nicotine is additive. Suppose that the amount of tar (measured in milligrams) is recorded for 25 cigarettes randomly selected from each of four brands (brand A, brand B, brand C and brand D). Referring to the Figure 1,

   (a) Discuss on distribution and normality of the data.

   (b) Using the ANOVA analysis, determine whether one brand is "lowest' in tar.

   (c) Analyze with Tukey's multiple comparison procedure to determine which differences between means are statistically significant.

[25 marks]

4. *Untuk mengkaji rokok, Suruhanjaya Perdagangan Persekutuan (FTC) menggunakan 'mesin rokok' yang mengukur kandungan tar, nikotin and karbon monoksida bagi setiap rokok. Karbon monoksida dikaitkan dengan penyakit hati, tar dikaitkan dengan kanser dan nikotin menyebabkan ketagihan. Katakan jumlah tar (diukur dalam milligram) direkodkan bagi 25 rokok yang dipilih secara rawak masing-masing daripada empat jenama (jenama A, jenama B, jenama C dan jenama D). Merujuk Rajah 1,*

(a) *Bincangkan taburan dan kenormalan data tersebut.*

(b) *Berdasarkan analisis varians, tentukan sama ada suatu jenama itu mengandungi tar yang terendah.*

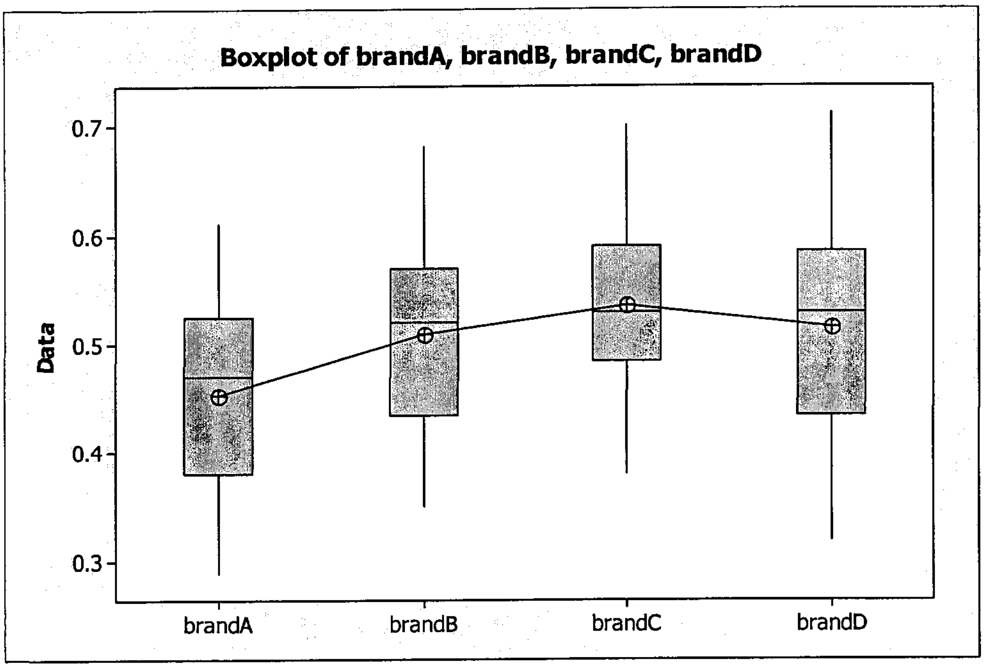
(c) *Menggunakan analisis Tukey untuk perbandingan berganda, tentukan perbezaan antara min yang signifikan secara statistik.*

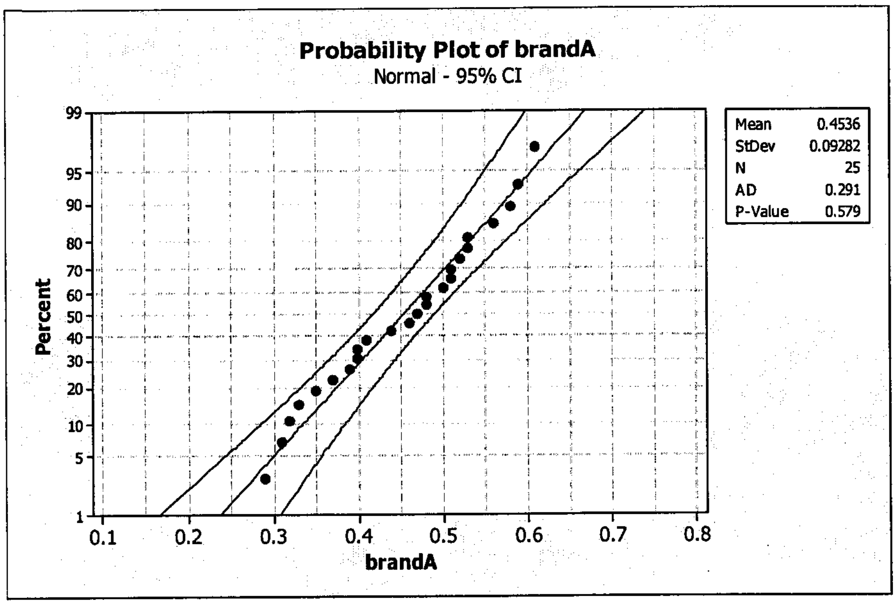
[25 markah]

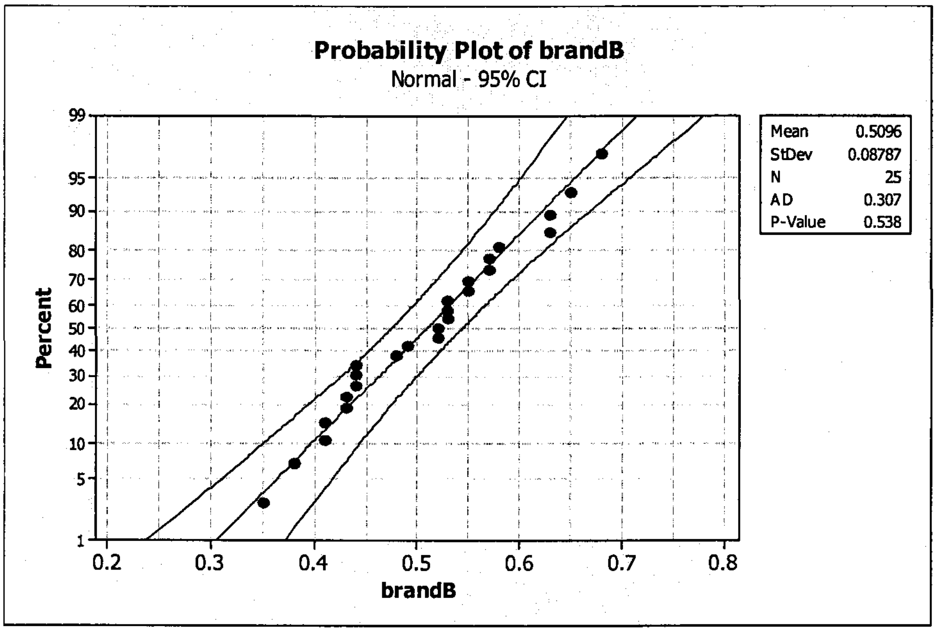
Figure 1 : Minitab output for Question 4

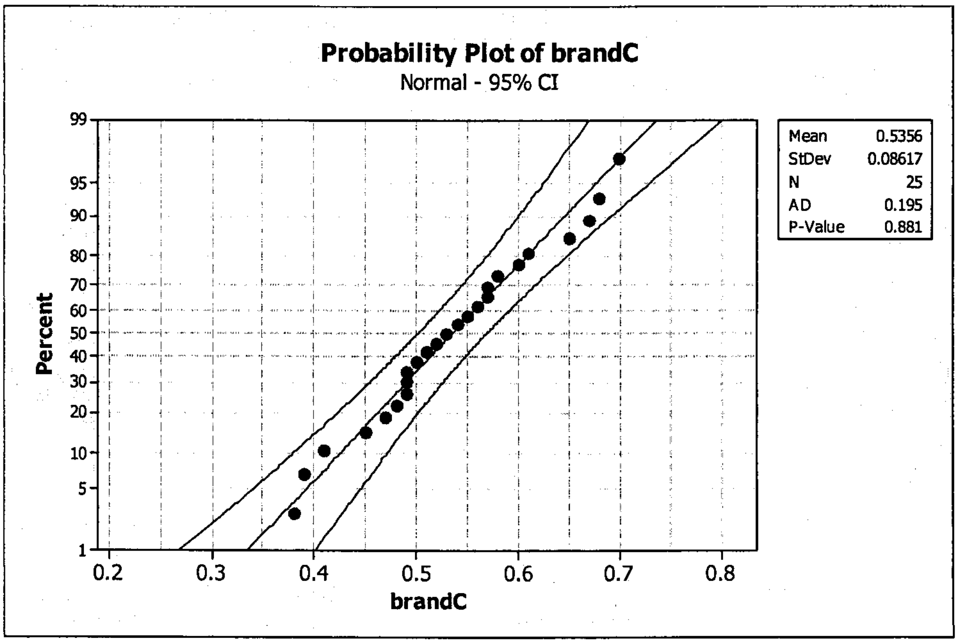
One-way ANOVA: brandA, brandB, brandC, brandD

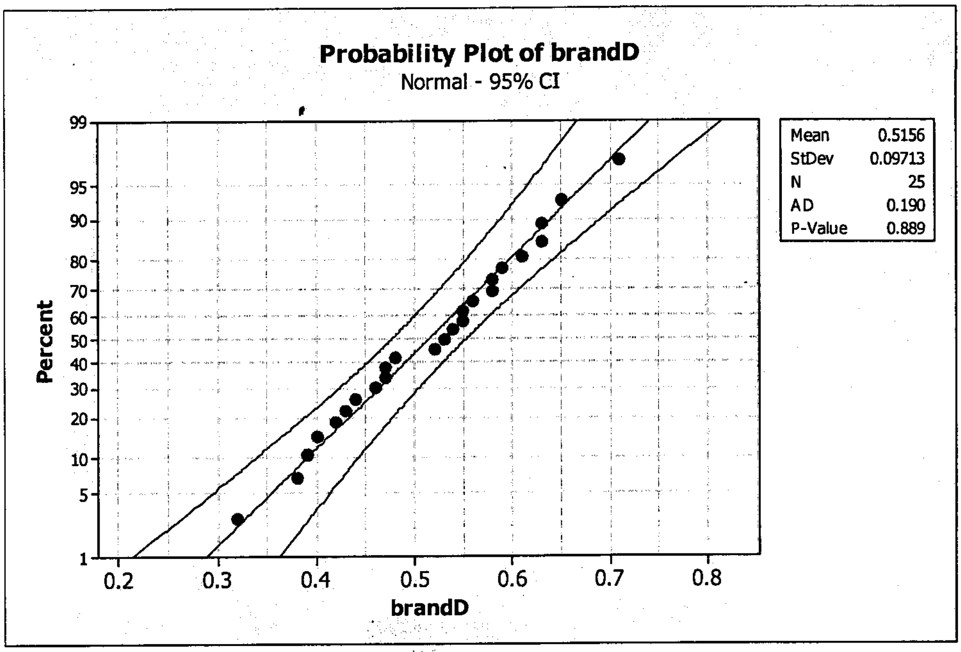
```
Source  DF       SS       MS     F      P
Factor   3  0.09260  0.03087  3.72  0.014
Error   96  0.79670  0.00830
Total   99  0.88930

S = 0.09110   R-Sq = 10.41%   R-Sq(adj) = 7.61%


                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level    N     Mean    StDev  ------+---------+---------+---------+---
brandA  25  0.45360  0.09282  (--------*--------)
brandB  25  0.50960  0.08787              (--------*--------)
brandC  25  0.53560  0.08617                     (--------*--------)
brandD  25  0.51560  0.09713                (--------*--------)
                              ------+---------+---------+---------+---
                                  0.440     0.480     0.520     0.560

Pooled StDev = 0.09110


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98.97%


brandA subtracted from:

          Lower   Center    Upper  ---+---------+---------+---------+------
brandB  -0.01141  0.05600  0.12341           (---------*---------)
brandC   0.01459  0.08200  0.14941               (---------*--------)
brandD  -0.00541  0.06200  0.12941            (---------*--------)
                                   ---+---------+---------+---------+------
                                   -0.070     0.000     0.070     0.140


brandB subtracted from:

          Lower   Center    Upper  ---+---------+---------+---------+------
brandC  -0.04141  0.02600  0.09341       (---------*--------)
brandD  -0.06141  0.00600  0.07341    (---------*--------)
                                   ---+---------+---------+---------+------
                                   -0.070     0.000     0.070     0.140


brandC subtracted from:

          Lower    Center    Upper  ---+---------+---------+---------+-----
brandD  -0.08741  -0.02000  0.04741  (--------*--------)
                                    ---+---------+---------+---------+-----
                                    -0.070     0.000     0.070     0.140
```


Boxplot of brandA, brandB, brandC, brandD
Data
0.7
0.6
0.5
0.4
0.3
brandA
brandB
brandC
brandD


Probability Plot of brandA
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
brandA
Mean 0.4536
StDev 0.09282
N 25
AD 0.291
P-Value 0.579


Probability Plot of brandB
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
brandB
Mean 0.5096
StDev 0.08787
N 25
AD 0.307
P-Value 0.538


Probability Plot of brandC
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
brandC
Mean 0.5356
StDev 0.08617
N 25
AD 0.195
P-Value 0.881


Probability Plot of brandD
Normal - 95% CI
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
brandD
Mean 0.5156
StDev 0.09713
N 25
AD 0.190
P-Value 0.889

- ooo O ooo -