Co-funded by the
Erasmus+ Programme
of the European Union

Name of Degree: Msc. Urban Agriculture

Semester: Ist

Course: Statistics

Name of author: Dr. Iliriana Miftari

Institutional affiliation: Faculty of Agriculture and Veterinary, University of Prishtina

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

Co-funded by the
Erasmus+ Programme
of the European Union

Statistics

LECTURE

Introduction and descriptive statistics

Below you find data based on a study of firewood consumption in the district of Sumbawanga, Tanzania by F.B.N. Sabuni (Thesis 1985). We have selected 12 observations of both firewood consumption (in cubic meter per family per year) and income (in Tanzanian shillings per month).

firewood	income
2,81	7800
1,60	3200
2,97	5900
1,90	4100
1,01	6100
3,35	6700
3,56	4900
3,30	5400
1,11	3200
2,49	3700
2,88	6700
1,20	2100

1) Punch the firewood data given above. Give names to the variables.

2) Save the dataset by: FILE → SAVE CURRENT WORKSHEET AS. Give the name yourself.

3) Sort the data according to increasing firewood consumption, and put the sorted data into 2 new columns. Use DATA → SORT, try the rest yourself. Remember that you must sort the income variable as well.

4) Make two histograms (of relative frequencies), one for firewood and one for income. Use GRAPH → HISTOGRAM → SIMPLE ok at Y-Scale type. Choose density ok. Give each histogram a title by clicking at labels. Save each histogram. Make your own name for it, but keep the three letters after the dot (mgf). Please note that graphics always should have these 3 letters after the dot.

5) Make two dot plots, one for firewood and one for income. Use GRAPH → DOTPLOT. By regarding the dot plots, guess the center of gravity (the sample mean).

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

6) Find the sample mean, sample median and sample standard deviation for both variables. Use STAT BASIC STATISTIC DESPLAY DESCRIPTIVE STATISTICS.

7) If you are interested in just one descriptive measure, use CALC ⟶ COLUMN STATISTICS. Find the sample mean of firewood consumption. Also find the sum of squares. What is the mathematical expression for sum of squares?

8) Construct two new variables, one measuring annual income and one measuring annual firewood consumption in kilograms (assume that one cubic meter is 700 kg). Use CALC CALCULATOR, try the rest yourself. Please note that all arithmetic's are done by calc-calculator.

9) Create a new variable by adding 1000 (T sh) to all income observations. What do you think are the values of the sample mean and the sample standard deviation for this new variable? Check it out.

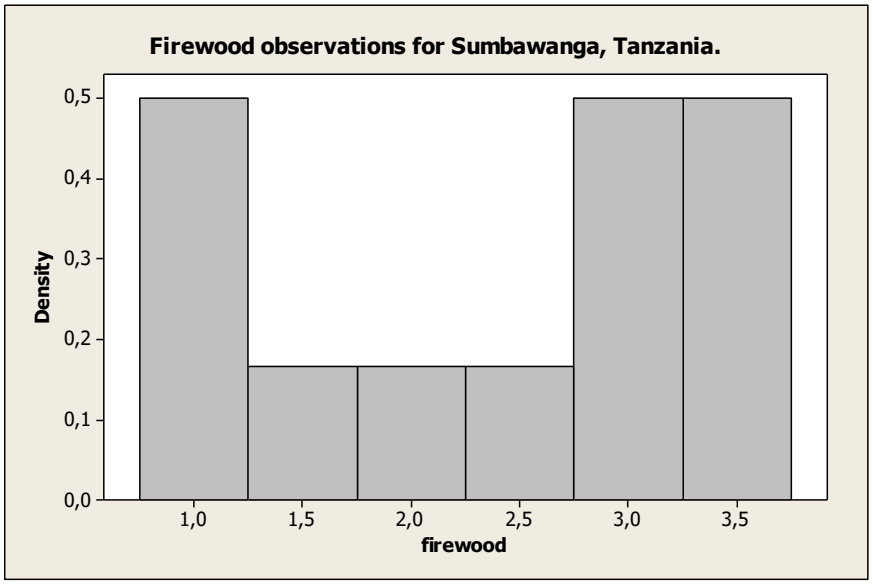
10) Save today's work. Use FILE-----SAVE PROJECT AS. Make a name for it.

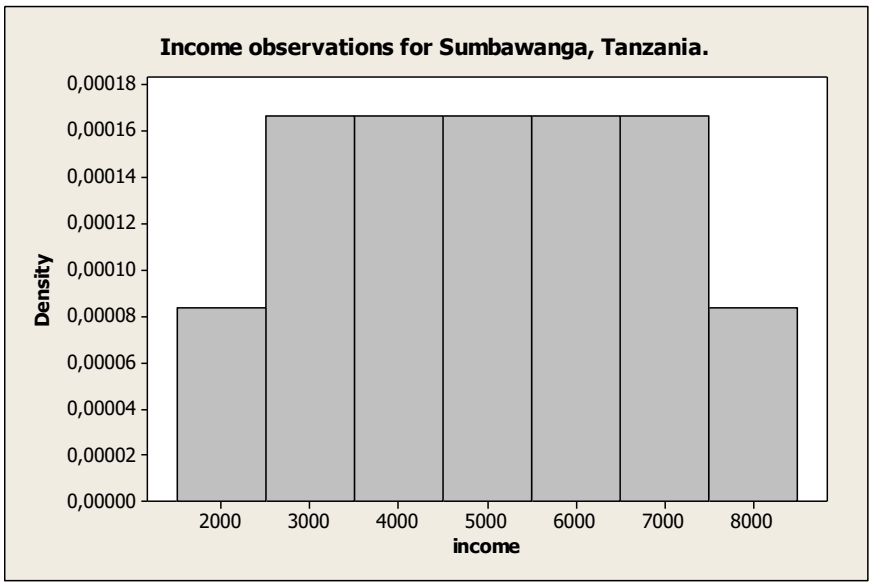
Minitab Lecture 1.
Data Display

Row	firewood	income
1	2,81	7800
2	1,60	3200
3	2,97	5900
4	1,90	4100
5	1,01	6100
6	3,35	6700
7	3,56	4900
8	3,30	5400
9	1,11	3200
10	2,49	3700
11	2,88	6700
12	1,20	2100

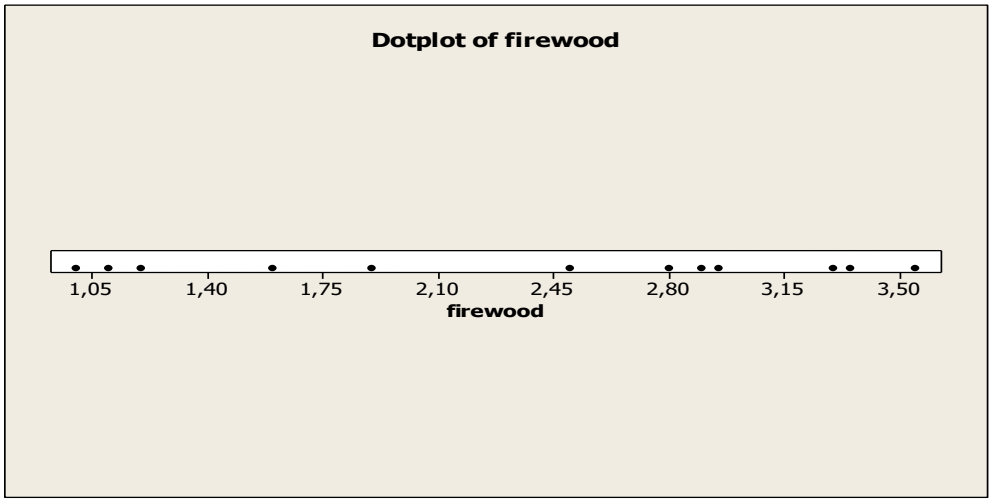
4) Histogram of firewood

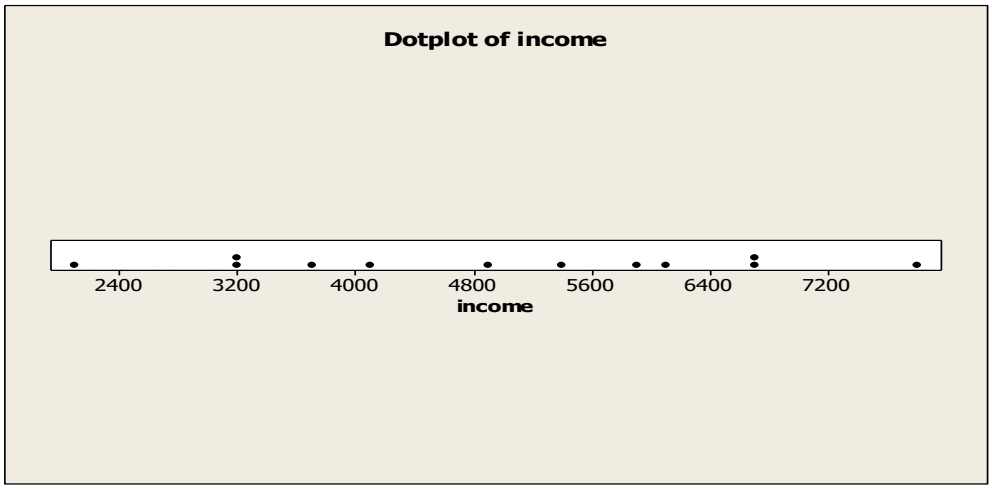
Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"


Firewood observations for Sumbawanga, Tanzania.
Density
0,0
0,1
0,2
0,3
0,4
0,5
1,0
1,5
2,0
2,5
3,0
3,5
firewood


Income observations for Sumbawanga, Tanzania.
Density
0,00000
0,00002
0,00004
0,00006
0,00008
0,00010
0,00012
0,00014
0,00016
0,00018
2000
3000
4000
5000
6000
7000
8000
income

5)





I guess the sample mean for firewood is 2.5 and for income: 5000.

6) Descriptive Statistics: firewood; income

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
firewood	12	0	**2,348**	0,271	**0,938**	1,010	1,300	**2,650**	3,218	3,560
income	12	0	**4983**	501	**1737**	2100	3325	**5150**	6550	7800

7)Mean of firewood

Mean of firewood = 2,34833

Sum of Squares of firewood

Sum of squares (uncorrected) of firewood = 75,8598 = $\sum_{i=1}^{n} x_i^2$

8) Annual income: expression: income*12, annual firewood in kg.: firewood*7000.

9) Descriptive Statistics: inc+1000

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
inc+1000	12	0	5983	501	1737	3100	4325	6150	7550	8800

The sample mean will increase by 1000 if the income observations increase by 1000. The sample standard deviation will not change.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

Data Display

Row	firewood	income	firesort	incom	annuinc	firekg	inc+1000
1	2,81	7800	1,01	6100	93600	1967	8800
2	1,60	3200	1,11	3200	38400	1120	4200
3	2,97	5900	1,20	2100	70800	2079	6900
4	1,90	4100	1,60	3200	49200	1330	5100
5	1,01	6100	1,90	4100	73200	707	7100
6	3,35	6700	2,49	3700	80400	2345	7700
7	3,56	4900	2,81	7800	58800	2492	5900
8	3,30	5400	2,88	6700	64800	2310	6400
9	1,11	3200	2,97	5900	38400	777	4200
10	2,49	3700	3,30	5400	44400	1743	4700
11	2,88	6700	3,35	6700	80400	2016	7700
12	1,20	2100	3,56	4900	25200	840	3100

Exercise 2.32

```
salary
   2450   2275   2425   4700   2650   2350   2475
```

Descriptive Statistics: salary

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
salary	7	0	2761	326	863	2275	2350	2450	2650	4700

a) The sample mean = 2761 and the median = 2450.
b) The median is preferable because one large observation influences the mean very much.

Exercise 2.38.

Descriptive Statistics: y2002

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
y2002	16	0	30,50	5,64	22,54	1,00	16,25	28,50	40,00	80,00

a) The sample mean = 30.5.
b) Large observations will influence the sample mean a lot.

Exercise 2.89.

Descriptive Statistics: differen

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
differen	16	0	-8,81	5,06	20,24	-67,00	-16,25	-6,00	2,00	24,00

a) The sample mean = -8.81 and the sample standard deviation is 20.24.
b) Large differences will influence the sample mean.

Data Display

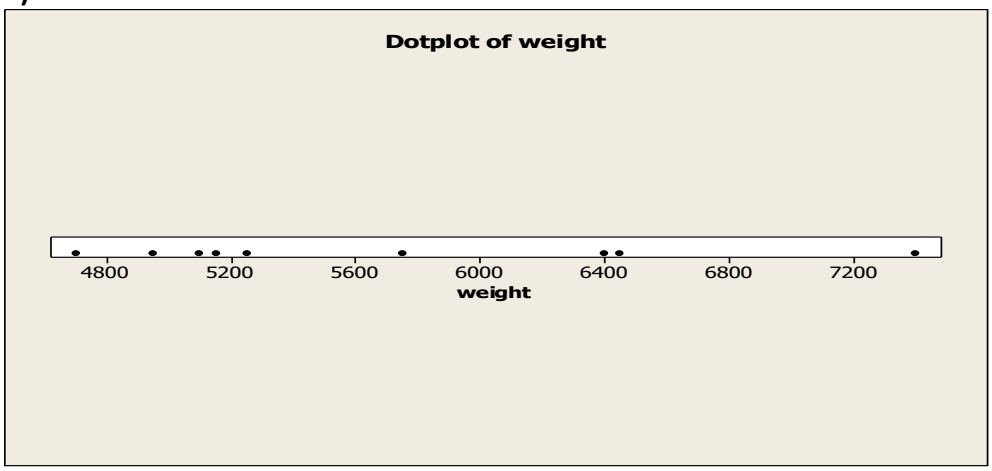
Row	y2002	y1997	differen
1	24	26	2
2	42	30	-12
3	1	8	7
4	2	9	7
5	15	15	0
6	26	12	-14
7	23	47	24
8	80	63	-17
9	1	3	2
10	31	23	-8
11	33	32	-1
12	53	20	-33
13	69	2	-67
14	34	15	-19
15	20	14	-6

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

16 34 28 -6

Minitab lecture 2.

1)



Observations have values from 4690 to 7410 kg. There is one elephant in the sample which is much heavier than the other elephants. The majority of the elephants have weights less than 5400 kg.

2)

Descriptive Statistics: weight

```
Variable  N  N*  Mean  SE Mean  StDev  Minimum    Q1  Median    Q3  Maximum
weight    9   0  5687      299    896     4690  5045    5230  6430     7410
```

$\bar{x} = 5687$ std($\bar{x}$) = 299

3) s=896

4)

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean     99% CI
weight    9  5687    896      299  (4684; 6689)
```

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean     95% CI
weight    9  5687    896      299  (4998; 6376)
```

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean     90% CI
weight    9  5687    896      299  (5131; 6242)
```

5)

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean    99,9% CI
weight    9  5687    896      299  (4180; 7193)
```

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

The interval is extremely wide because the chance that this interval will not cover E(X) is just 0.1%.

6) The length of the interval is: $2t_{\frac{\alpha}{2},8} \cdot \frac{s}{\sqrt{n}} = 900$ which gives:

$$t_{\frac{\alpha}{2},8} = 900/(2\frac{s}{\sqrt{n}}) = \frac{900}{2 \cdot 299} = 1.505$$

Calc $\rightarrow$ Probability Distributions $\rightarrow$ t…
Degrees of freedom: 8
Input constant: -1,505

Cumulative Distribution Function

```
Student's t distribution with 8 DF

     x  P( X <= x )
-1,505    0,0853699
```

$\alpha/2=0.08537$ and $\alpha = 2 \cdot 0.08537 = 0.1704$.
We have to calculate a 1-0.1704 = 0.8296 confidence interval for E(X).

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean    82,96% CI
weight    9  5687    896      299  (5237; 6137)
```

The chance that this interval will cover E(X) is 82.96%.

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean    82,95% CI
weight    9  5687    896      299  (5237; 6137)
```

7)

Descriptive Statistics: weight

```
Variable   N  N*  Mean  SE Mean  StDev  Minimum    Q1  Median    Q3  Maximum
weight    14   0  5541      208    778     4690  4953    5265  6105     7410
```

s=778

One-Sample T: weight

```
Variable   N  Mean  StDev  SE Mean     99% CI
weight    14  5541    778      208  (4915; 6168)
```

One-Sample T: weight

```
Variable   N  Mean  StDev  SE Mean     95% CI
weight    14  5541    778      208  (5092; 5991)
```

One-Sample T: weight

```
Variable   N  Mean  StDev  SE Mean      90% CI
weight    14  5541    778      208  (5173; 5910)
```

One-Sample T: weight

```
Variable   N  Mean  StDev  SE Mean    99,9% CI
weight    14  5541    778      208  (4664; 6419)
```

The intervals are shorter now because the sample size is larger and std($\bar{x}$)= 208now, and it was 299 before.

8)

One-Sample T: weight

```
Test of mu = 5700 vs < 5700
                                    99,9%
                                    Upper
Variable  N  Mean  StDev  SE Mean  Bound      T      P
weight    9  5687    896      299   7032  -0,04  0,483
```

The p-value is 0.483 and we retain H_0 at a 5% level of significance. We can't state that the elephants in the park have reduced their mean weight during the dry period.

9)

One-Sample T: weight

```
Test of mu = 5800 vs < 5800
                                    99,9%
                                    Upper
Variable  N  Mean  StDev  SE Mean  Bound      T      P
weight    9  5687    896      299   7032  -0,38  0,357
```

The null hypothesis is retained at a 5% level of significance, because the p-value = 0.357>0.05. The sample mean is 5687, which is closer to 5700 than to 5800. If the distance between the sample mean and the test mean increases, then the p-value will decrease.

10)

One-Sample T: weight

```
Test of mu = 6200 vs not = 6200


Variable  N  Mean  StDev  SE Mean    99,9% CI        T      P
weight    9  5687    896      299  (4180; 7193)  -1,72  0,124
```

The p-value= 0.124>0.05 and we retain H_0 at a 5% level of significance. We have not proven that the elephants in the park have a mean weight which is different from 6200.

11) The value in H_0: μ=6200 is not in the 82.96% confidence interval for μ, but it is inside all other intervals which are calculated when n=9. This shows that we must choose α large to be able to reject H_0.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

12) We have: 1-p-value=0.876. This means: a 87% confidence interval will not cover 6200, a 88% confidence interval will cover 6200.

One-Sample T: weight

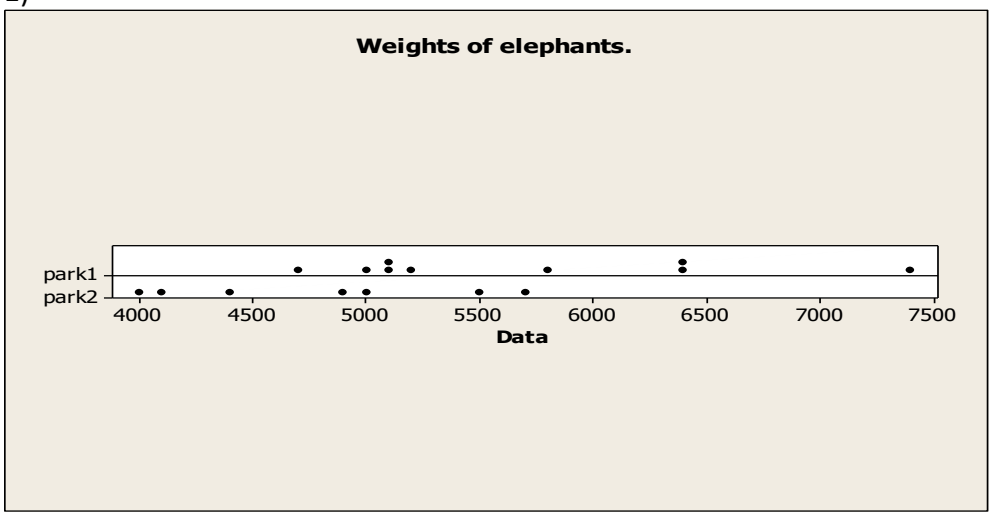
```
Variable  N  Mean  StDev  SE Mean     87% CI
weight    9  5687    896      299  (5182; 6191)
```

One-Sample T: weight

```
Variable  N  Mean  StDev  SE Mean     88% CI
weight    9  5687    896      299  (5167; 6207)
```

Minitab lecture 3.

1)



These dotplots indicate that male elephants in park 1 are at the average heavier than male elephants in park 2.

2)

Descriptive Statistics: park1; park2

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
park1	9	0	5687	299	896	4690	5045	5230	6430	7410
park2	7	0	4813	249	659	4040	4120	4930	5470	5740

Park 1: $\bar{x} = 5687$ $s_1 = 896$ Park 2: $\bar{y} = 4813$ $s_2 = 659$

An estimate of E(X) – E(Y) is: 5687 – 4813 = 874. We have: $\frac{s_1}{s_2} = \frac{896}{659} = 1.36$ which is inside [0.5 , 2] and we assume that the two populations have equal standard deviation.

3)

We have two groups with no specific relationship between one observation in group 1 and one observation in group 2. We have no matched pairs.

Two-Sample T-Test and CI: park1; park2

```
Two-sample T for park1 vs park2

       N  Mean  StDev  SE Mean
park1  9  5687    896      299
park2  7  4813    659      249


Difference = mu (park1) - mu (park2)
Estimate for difference:  874
95% CI for difference:  (5; 1742)
T-Test of difference = 0 (vs not =): T-Value = 2,16  P-Value = 0,049  DF = 14
Both use Pooled StDev = 803,4454
```

We can use the confidence interval to test H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 \neq 0$. The 95% confidence interval for $\mu_1 - \mu_2$ does not contain 0 , and we can reject H_0 at a 5% level of significance.

4)

The model : We assume: Park 1: $X_1, X_2, \ldots, X_9$ are $N(\mu_1, \sigma)$ and

park 2: $Y_1, Y_2, \ldots, Y_7$ are $N(\mu_2, \sigma)$ and all observations are independent.

Two-Sample T-Test and CI: park1; park2

```
Two-sample T for park1 vs park2

       N  Mean  StDev  SE Mean
park1  9  5687    896      299
park2  7  4813    659      249


Difference = mu (park1) - mu (park2)
Estimate for difference:  874
95% lower bound for difference:  161
T-Test of difference = 0 (vs >): T-Value = 2,16  P-Value = 0,024  DF = 14
Both use Pooled StDev = 803,4454
```

We test H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 > 0$. We find t = 2.16 and $t_{0.05,14}$=1.761 < t and we can reject H_0 at a 5% level of significance. The p-value for this test is 0.024. This tells us that the smallest level of significance we can choose and still be able to reject H_0 is 2.4%.

5)

H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 > 200$.

Two-Sample T-Test and CI: park1; park2

```
Two-sample T for park1 vs park2

       N  Mean  StDev  SE Mean
park1  9  5687    896      299
park2  7  4813    659      249


Difference = mu (park1) - mu (park2)
Estimate for difference:  874
95% lower bound for difference:  161
T-Test of difference = 200 (vs >): T-Value = 1,66  P-Value = 0,059  DF = 14
Both use Pooled StDev = 803,4454
```

The p-value = 0.059 and we can reject H_0 at a 5.9% level of significance. If we want a 5% level of significance, we must retain H_0.

6)

Data → Stack → Columns. Use column C3 as storage column for the observations, C4 as subscript column, C4 tells which park the observations are from.

Data Display

Row	park1	park2	weights	park
1	7410	5010	7410	park1
2	5230	4120	5230	park1
3	6440	4040	6440	park1
4	4970	4930	4970	park1
5	4690	5740	4690	park1
6	6420	5470	6420	park1
7	5130	4380	5130	park1
8	5770		5770	park1
9	5120		5120	park1
10			5010	park2
11			4120	park2
12			4040	park2
13			4930	park2
14			5740	park2
15			5470	park2
16			4380	park2

Descriptive Statistics: weights

Variable	park	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
weights	park1	9	0	5687	299	896	4690	5045	5230	6430
	park2	7	0	4813	249	659	4040	4120	4930	5470

Variable	park	Maximum
weights	park1	7410
	park2	5740

7)

Two-Sample T-Test and CI: weights; park

```
Two-sample T for weights

park   N  Mean  StDev  SE Mean
park1  9  5687    896      299
park2  7  4813    659      249

Difference = mu (park1) - mu (park2)
Estimate for difference:  874
95% CI for difference:  (5; 1742)
T-Test of difference = 0 (vs not =): T-Value = 2,16  P-Value = 0,049  DF = 14
Both use Pooled StDev = 803,4454
```

Two-Sample T-Test and CI: weights; park

```
Two-sample T for weights

park   N  Mean  StDev  SE Mean
park1  9  5687    896      299
park2  7  4813    659      249

Difference = mu (park1) - mu (park2)
Estimate for difference:  874
95% lower bound for difference:  161
T-Test of difference = 0 (vs >): T-Value = 2,16  P-Value = 0,024  DF = 14
Both use Pooled StDev = 803,4454
```

See the answers in 3 and 4.

8) This is a matched pairs situation because we observe the same person twice. There is a relationship between the i-th observation at day 2 and day 4.
An independent sampling would have been: Assume Day 2: $X_1, X_2,\ldots,X_7$ are $N(\mu_1,\sigma)$ and day 4: $Y_1, Y_2,\ldots,Y_7$ are $N(\mu_2,\sigma)$, all observations are independent. We cannot assume that all our observations are independent. X_1 and Y_1 are dependent and so on.
9)

Descriptive Statistics: day2; day4

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
day2	7	0	257,0	10,2	27,0	218,0	238,0	257,0	270,0	302,0
day4	7	0	236,00	7,00	18,53	201,00	232,00	233,00	248,00	260,00

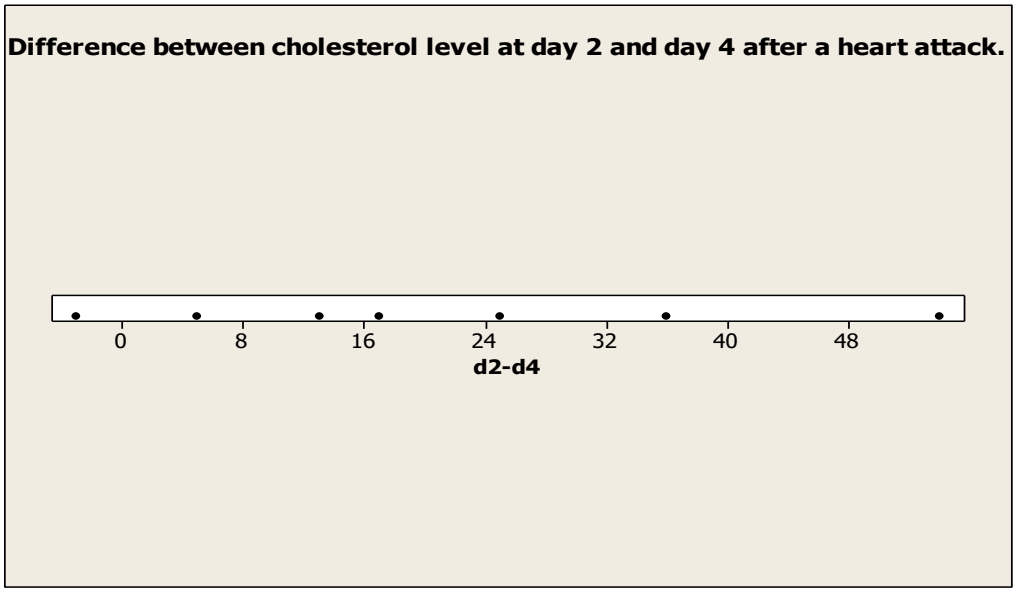
Day 2: $\bar{x} = 257$ Day 4: $\bar{y} = 236$ This could indicate that there is a reduction in cholesterol level from day 2 to day 4.
10)
We calculate $d_i = x_i - y_i$:

Data Display

Row	day2	day4	d2-d4	patient
1	218	201	17	1
2	238	233	5	2
3	270	245	25	3
4	269	233	36	4
5	302	248	54	5
6	257	260	-3	6
7	245	232	13	7

11)

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"



A negative value tells us that there has been an increase in the cholesterol level from day2 to day 4.

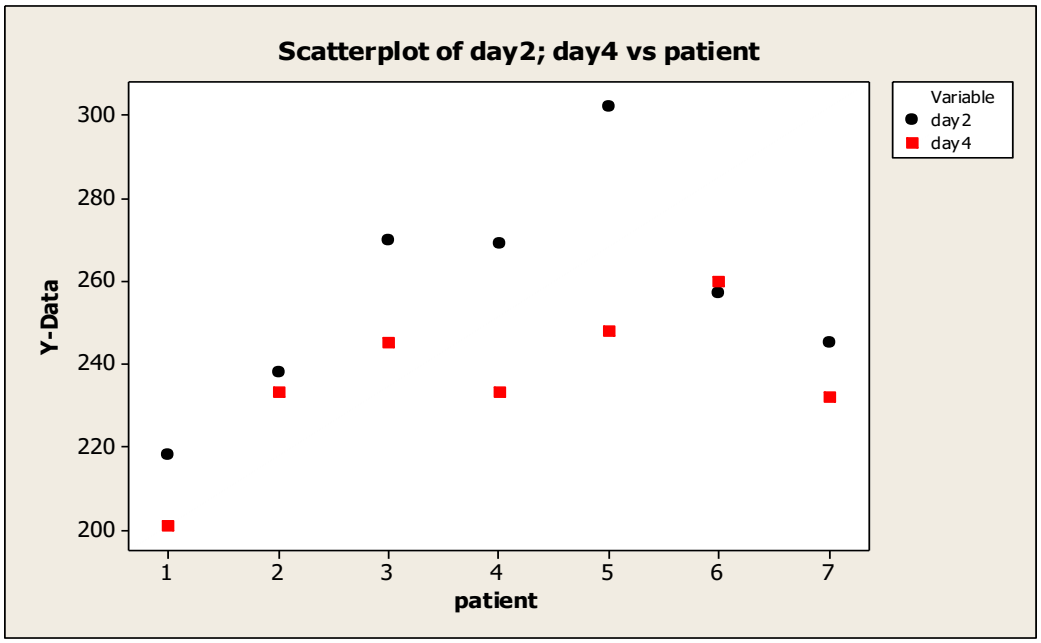
12)

Graph → Scatterplot → Simple

Specify:

Y	X
day2	patient
day4	patient

Click at Multiple Graphs, mark overlaid on the same graph.



13)

Paired T-Test and CI: day2; day4

```
Paired T for day2 - day4

            N   Mean  StDev  SE Mean
day2        7  257,0   27,0     10,2
day4        7  236,0   18,5      7,0
Difference  7  21,00  19,33     7,31


99% CI for mean difference: (-6,09; 48,09)
T-Test of mean difference = 0 (vs not = 0): T-Value = 2,87  P-Value = 0,028
```

There is a 99% chance that this specific interval will cover E(D) = E(X) – E(Y).

14)

We assume: $D_1, D_2, \ldots D_7$ are independent $N(\delta, \sigma_D)$

Paired T-Test and CI: day2; day4

```
Paired T for day2 - day4

            N   Mean  StDev  SE Mean
day2        7  257,0   27,0     10,2
day4        7  236,0   18,5      7,0
Difference  7  21,00  19,33     7,31


99% lower bound for mean difference: -1,96
T-Test of mean difference = 0 (vs > 0): T-Value = 2,87  P-Value = 0,014
```

We test: H_0: $\delta = 0$ against H_1: $\delta > 0$. The p-value = 0.014 < 0.05 and we can reject H_0 at a 5% level of significance.

15)

Paired T-Test and CI: day2; day4

```
Paired T for day2 - day4

            N   Mean  StDev  SE Mean
day2        7  257,0   27,0     10,2
day4        7  236,0   18,5      7,0
Difference  7  21,00  19,33     7,31


99% lower bound for mean difference: -1,96
T-Test of mean difference = 10 (vs > 10): T-Value = 1,51  P-Value = 0,091
```

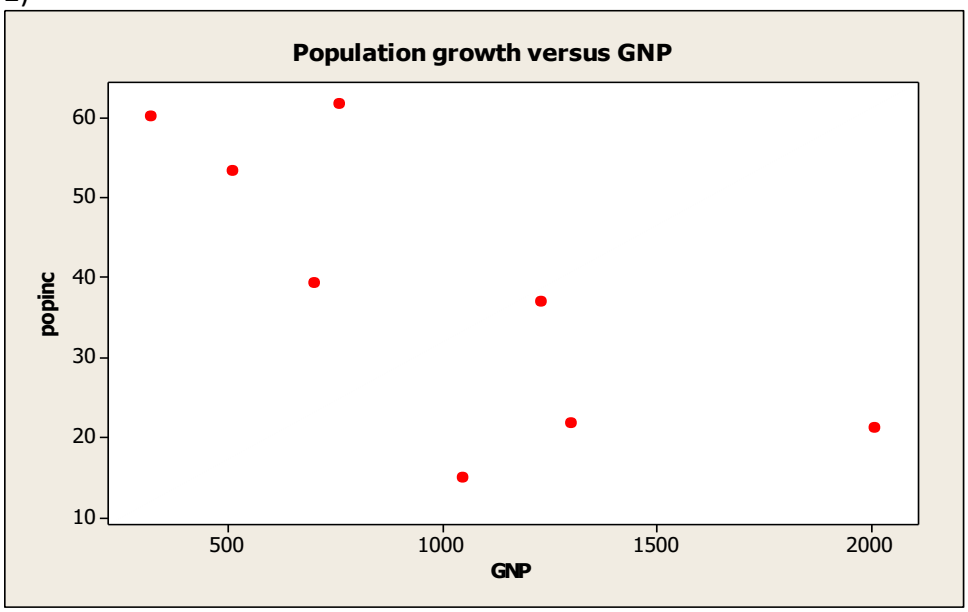
We test: H_0: $\delta = 10$ against H_1: $\delta > 10$. The p-value = 0.091 > 0.05 and we retain H_0 at a 5% level of significance. We can't state that the reduction is more than 10.

Minitab lecture 4.

1)**Data Display**

Row	popinc	GNP
1	53,2	510
2	14,9	1050
3	39,3	700
4	60,1	320
5	21,7	1300
6	61,6	760
7	21,1	2010
8	36,9	1230

2)



I guess r = - 0.5

Correlations: popinc; GNP

```
Pearson correlation of popinc and GNP = -0,752
P-Value = 0,032
```

The correlation coefficient gives a measure of the strength of the linear relationship between population increase and GNP.

3)

The model is: $Y_i = \beta_0 + \beta_1 x_i + e_i$, we assume $e_1,....,e_n$ are independent N(0 , σ)

Y = population increase

x = GNP

We have 3 unknown parameters in this model. The model seems to fit the data quite well.

Interpretation of the parameters:

β_0 = the mean population increase for countries with GNP = 0.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

β_1 = the mean increase of the population increase in a country if GNP increases by 1.
σ = the standard deviation of an error-term.

4)

Regression Analysis: popinc versus GNP

```
The regression equation is
popinc = 63,9 - 0,0257 GNP

Predictor       Coef    SE Coef       T      P
Constant       63,95      10,19    6,27  0,001
GNP        -0,025735   0,009222   -2,79  0,032

S = 13,0871   R-Sq = 56,5%   R-Sq(adj) = 49,2%

Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  1333,7  1333,7  7,79  0,032
Residual Error    6  1027,6   171,3
Total             7  2361,3
```

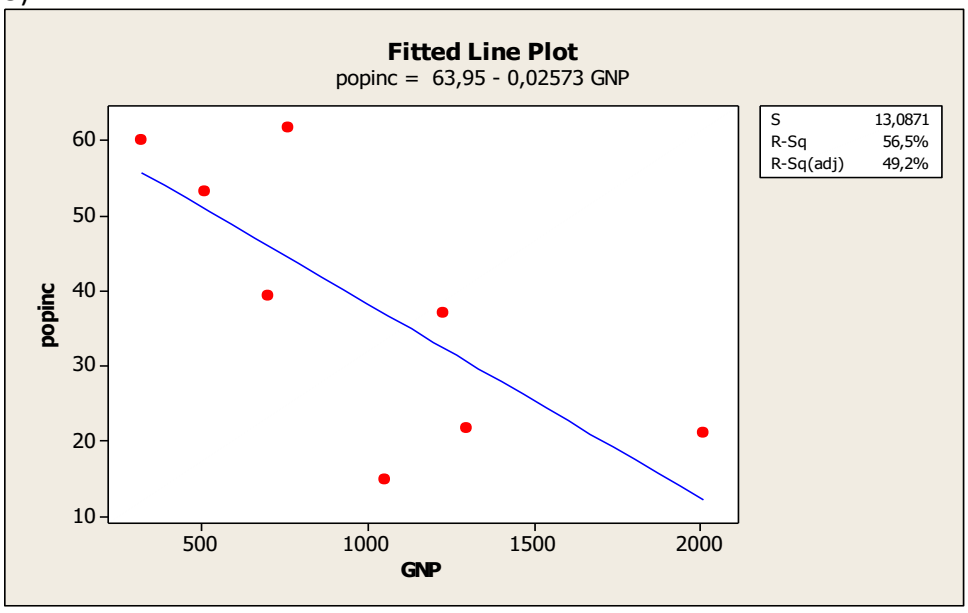
We find $\hat{\beta}_0$ = 63.95 $\hat{\beta}_1$ = - 0.0257 and $\hat{\sigma}$ = s = 13.09

If GNP is 0 then the population increase has the estimate $\hat{\beta}_0$ = 63.95.

If GNP increases by 1 then the estimate of the population increase decreases by -$\hat{\beta}_1$ = 0.0257

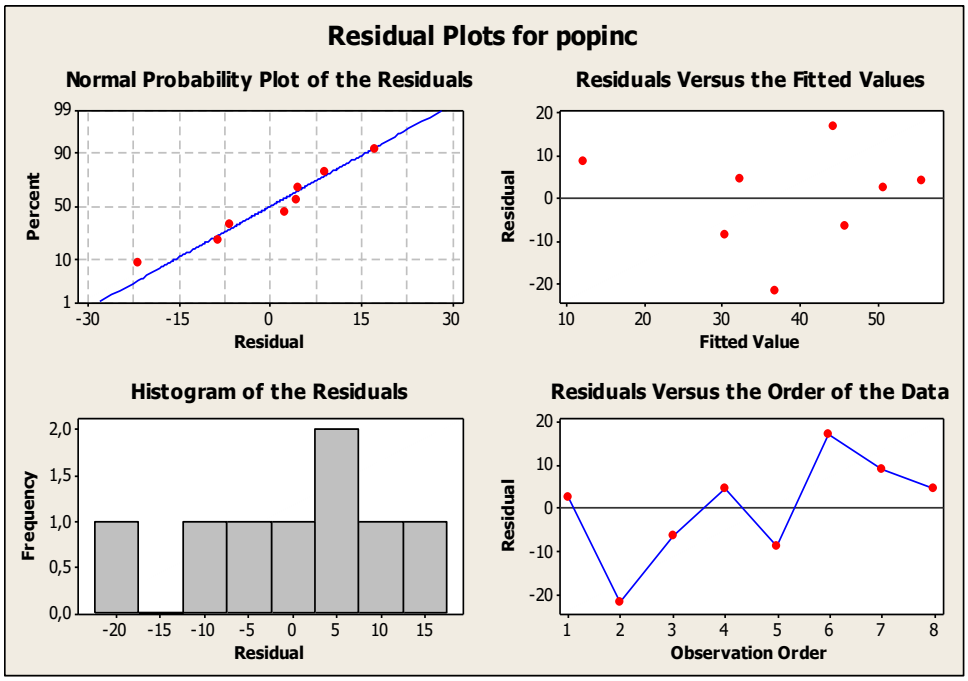
An estimate of the standard deviation of the error term is $\hat{\sigma}$ = s = 13.09, and this is also an estimate of the standard error of one Y-observation.

5)



6)

To make a residual plot, one has to repeat question 4, then click at Graphs and specify Four in one.



A residual is an estimate of an error term , $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ = the difference between the observation of the response and the fitted line when x = x_i.

A fitted value = $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ = is the y-value on the fitted line when x = x_i. When x = x_i a prediction of the y-value is $\hat{Y}_i$.

7)

We test H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. The p-value for this test is 0.032<0.05 and we reject H_0 at a 5% level of significance. This is the smallest (α) level of significance we can choose and still reject H_0.

8)

We test H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$. We have : $t_{\hat{\beta}_1}$ = - 2.79 and - $t_{0.05,6}$= -1.943> $t_{\hat{\beta}_1}$ We can reject H_0 at a 5% level of significance.

This supports the claim that increased standard of living is a good method to prevent population increase.

9)Lower limit = $\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2} s_{\hat{\beta}_1}$ = - 0.0257 - 2.447 • 0.0092 = - 0.0482

Upper limit = $\hat{\beta}_1 + t_{\frac{\alpha}{2},n-2} s_{\hat{\beta}_1}$ = - 0.0257 + 2.447 • 0.0092 = - 0.0032

[- 0.0482 , - 0.0032] is a 95% confidence interval for β_1. The length of the interval is

- 0.0032 + 0.0482 = 0.0450

Interpretation of the interval: The chance that this interval will cover the true β_1 is 95%.

10) **Regression Analysis: popinc versus GNP**

```
The regression equation is
popinc = 63,9 - 0,0257 GNP

Predictor       Coef   SE Coef      T      P
Constant       63,95     10,19   6,27  0,001
GNP        -0,025735  0,009222  -2,79  0,032

S = 13,0871   R-Sq = 56,5%   R-Sq(adj) = 49,2%

Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  1333,7  1333,7  7,79  0,032
Residual Error    6  1027,6   171,3
Total             7  2361,3

Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI          95% PI
  1  53,65    7,11  (36,26; 71,05)  (17,21; 90,10)

Values of Predictors for New Observations

New
Obs  GNP
  1  400
```

A prediction of the population increase for Kenya is: 53.65. A 95% prediction interval for Y in Kenya is [17.21 , 90.10].

The interval is wide because GNP for Kenya is far away from the sample mean of GNP.

11) **Descriptive Statistics: GNP**

```
Variable  N  N*  Mean  SE Mean  StDev  Minimum   Q1  Median    Q3  Maximum
GNP       8   0   985      190    536      320  558     905  1283     2010
```

We have : $\bar{x}$ = 985. The shortest 95% prediction interval will be when GNP = 985.

Regression Analysis: popinc versus GNP

```
The regression equation is
popinc = 63,9 - 0,0257 GNP

Predictor       Coef   SE Coef      T      P
Constant       63,95     10,19   6,27  0,001
GNP        -0,025735  0,009222  -2,79  0,032

S = 13,0871   R-Sq = 56,5%   R-Sq(adj) = 49,2%

Analysis of Variance
```

```
Source           DF      SS      MS     F      P
Regression        1  1333,7  1333,7  7,79  0,032
Residual Error    6  1027,6   171,3
Total             7  2361,3

Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI          95% PI
  1  38,60    4,63  (27,28; 49,92)  (4,63; 72,57)

Values of Predictors for New Observations

New
Obs  GNP
  1  985
```

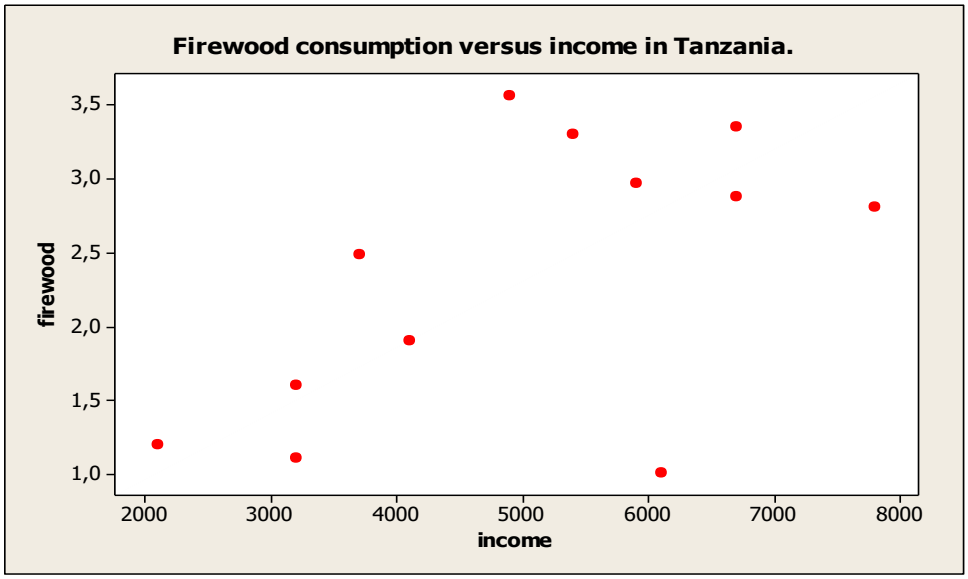
The 95% prediction interval is:
[4.63,72.57]. This is the shortest 95% prediction interval we can have here.

Minitab lecture 5.

1)

Data Display

Row	firewood	income	distance	famsize
1	2,81	7800	1,3	8
2	1,60	3200	1,7	4
3	2,97	5900	1,1	6
4	1,90	4100	3,2	4
5	1,01	6100	7,9	7
6	3,35	6700	1,1	7
7	3,56	4900	1,0	5
8	3,30	5400	1,0	6
9	1,11	3200	6,6	3
10	2,49	3700	1,3	5
11	2,88	6700	1,2	8
12	1,20	2100	2,7	4



I don't think income can explain much of the variation in firewood consumption. The observations are far away from a straight line.

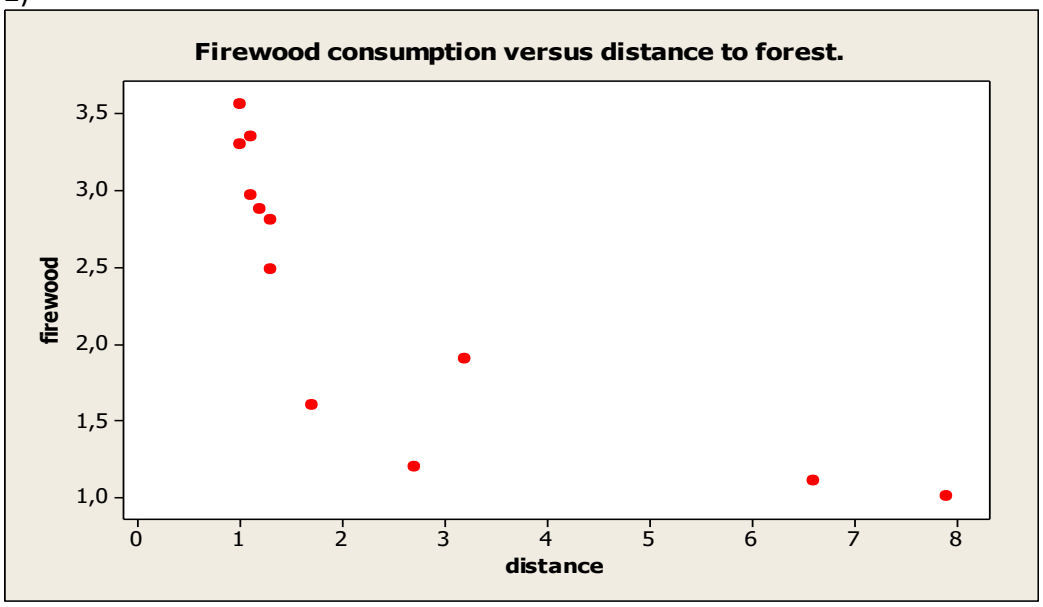
Regression Analysis: firewood versus income

```
The regression equation is
firewood = 0,791 + 0,000312 income

Predictor       Coef    SE Coef     T      P
Constant      0,7914     0,7319  1,08  0,305
income     0,0003124  0,0001393  2,24  0,049

S = 0,802704   R-Sq = 33,5%   R-Sq(adj) = 26,8%

Analysis of Variance

Source           DF      SS      MS     F      P
Regression        1  3,2404  3,2404  5,03  0,049
Residual Error   10  6,4433  0,6443
Total            11  9,6838

Unusual Observations

Obs  income  firewood    Fit  SE Fit  Residual  St Resid
  5    6100     1,010  2,697   0,279    -1,687     -2,24R

R denotes an observation with a large standardized residual.

Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI          95% PI
  1  2,354   0,232  (1,837; 2,870)  (0,492; 4,215)

Values of Predictors for New Observations

New
Obs  income
  1    5000
```

We assume : $Y_i=\beta_0+\beta_1 x_i+e_i$ i= 1,.....,12 The error terms are independent N(0,σ)
Y= Firewood consumption x= Income.
We test H_0 : β_1= 0 against H_1: $\beta_1 \neq 0$ P-value = 0.049 < 0.05 and we reject H_0 at 5% level.
The coefficient of determination is 33.5% , and we evaluate this as medium. Income explains 33.5% of the variation in the firewood consumption.
Interpretation of the fitted model: If income = 0 then we estimate the firewood consumption by 0.791 m^3.
If income increases by 1 then we estimate the firewood consumption to increase by 0.000312 m^3. We predict the firewood consumption for a family having an income of 5000 Tsh by :

$$\hat{Y} = 0.791+0.000312 \cdot 5000=2.35$$

2)



The relationship seems to be a curve.

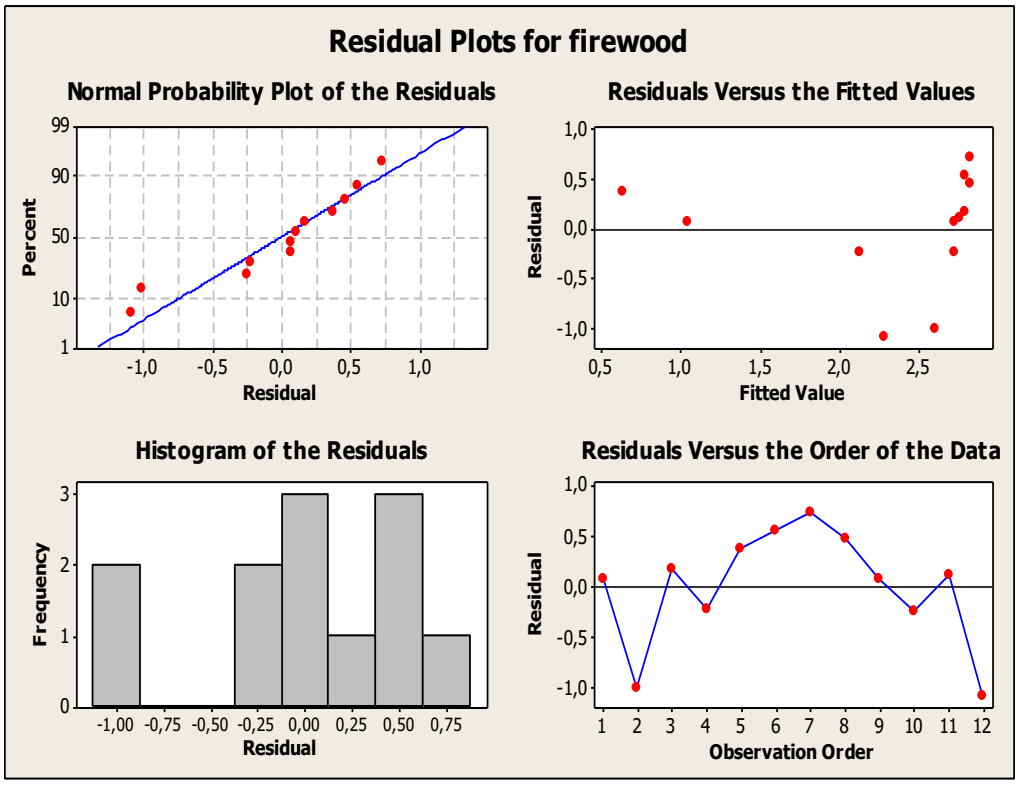
Correlations: firewood; distance

```
Pearson correlation of firewood and distance = -0,795
P-Value = 0,002
```

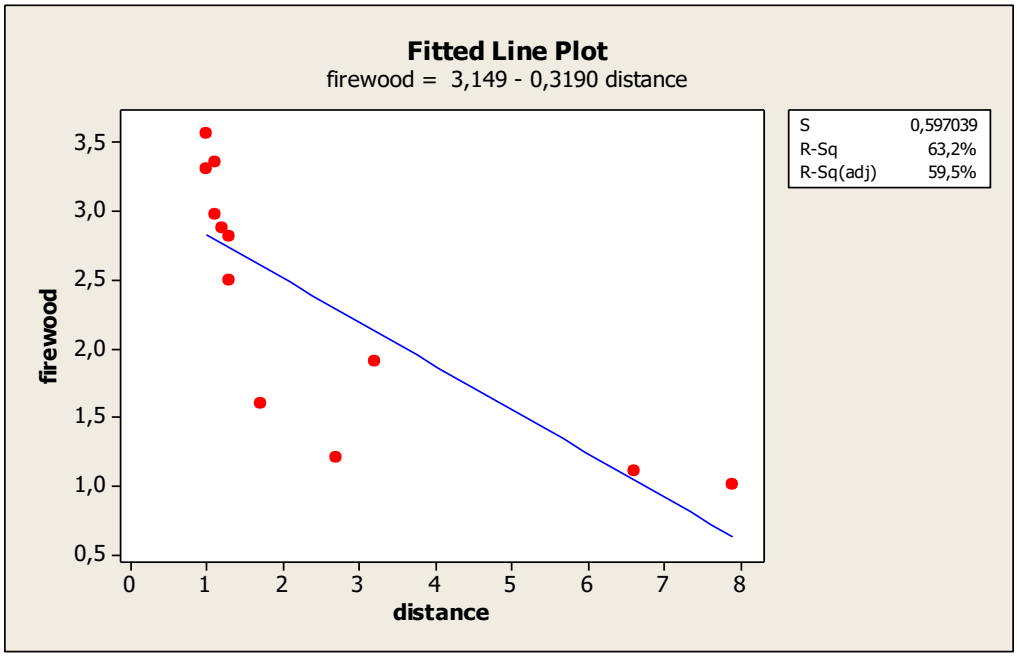
Regression Analysis: firewood versus distance

```
The regression equation is
firewood = 3,15 - 0,319 distance

Predictor      Coef  SE Coef      T      P
Constant     3,1485   0,2588  12,16  0,000
distance   -0,31901  0,07699  -4,14  0,002
S = 0,597039   R-Sq = 63,2%   R-Sq(adj) = 59,5%

Analysis of Variance

Source          DF      SS      MS      F      P
```

```
Regression       1   6,1192   6,1192   17,17   0,002
Residual Error  10   3,5646   0,3565
Total           11   9,6838

Unusual Observations

Obs  distance  firewood     Fit  SE Fit  Residual  St Resid
  5      7,90     1,010   0,628   0,449     0,382      0,97 X

X denotes an observation whose X value gives it large influence.

Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI           95% PI
  1   2,128   0,180   (1,726; 2,530)   (0,738; 3,517)

Values of Predictors for New Observations

New
Obs  distance
  1      3,20
```



We can see a pattern in the residual plot. This means: the linear model has a poor fit



We predict the firewood consumption for a family having distance to forest : 3.2 km. by :

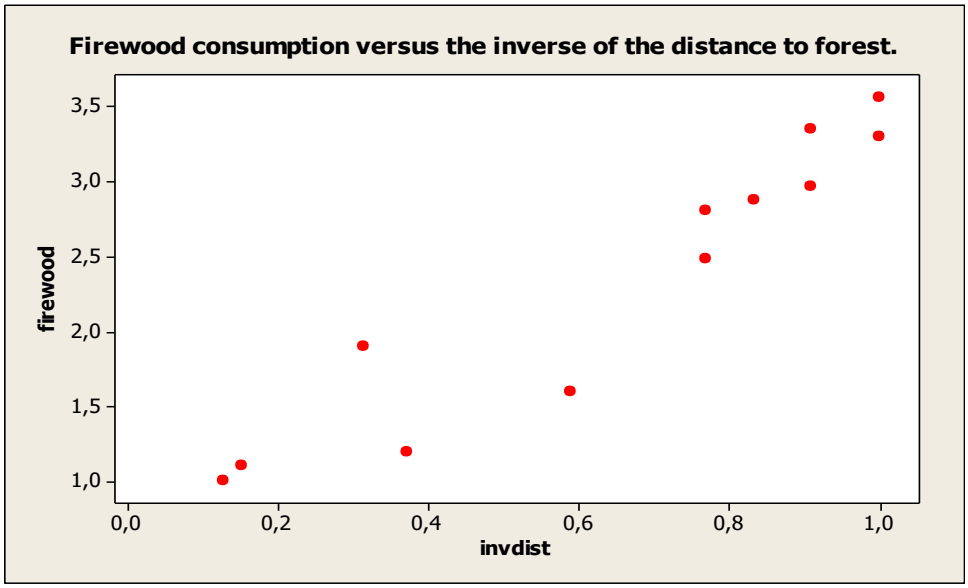
$\hat{Y}$ = 3.15 - 0.319 · **3.2=2.13**

3)

Data Display

Row	firewood	income	distance	famsize	invdist
1	2,81	7800	1,3	8	0,76923
2	1,60	3200	1,7	4	0,58824
3	2,97	5900	1,1	6	0,90909
4	1,90	4100	3,2	4	0,31250
5	1,01	6100	7,9	7	0,12658
6	3,35	6700	1,1	7	0,90909
7	3,56	4900	1,0	5	1,00000
8	3,30	5400	1,0	6	1,00000
9	1,11	3200	6,6	3	0,15152
10	2,49	3700	1,3	5	0,76923
11	2,88	6700	1,2	8	0,83333
12	1,20	2100	2,7	4	0,37037

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"



Correlations: firewood; invdist

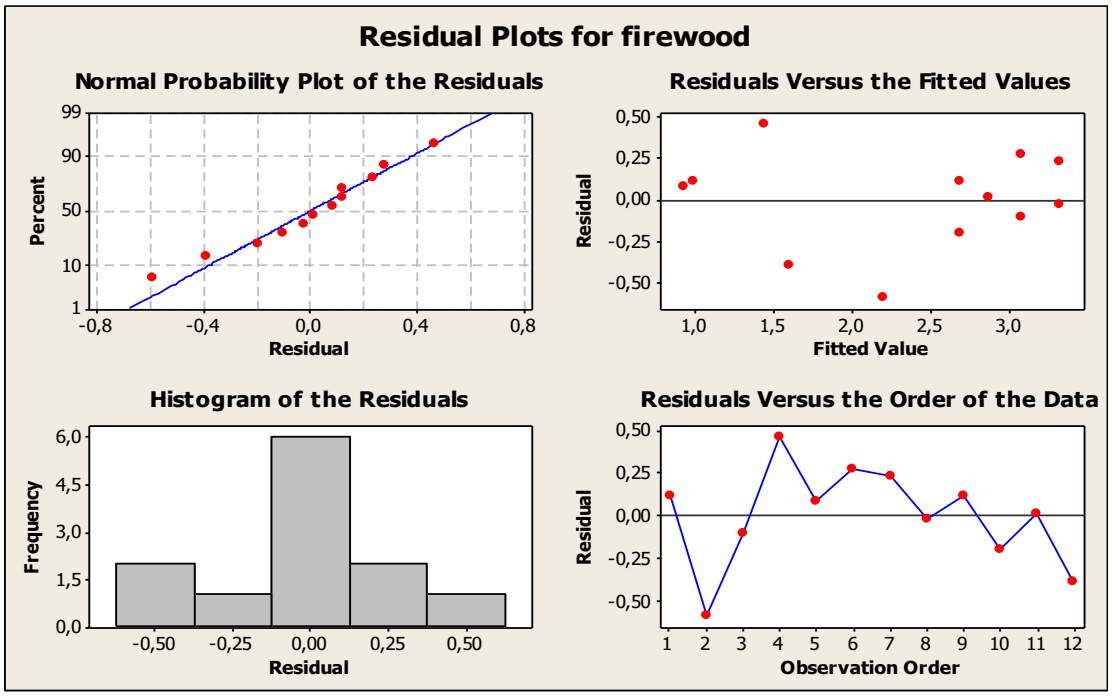
```
Pearson correlation of firewood and invdist = 0,950
P-Value = 0,000
```

4)

Regression Analysis: firewood versus invdist

```
The regression equation is
firewood = 0,577 + 2,75 invdist

Predictor     Coef  SE Coef      T      P
Constant   0,5773   0,2040   2,83  0,018
invdist    2,7460   0,2850   9,64  0,000

S = 0,306850   R-Sq = 90,3%   R-Sq(adj) = 89,3%

Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  8,7422  8,7422  92,85  0,000
Residual Error   10  0,9416  0,0942
Total            11  9,6838

Unusual Observations

Obs  invdist  firewood     Fit  SE Fit  Residual  St Resid
  2     0,59    1,6000  2,1926  0,0900   -0,5926     -2,02R

R denotes an observation with a large standardized residual.

Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI            95% PI
  1  1,4355  0,1297  (1,1465; 1,7245)  (0,6932; 2,1777)

Values of Predictors for New Observations

New
```

```
Obs  invdist
  1    0,313
```



Here we have a residual plot showing random errors. From the regression analysis we have R-sq = 90.3% .

This indicates that we have a much better model now, compared to the model in question 2.

5)

If distance = 3.2 then 1/distance = 0.3125.

We predict the firewood consumption for a family having distance to forest : 3.2 km. by :

$\hat{Y} = 0.577 + 2.75 \cdot$ **0.3125=1.44 m³**

In question 2 we predicted Y by: $\hat{Y}$ = **2.13 m³**

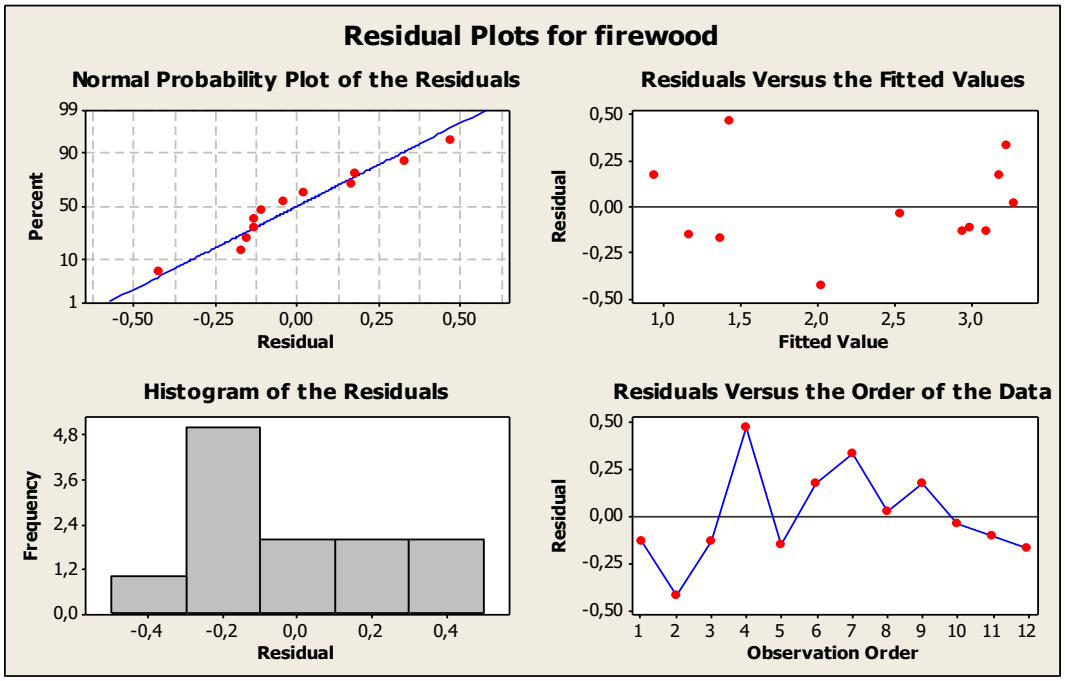
6)

We assume the model $Y_i=\beta_0+\beta_1 x_{i1}+ \beta_2 x_{i2} +e_i$ i= 1,.....,12 The error terms are independent N(0,σ). Y= Firewood consumption x_1= Income x_2= 1/distance.

Regression Analysis: firewood versus income; invdist

```
The regression equation is
firewood = 0,235 + 0,000100 income + 2,50 invdist

Predictor         Coef      SE Coef      T       P
Constant        0,2351       0,2580   0,91   0,386
income     0,00010017   0,00005342   1,87   0,094
invdist         2,5028       0,2859   8,75   0,000

S = 0,274288   R-Sq = 93,0%   R-Sq(adj) = 91,5%

Analysis of Variance

Source           DF       SS       MS       F       P
Regression        2   9,0067   4,5033   59,86   0,000
```

```
Residual Error    9  0,6771  0,0752
Total            11  9,6838

Source    DF  Seq SS
income     1  3,2404
invdist    1  5,7662

Predicted Values for New Observations

New
Obs      Fit  SE Fit        95% CI             95% PI
  1  1,5180  0,1240  (1,2375; 1,7985)  (0,8371; 2,1990)

Values of Predictors for New Observations

New
Obs  income  invdist
  1    5000    0,313
```



We predicted Y by : $\hat{Y} = 0.235 + 0.000100 \cdot 5000 + 2.50 \cdot 0.3125 =$ **1.52** This is almost identical to $\hat{Y}$ in question 5 , but different from $\hat{Y}$ in question 2.

7)

We test H_0 : $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$, and assume $\beta_2 \neq 0$. P-value = 0.094 > 0.05 and we retain H_0 at 5% level of significance.

We test H_0 : $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$, and assume $\beta_1 \neq 0$. P-value = 0.000 < 0.05 and we reject H_0 at 5% level of significance. This means that we can remove income from our model.

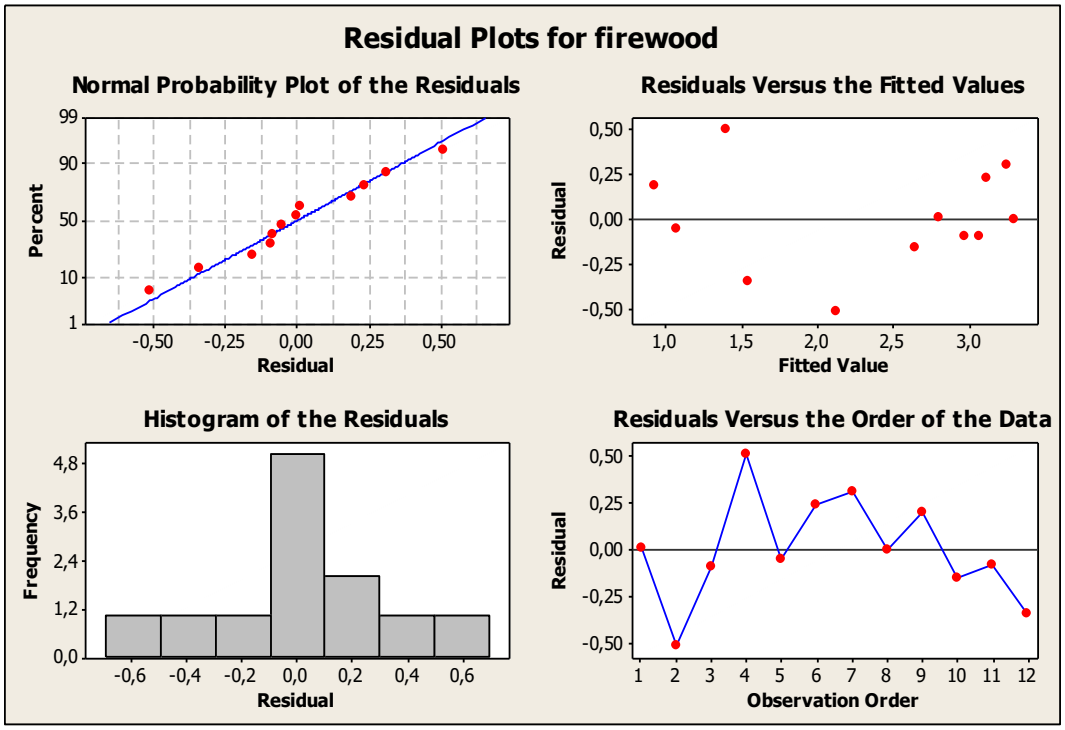
We will now use the model: $Y_i = \beta_0 + \beta_1 x_i + e_i$ i= 1,…..,12 The error terms are independent N(0,σ). Y= Firewood consumption x= 1/distance.

8)

We assume the model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ i= 1,…….,12 The error terms are independent N(0,σ). Y= Firewood consumption x_1= 1/distance x_2=family size.

Regression Analysis: firewood versus invdist; famsize

```
The regression equation is
firewood = 0,364 + 2,62 invdist + 0,0523 famsize


Predictor     Coef  SE Coef     T      P
Constant    0,3638   0,3296  1,10  0,298
invdist     2,6242   0,3243  8,09  0,000
famsize    0,05233  0,06279  0,83  0,426

S = 0,311649   R-Sq = 91,0%   R-Sq(adj) = 89,0%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        2  8,8096  4,4048  45,35  0,000
Residual Error    9  0,8741  0,0971
Total            11  9,6838


Source   DF  Seq SS
invdist   1  8,7422
famsize   1  0,0674


Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI            95% PI
  1  1,4978  0,1515  (1,1551; 1,8404)  (0,7139; 2,2816)


Values of Predictors for New Observations

New
Obs  invdist  famsize
  1    0,313     6,00
```



We use the model $Y_i=\beta_0+\beta_1 x_{i1}+ \beta_2 x_{i2} +e_i$ to predict Y when $x_1= 0.3125$ and $x_2=6$:

$\hat{Y} = 0.364 + 2.62 \cdot 0.3125 + 0.0523 \cdot 6 = 1.50$

We test $H_0 : \beta_1= 0$ against $H_1: \beta_1 \neq 0$, and assume $\beta_2 \neq 0$. P-value = 0.000 < 0.05 and we reject H_0 at 5% level of significance.

We test $H_0 : \beta_2= 0$ against $H_1: \beta_2 \neq 0$, and assume $\beta_1 \neq 0$. P-value = 0.426 > 0.05 and we retain H_0 at 5% level of significance.

This means that we can remove family size from our model. We will now use the model: $Y_i=\beta_0+\beta x_i +e_i$ i= 1,.....,12 The error terms are independent $N(0,\sigma)$

Y= Firewood consumption x= 1/distance.

Minitab lecture 6.

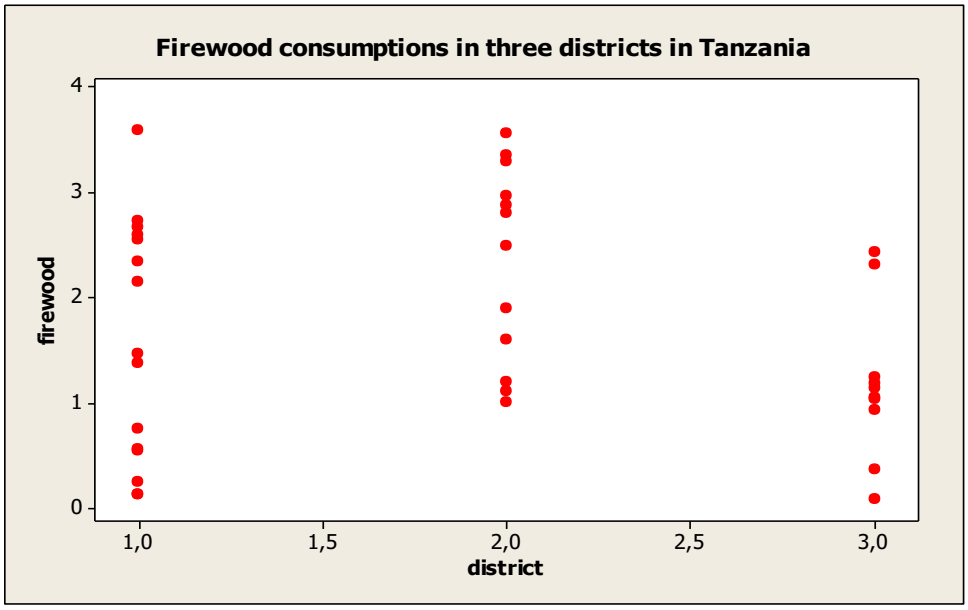
1)

Data Display

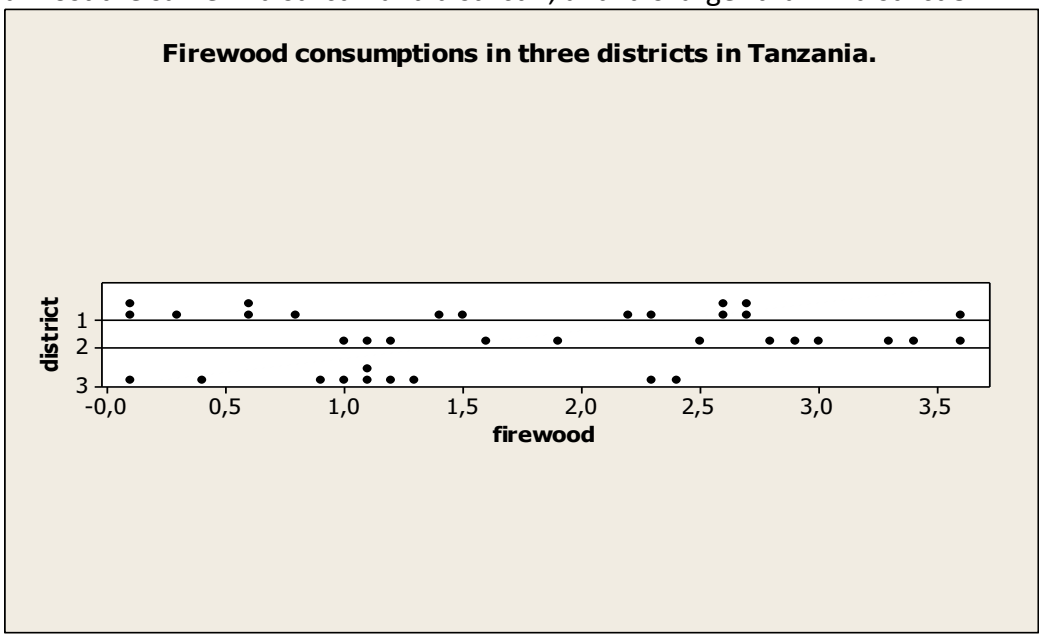
Row	urban	rural	Kigoma	firewood	district	RESI1	FITS1
1	0,55	2,81	0,37	0,55	1	-1,04267	1,59267
2	2,16	1,60	1,19	2,16	1	0,56733	1,59267
3	2,60	2,97	1,14	2,60	1	1,00733	1,59267
4	2,34	1,90	0,93	2,34	1	0,74733	1,59267
5	2,68	1,01	1,25	2,68	1	1,08733	1,59267
6	1,38	3,35	0,09	1,38	1	-0,21267	1,59267
7	0,13	3,56	2,44	0,13	1	-1,46267	1,59267
8	0,25	3,30	2,31	0,25	1	-1,34267	1,59267
9	2,73	1,11	1,06	2,73	1	1,13733	1,59267
10	0,57	2,49	1,04	0,57	1	-1,02267	1,59267
11	3,59	2,88		3,59	1	1,99733	1,59267
12	1,47	1,20		1,47	1	-0,12267	1,59267
13	0,14			0,14	1	-1,45267	1,59267
14	0,75			0,75	1	-0,84267	1,59267
15	2,55			2,55	1	0,95733	1,59267
16				2,81	2	0,46167	2,34833
17				1,60	2	-0,74833	2,34833
18				2,97	2	0,62167	2,34833
19				1,90	2	-0,44833	2,34833
20				1,01	2	-1,33833	2,34833
21				3,35	2	1,00167	2,34833
22				3,56	2	1,21167	2,34833
23				3,30	2	0,95167	2,34833
24				1,11	2	-1,23833	2,34833
25				2,49	2	0,14167	2,34833
26				2,88	2	0,53167	2,34833
27				1,20	2	-1,14833	2,34833
28				0,37	3	-0,81200	1,18200
29				1,19	3	0,00800	1,18200
30				1,14	3	-0,04200	1,18200
31				0,93	3	-0,25200	1,18200
32				1,25	3	0,06800	1,18200
33				0,09	3	-1,09200	1,18200
34				2,44	3	1,25800	1,18200
35				2,31	3	1,12800	1,18200
36				1,06	3	-0,12200	1,18200
37				1,04	3	-0,14200	1,18200

2)

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"



The plot tells me that the firewood consumption has a larger sample mean in district 2 compared to district 1 and district 3. The minimum consumption of firewood is larger in district 2 compared to the two other districts. The maximum consumption of firewood is almost the same in district 1 and district 2, and it is larger than in district 3.



3)

Descriptive Statistics: urban; rural; Kigoma

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
urban	15	0	1,593	0,295	1,141	0,130	0,550	1,470	2,600	3,590
rural	12	0	2,348	0,271	0,938	1,010	1,300	2,650	3,218	3,560
Kigoma	10	0	1,182	0,231	0,731	0,0900	0,790	1,100	1,515	2,440

District 3 has a smaller sample mean than the two other districts.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

4) I believe there are differences in the population standard deviations, but the differences are so small that we can look at them as equal. No sample standard deviation is twice one of the other standard deviations.
5) Let Y_{ij} = firewood consumption in district i , observation number j, i = 1,2,3 and j = 1, . ,n_i. We assume $Y_{ij} = \mu_i + e_{ij}$ and e_{ij} is N(0,σ) , all the error-terms are independent. This is the same as assuming : Y_{ij} is N(μ_i,σ). This is a one way design because we have just one systematic effect in the model , the effect of district.
6) Yes, I believe so. The observations in one group are very spread out and we have many almost identical observations in the three groups.
7 and 8)We test H_0: $\mu_1 = \mu_2 = \mu_3$ against H_1: At least one μ_i is different from the other population means. We reject H_0 at a 5% level of significance because p-value=0.026<0.05.

One-way ANOVA: firewood versus district

```
Source     DF      SS     MS     F      P
district    2   7,874  3,937  4,09  0,026
Error      34  32,728  0,963
Total      36  40,602

S = 0,9811   R-Sq = 19,39%   R-Sq(adj) = 14,65%

                           Individual 95% CIs For Mean Based on
                           Pooled StDev
Level   N    Mean   StDev   -+---------+---------+---------+--------
1      15  1,5927  1,1412            (--------*-------)
2      12  2,3483  0,9383                      (--------*---------)
3      10  1,1820  0,7312   (----------*---------)
                            -+---------+---------+---------+--------
                           0,60      1,20      1,80      2,40
Pooled StDev = 0,9811

Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons among Levels of district

Simultaneous confidence level = 88,01%

district = 1 subtracted from:

district    Lower   Center   Upper  -------+---------+---------+---------+--
2         -0,0166   0,7557  1,5279                   (-----*------)
3         -1,2247  -0,4107  0,4033         (------*-----)
                            -------+---------+---------+---------+--
                                -1,2       0,0       1,2       2,4

district = 2 subtracted from:

district    Lower   Center    Upper  -------+---------+---------+---------+--
3         -2,0201  -1,1663  -0,3126  (------*------)
                                     -------+---------+---------+---------+--
                                         -1,2       0,0       1,2       2,4
```

One-way ANOVA: firewood versus district

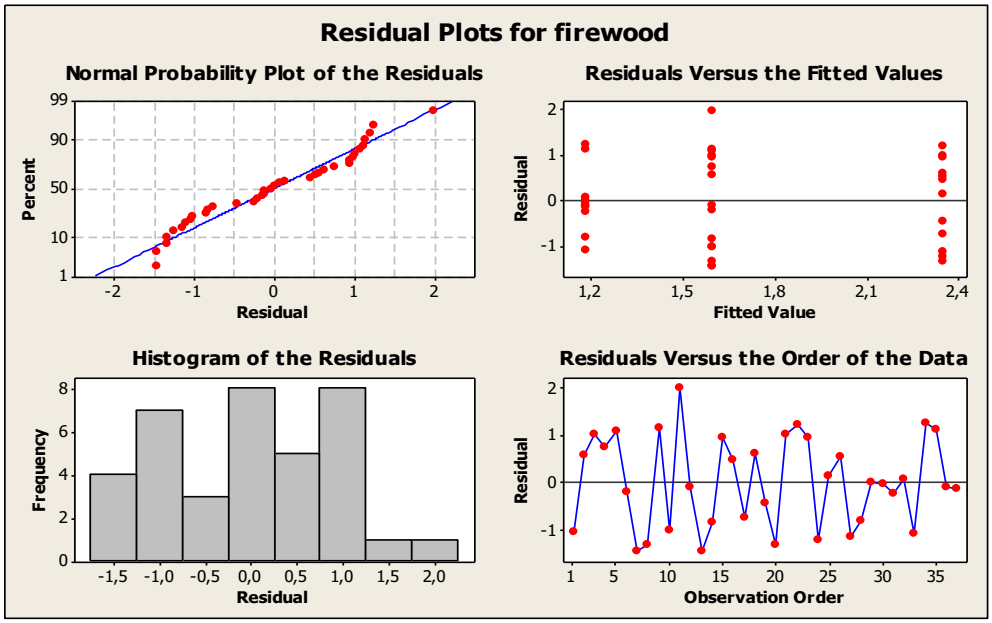
```
Source     DF      SS     MS     F      P
```

```
district   2   7,874  3,937  4,09  0,026
Error     34  32,728  0,963
Total     36  40,602

S = 0,9811   R-Sq = 19,39%   R-Sq(adj) = 14,65%

                           Individual 95% CIs For Mean Based on Pooled StDev
Level    N    Mean   StDev   -+---------+---------+---------+--------
1       15  1,5927  1,1412            (--------*-------)
2       12  2,3483  0,9383                      (--------*---------)
3       10  1,1820  0,7312   (----------*---------)
                             -+---------+---------+---------+--------
                           0,60      1,20      1,80      2,40

Pooled StDev = 0,9811

Fisher 99% Individual Confidence Intervals
All Pairwise Comparisons among Levels of district

Simultaneous confidence level = 97,36%

district = 1 subtracted from:

district    Lower   Center   Upper  ---------+---------+---------+---------+
2         -0,2811   0,7557  1,7924                    (-------*--------)
3         -1,5035  -0,4107  0,6822         (---------*--------)
                                   ---------+---------+---------+---------+
                                         -1,2       0,0       1,2       2,4

district = 2 subtracted from:

district    Lower   Center    Upper  ---------+---------+---------+---------+
3         -2,3125  -1,1663  -0,0202  (--------*---------)
                                    ---------+---------+---------+---------+
                                          -1,2       0,0       1,2       2,4
```

If we reduce the level of confidence, we increase the level of significance α, and the intervals become narrower. If we choose α = 1%, the simultaneous confidence level is 97.36%. The intervals become wider.

9) A fitted value = $\hat{\mu}_i = \bar{y}_{i.}$ and a residual = $y_{ij} - \bar{y}_{i.}$

Residuals vs Fits for firewood



We can see from the plot if the residuals for the three groups are very different. In this plot the residuals for the three groups are similar. It looks as the assumptions about the model are reasonable. The error-terms seem to be independent N(0,σ) with the same σ for the three groups.

10) We can see that there are only small differences within each group, but large differences from one group to another.

The F-value will become large and we can state that there are differences in mean yield for the 4 groups.

11)

One-way ANOVA: yield versus variety

```
Source    DF        SS       MS          F      P
variety    3  129322,1  43107,4  14719,59  0,000
Error     14      41,0      2,9
Total     17  129363,1

S = 1,711   R-Sq = 99,97%   R-Sq(adj) = 99,96%

                            Individual 95% CIs For Mean Based on
                            Pooled StDev
Level      N     Mean  StDev  -------+---------+---------+---------+--
variety1   5   936,40   1,14     *
variety2   4   921,00   1,41  (*
variety3   4   945,50   0,58      (*
variety4   5  1122,80   2,68                                      *
                            -------+---------+---------+---------+--
                                 960      1020      1080      1140

Pooled StDev = 1,71


Fisher 99% Individual Confidence Intervals
All Pairwise Comparisons among Levels of variety

Simultaneous confidence level = 95,61%
```

```
variety = variety1 subtracted from:

variety    Lower  Center   Upper  -------+---------+---------+---------+--
variety2  -18,82  -15,40  -11,98                (*
variety3    5,68    9,10   12,52                  (*
variety4  183,18  186,40  189,62                                (*
                                  -------+---------+---------+---------+--
                                      -120         0       120       240


variety = variety2 subtracted from:

variety    Lower  Center   Upper  -------+---------+---------+---------+--
variety3   20,90   24,50   28,10                   *
variety4  198,38  201,80  205,22                              *
                                  -------+---------+---------+---------+--
                                      -120         0       120       240


variety = variety3 subtracted from:

variety    Lower  Center   Upper  -------+---------+---------+---------+--
variety4  173,88  177,30  180,72                            (*
                                  -------+---------+---------+---------+--
                                      -120         0       120       240
```

We choose the individual level of significance as 1%. Then simultaneous confidence level is 95.61%. The intervals are narrow because we have S-pooled = 1.71 and this is very small compared to the differences between the sample means for the 4 groups. No interval contains 0. This means: We can state that all population means are different.

Minitab lecture 7.

1)

Data Display

Row	sit	Medac	PNK_50	Smiths	Biolite	Growth	Fertil	Sites
1	1	8,0	9,5	7,0	8,5	8,0	1	1
2	2	7,5	8,5	7,0	8,0	7,5	1	2
3	3	9,0	8,5	9,5	7,5	9,0	1	3
4	4	7,5	8,0	7,5	8,0	7,5	1	4
5	5	7,0	7,5	9,0	9,5	7,0	1	5
6						9,5	2	1
7						8,5	2	2
8						8,5	2	3
9						8,0	2	4
10						7,5	2	5
11						7,0	3	1
12						7,0	3	2
13						9,5	3	3
14						7,5	3	4
15						9,0	3	5
16						8,5	4	1
17						8,0	4	2
18						7,5	4	3
19						8,0	4	4
20						9,5	4	5

2)

Descriptive Statistics: Medac; PNK_50; Smiths; Biolite

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Medac	5	0	7,800	0,339	0,758	7,000	7,250	7,500	8,500	9,000
PNK_50	5	0	8,400	0,332	0,742	7,500	7,750	8,500	9,000	9,500
Smiths	5	0	8,000	0,524	1,173	7,000	7,000	7,500	9,250	9,500
Biolite	5	0	8,300	0,339	0,758	7,500	7,750	8,000	9,000	9,500

3)

Let Y_{ij} = observed annual growth of tree given fertilizer i, site j. We assume $Y_{ij} = \mu_i + e_{ij}$ and e_{ij} is $N(0,\sigma)$, all error terms are independent.

We test the hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against H_1: at least two fertilizers have different effects on mean growth of redbud trees.

One-way ANOVA: Growth versus Fertil

Source	DF	SS	MS	F	P
Fertil	3	1,138	0,379	0,49	0,692
Error	16	12,300	0,769		
Total	19	13,438			

S = 0,8768 R-Sq = 8,47% R-Sq(adj) = 0,00%

Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
1	5	7,8000	0,7583	(----------*----------)
2	5	8,4000	0,7416	(----------*----------)
3	5	8,0000	1,1726	(----------*----------)
4	5	8,3000	0,7583	(----------*----------)

```
----+---------+---------+---------+-----
  7,20      7,80      8,40      9,00

Pooled StDev = 0,8768
```

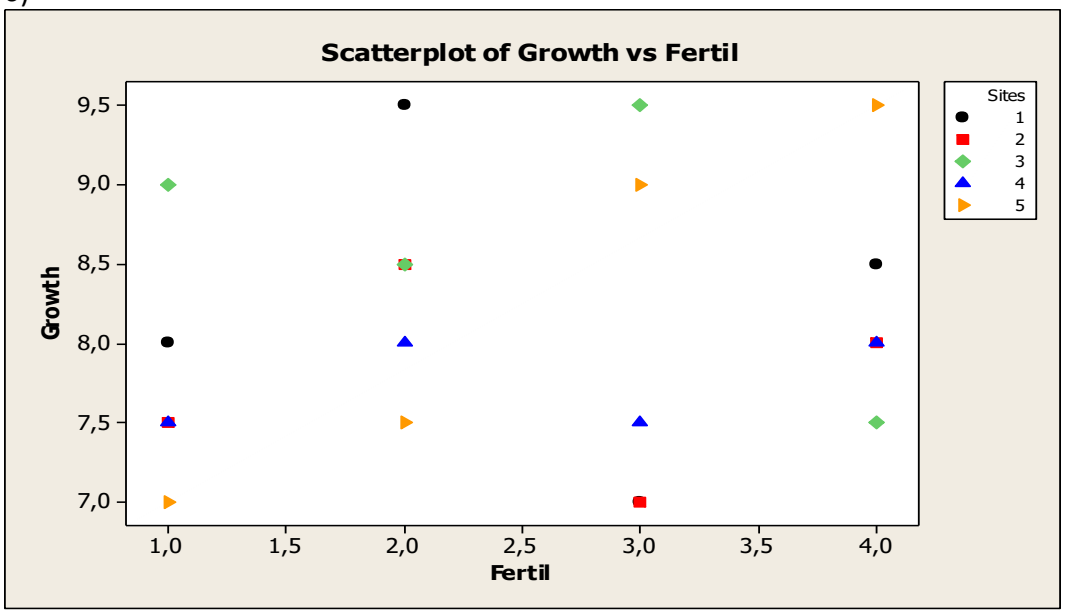
We retain H_0 at a 5% level of significance because the p-value = 0.692 > 0.05.

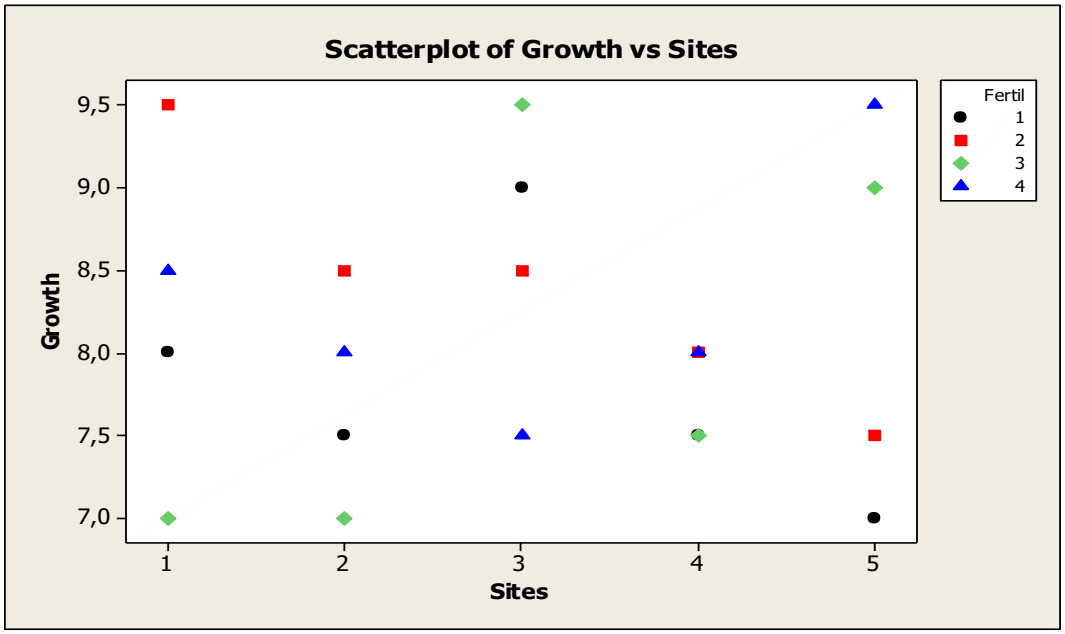
5)

Descriptive Statistics: Growth

Variable	Sites	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Growth	1	4	0	8,250	0,520	1,041	7,000	7,250	8,250	9,250	9,500
	2	4	0	7,750	0,323	0,645	7,000	7,125	7,750	8,375	8,500
	3	4	0	8,625	0,427	0,854	7,500	7,750	8,750	9,375	9,500
	4	4	0	7,750	0,144	0,289	7,500	7,500	7,750	8,000	8,000
	5	4	0	8,250	0,595	1,190	7,000	7,125	8,250	9,375	9,500

6)





This first plot gives the best information of the differences between the fertilizers.

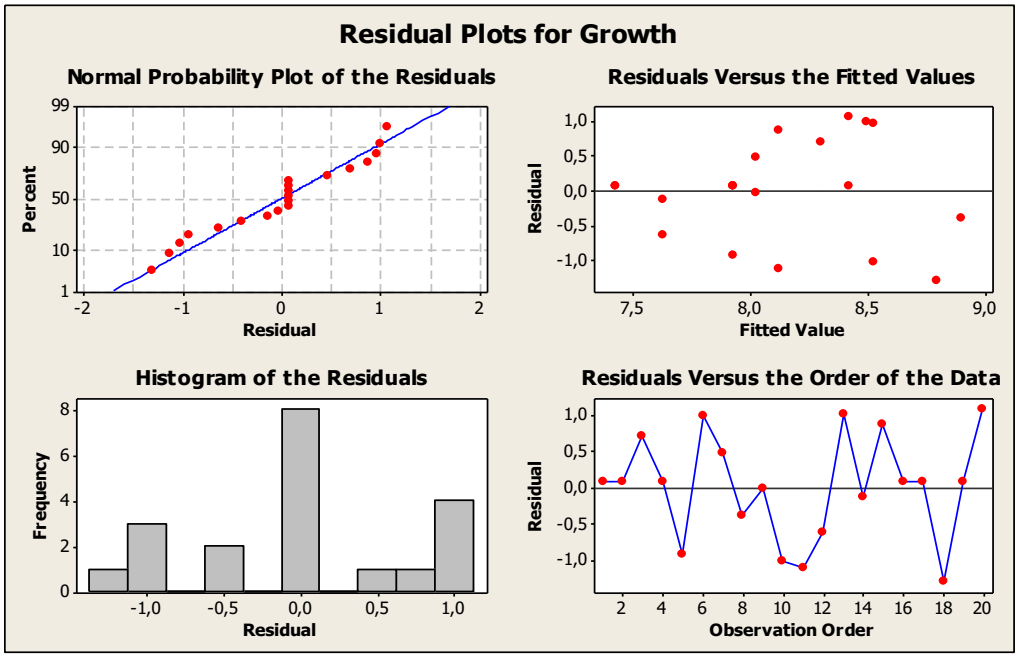
7)

Let Y_{ij} = observed annual growth of tree given fertilizer i, site j. We assume $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ and e_{ij} is $N(0,\sigma)$, all error terms are independent. α_i = effect of fertilizer i and β_j = effect of site j. We test the hypotheses: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ against H_1 : at least two fertilizers have different effects on mean growth of redbud trees.

General Linear Model: Growth versus Fertil; Sites

```
Factor  Type   Levels  Values
Fertil  fixed       4  1; 2; 3; 4
Sites   fixed       5  1; 2; 3; 4; 5

Analysis of Variance for Growth, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS     F      P
Fertil   3   1,1375   1,1375  0,3792  0,45  0,720
Sites    4   2,2500   2,2500  0,5625  0,67  0,624
Error   12  10,0500  10,0500  0,8375
Total   19  13,4375

S = 0,915150   R-Sq = 25,21%   R-Sq(adj) = 0,00%

Least Squares Means for Growth

Fertil   Mean  SE Mean
1       7,800   0,4093
2       8,400   0,4093
3       8,000   0,4093
4       8,300   0,4093
```

The p-value = 0.72 > 0.05 and we retain H_0.



8) We test $H_0 : \alpha_3 = \alpha_4$ against $H_1 : \alpha_3 < \alpha_4$. We calculate :

$$t = \frac{\bar{y}_{3.} - \bar{y}_{4.}}{S\sqrt{\frac{2}{b}}} = \frac{8 - 8.3}{0.91515\sqrt{\frac{2}{5}}} = -0.518321$$

We find $-t_{0.05,12} = -1.782 < t$ and we retain H_0.

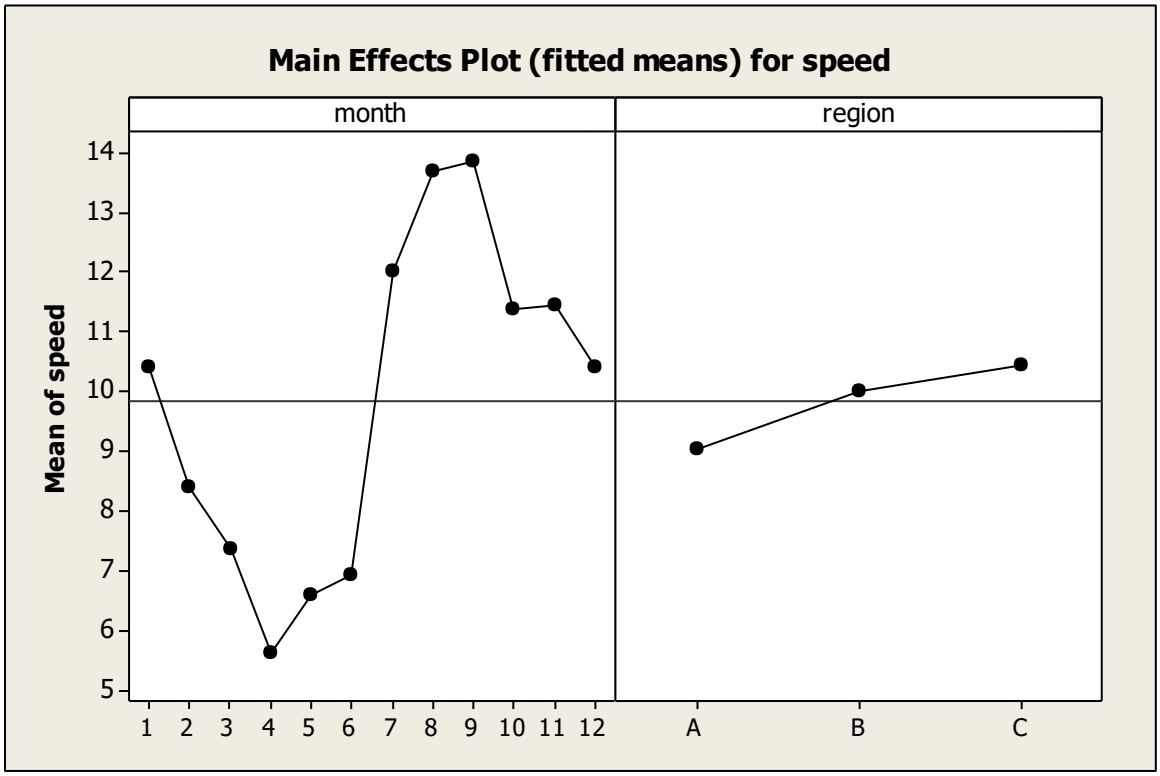
Exam problem 2 December 9. 2002.

Let Y_{ij} = the mean wind speed measured in month j, region i. We assume $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ and e_{ij} is $N(0,\sigma)$, all error terms are independent. α_i = effect of region i and β_j = effect of month j.

General Linear Model: speed versus region; month

```
Factor  Type   Levels  Values
region  fixed       3  A; B; C
month   fixed      12  1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12


Analysis of Variance for speed, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS  Adj MS      F      P
region   2   12,365   12,365   6,183   5,35  0,013
month   11  258,328  258,328  23,484  20,33  0,000
Error   22   25,415   25,415   1,155
Total   35  296,108


S = 1,07481   R-Sq = 91,42%   R-Sq(adj) = 86,35%


Unusual Observations for speed
```

```
Obs    speed      Fit  SE Fit  Residual  St Resid
  2   5,7000   7,6083  0,6703   -1,9083     -2,27 R
 34  10,2000  11,9750  0,6703   -1,7750     -2,11 R

R denotes an observation with a large standardized residual.


Bonferroni 95,0% Simultaneous Confidence Intervals
Response Variable speed
All Pairwise Comparisons among Levels of region
region = A  subtracted from:

region    Lower  Center  Upper  -------+---------+---------+---------
B       -0,1620  0,9750  2,112       (-----------*----------)
C        0,2630  1,4000  2,537            (----------*----------)
                                -------+---------+---------+---------
                                     0,0       1,0       2,0


region = B  subtracted from:

region    Lower  Center  Upper  -------+---------+---------+---------
C       -0,7120  0,4250  1,562  (----------*-----------)
                                -------+---------+---------+---------
                                     0,0       1,0       2,0


Bonferroni Simultaneous Tests
Response Variable speed
All Pairwise Comparisons among Levels of region
region = A  subtracted from:

        Difference       SE of           Adjusted
region    of Means  Difference  T-Value   P-Value
B           0,9750      0,4388    2,222    0,1106
C           1,4000      0,4388    3,191    0,0127


region = B  subtracted from:

        Difference       SE of           Adjusted
region    of Means  Difference  T-Value   P-Value
C           0,4250      0,4388   0,9686     1,000
```



We test the hypotheses: $H_0 : \alpha_A = \alpha_B = \alpha_C = 0$ against
H_1 : at least two regions are different with respect to mean wind speed.
The p-value = 0.13 < 0.05 and we reject H_0. At least two regions are different with respect to mean wind speed.

b) We test: $H_0 : \alpha_A = \alpha_B$ against $H_1 : \alpha_A < \alpha_B$

$t = \frac{\bar{y}_{A.} - \bar{y}_{B.}}{S\sqrt{\frac{2}{b}}}$ = -2.22< $-t_{0.025,22}$= -2.074 and we reject H_0.

We test: $H_0 : \alpha_B = \alpha_C$ against $H_1 : \alpha_B < \alpha_C$

$t = \frac{\bar{y}_{B.} - \bar{y}_{C.}}{S\sqrt{\frac{2}{b}}}$ = -0.97> $-t_{0.025,22}$= -2.074 and we retain H_0.

We can't prove that region B and C are different with respect to mean wind speed. Region B should be preferred.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

Solution to Minitab lecture 8.

1)

Data Display

Row	deaths	survive
1	1	22
2	3	40
3	4	22
4	6	8
5	3	1

2)

Chi-Square Test: deaths; survive

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
```

	deaths	survive	Total
1	1	22	23
	3,55	19,45	
	1,836	0,336	
2	3	40	43
	6,65	36,35	
	2,000	0,366	
3	4	22	26
	4,02	21,98	
	0,000	0,000	
4	6	8	14
	2,16	11,84	
	6,802	1,243	
5	3	1	4
	0,62	3,38	
	9,177	1,678	
Total	17	93	110

```
Chi-Sq = 23,437; DF = 4
WARNING: 1 cells with expected counts less than 1. Chi-Square approximation
     probably invalid.
5 cells with expected counts less than 5.
```

3)

There are cells with expected counts less than 5 in row 1, 3, 4 and 5. We combine row 1 and 2. We also combine row 3, 4 and 5. Now all expected counts are more than 5 in the table.

Data Display

Row	dead	sur
1	4	62
2	13	31

Chi-Square Test: dead; sur

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

        dead    sur  Total
    1      4     62     66
       10,20  55,80
       3,769  0,689

    2     13     31     44
        6,80  37,20
       5,653  1,033

Total     17     93    110

Chi-Sq = 11,144; DF = 1; P-Value = 0,001
```

We test : H_0 : $p_{ij} = p_{i.} \cdot p_{.j}$ against H_1: $p_{ij} \neq p_{i.} \cdot p_{.j}$ i = 1 = death, i = 2 = survive,
j = 1 = age [0,44] , j = 2 = age [45, $\rightarrow$).

p_{1j} = P(to die if the age of the patient is in interval number j)

p_{2j} = P(to survive if the age of the patient is in interval number j)

$p_{.j} = p_{1j} + p_{2j}$ j = 1, 2 $p_{i.} = p_{i1} + p_{i2}$ i = 1 , 2

4)

This is a test of independence. The p-value is 0.001 < 0.05 and we reject H_0. We find dependence between the outcome of the SARS virus and age.

5)

Let X = the number of tram passengers without a valid ticket out of 70.
X has a Binomial distribution, n = 70 p = P("no valid ticket")
We test : H_0: p = 0.1 against H_1: p> 0.1

Test and CI for One Proportion

```
Test of p = 0,1 vs p > 0,1

                                                    Exact
Sample      X      N  Sample p  95,0% Lower Bound  P-Value
1          11     70  0,157143           0,090698    0,087
```

Test and CI for One Proportion

```
Test of p = 0,1 vs p > 0,1

Sample      X      N  Sample p  95,0% Lower Bound  Z-Value  P-Value
1          11     70  0,157143           0,085594     1,59    0,056
```

Both the exact test and the z – test give p-value > 0.05. We retain H_0 at a 5% level of significance. The p – value for the exact test is 0.087 and for the z-test it is 0.056.

Project number: 586304-EPP-1-2017-1-BA-EPPKA2-CBHE-JP "This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein"

6)

Test and CI for One Proportion

```
Test of p = 0,1 vs p not = 0,1

                                                          Exact
Sample      X      N  Sample p          90,0% CI       P-Value
1          11     70  0,157143  (0,090698; 0,246679)    0,158
```

Test and CI for One Proportion

```
Test of p = 0,1 vs p not = 0,1

Sample      X      N  Sample p          90,0% CI       Z-Value  P-Value
1          11     70  0,157143  (0,085594; 0,228692)      1,59    0,111
```

7) Let Y = the number of tram passengers without a valid ticket out of 80.

Y has a Binomial distribution, n = 80 p_2 = P("no valid ticket")

We test : H_0: p = p_2 against H_1: p > p_2

Test and CI for Two Proportions

```
Sample      X      N  Sample p
1          11     70  0,157143
2          10     80  0,125000

Estimate for p(1) - p(2):  0,0321429
95% lower bound for p(1) - p(2):  -0,0617625
Test for p(1) - p(2) = 0 (vs > 0):  Z = 0,57  P-Value = 0,286
```

The p-value = 0.286 > 0.05 and we retain H_0. We can't say that the proportion of tram passengers in Oslo without a valid ticket decreased from the day of the first to the day of the second inspection.

8)

Test and CI for Two Proportions

```
Sample      X      N  Sample p
1          11     70  0,157143
2          10     80  0,125000

Estimate for p(1) - p(2):  0,0321429
90% CI for p(1) - p(2):  (-0,0617625; 0,126048)
Test for p(1) - p(2) = 0 (vs not = 0):  Z = 0,57  P-Value = 0,571
```

9)

Problem 2. 26/5 – 00

Data Display

```
Row  rich  medium  poor
  1    12       7     3
  2     6       5     5
  3     3      13    22
```

Chi-Square Test: rich; medium; poor

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

         rich  medium   poor  Total
    1      12       7      3     22
         6,08    7,24   8,68
        5,767   0,008  3,721

    2       6       5      5     16
         4,42    5,26   6,32
        0,564   0,013  0,274

    3       3      13     22     38
        10,50   12,50  15,00
        5,357   0,020  3,267

Total      21      25     30     76

Chi-Sq = 18,991; DF = 4; P-Value = 0,001
1 cells with expected counts less than 5.
```

We test H_0: gardening and economic situation are independent, against
H_1: gardening and economic situation are dependent.
The p-value = 0.001<0.05 and we reject H_0. Gardening and economic situation are dependent.

Problem 3.11 in the text-book:

Data Display

Row	died	survived	count	outcome	hospital	condition
1	15	685	15	died	R	good
2	16	584	16	died	C	good
3	75	1425	75	died	R	poor
4	7	93	7	died	C	poor
5			685	survived	R	good
6			584	survived	C	good
7			1425	survived	R	poor
8			93	survived	C	poor

Stat →Tables →Cross Tabulation and Chi-Square
Rows: hospital
Columns: outcome
Layers: condition
Frequencies are in: count
Click at Chi-square.

Tabulated statistics: hospital; outcome; condition

```
Using frequencies in count
```

Results for condition = good

Rows: hospital Columns: outcome

	died	survived	All
C	16	584	600
	14,3	585,7	600,0
	0,20017	0,00489	*
R	15	685	700
	16,7	683,3	700,0
	0,17157	0,00419	*
All	31	1269	1300
	31,0	1269,0	1300,0
	*	*	*

Cell Contents: Count
Expected count
Contribution to Chi-square

Pearson Chi-Square = 0,381; DF = 1; P-Value = 0,537
Likelihood Ratio Chi-Square = 0,379; DF = 1; P-Value = 0,538

Results for condition = poor

Rows: hospital Columns: outcome

	died	survived	All
C	7	93	100
	5,1	94,9	100,0
	0,68598	0,03706	*
R	75	1425	1500
	76,9	1423,1	1500,0
	0,04573	0,00247	*
All	82	1518	1600
	82,0	1518,0	1600,0
	*	*	*

Cell Contents: Count
Expected count
Contribution to Chi-square

Pearson Chi-Square = 0,771; DF = 1; P-Value = 0,380
Likelihood Ratio Chi-Square = 0,701; DF = 1; P-Value = 0,403

Let pC=P(survival) in community hospital, pR=P(survival) in research hospital.
We Test H_0: pC=pR against H_1: pC≠pR
for patients in a good health condition.
The p-value = 0.537> 0.05 and we retain H_0.
****
We Test H_0: pC=pR against H_1: pC≠pR
for patients in a poor health condition.
The p-value = 0.380> 0.05 and we retain H_0.

We have not proven that the probability of survival is different for the two kinds of hospitals.

1.