Research Methodology Statistics

Maha Omair
Teaching Assistant
Department of Statistics, College of science
King Saud University

Why we need statistical data analysis?

Investigations in diverse fields like agriculture, medicine, physics, biology, chemistry etc. require collection of “observations”. Observations are almost always subject to random error. Hence statistical methods have to be employed to collect as well as to analyze the data.

Statistical data analysis

Studying a problem through the use of statistical data analysis usually involves four basic steps:

1. Defining the problem.
2. Collecting the data.
3. Analyzing the data.
4. Conclusions and recommendations.

Defining the problem

An exact definition of the problem is imperative in order to obtain accurate data about it. It is extremely difficult to gather data without a clear definition of the problem.

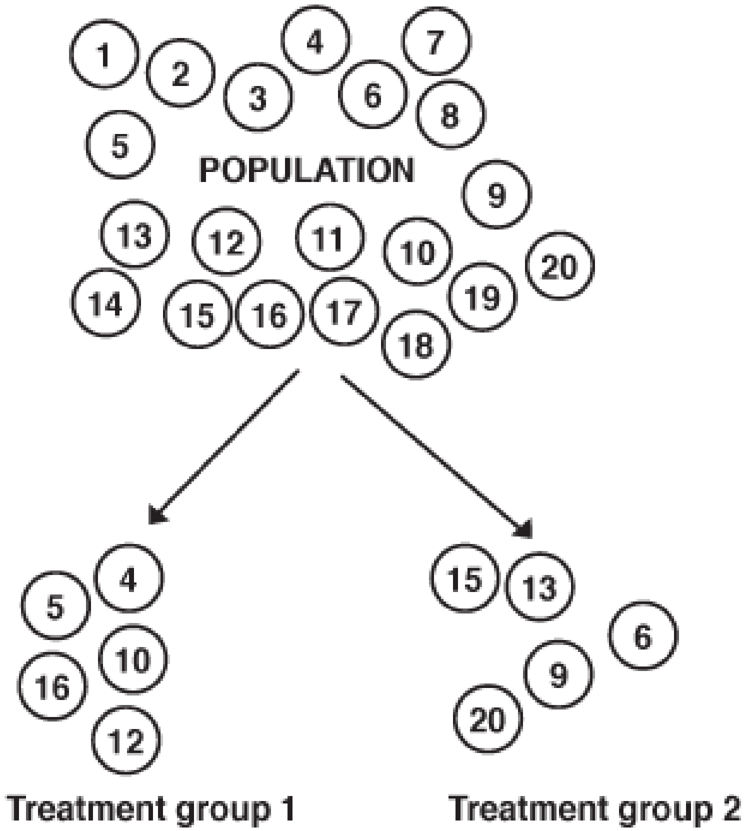
Collecting the data

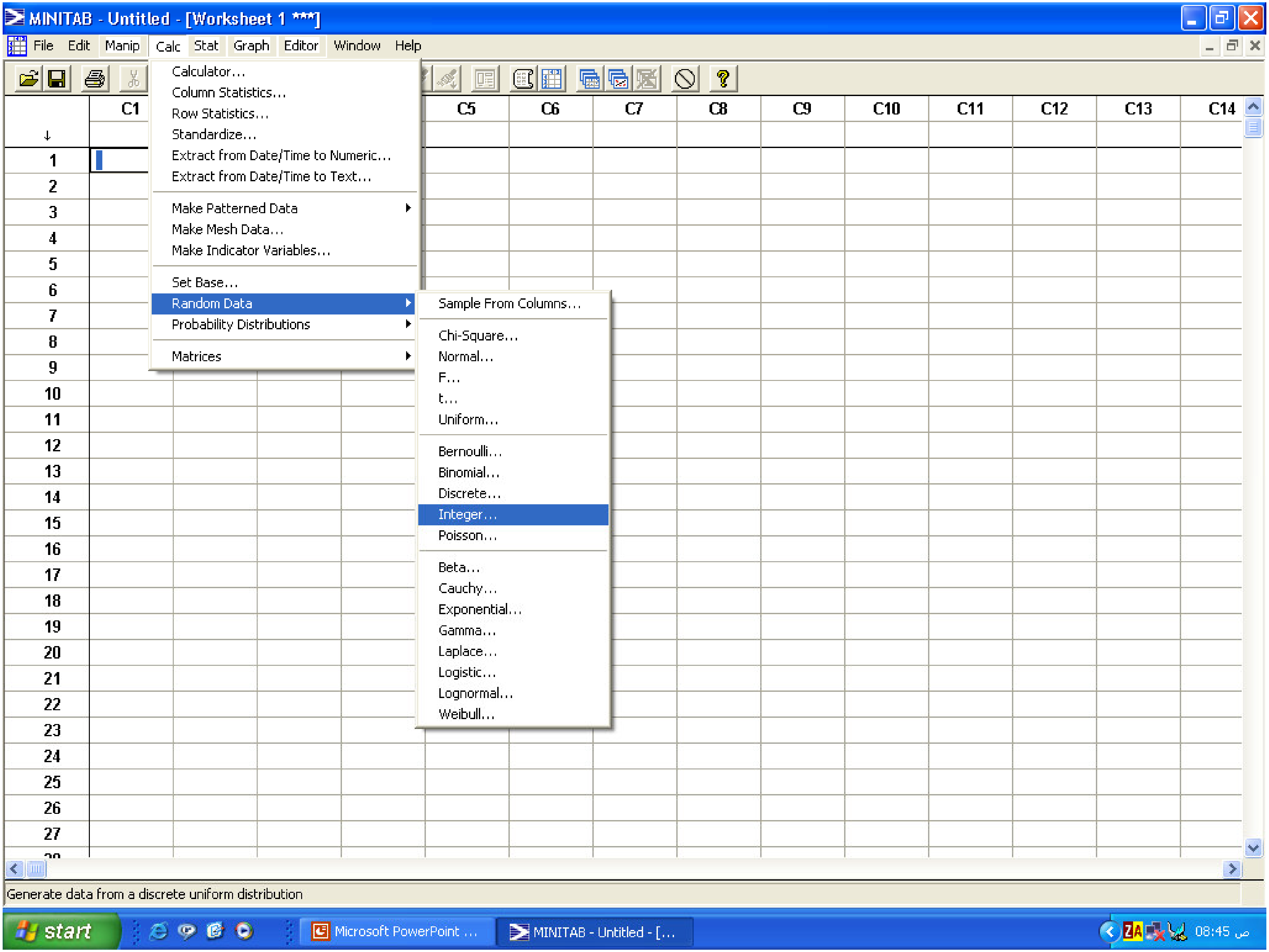
The three basic principles of experimental design are:

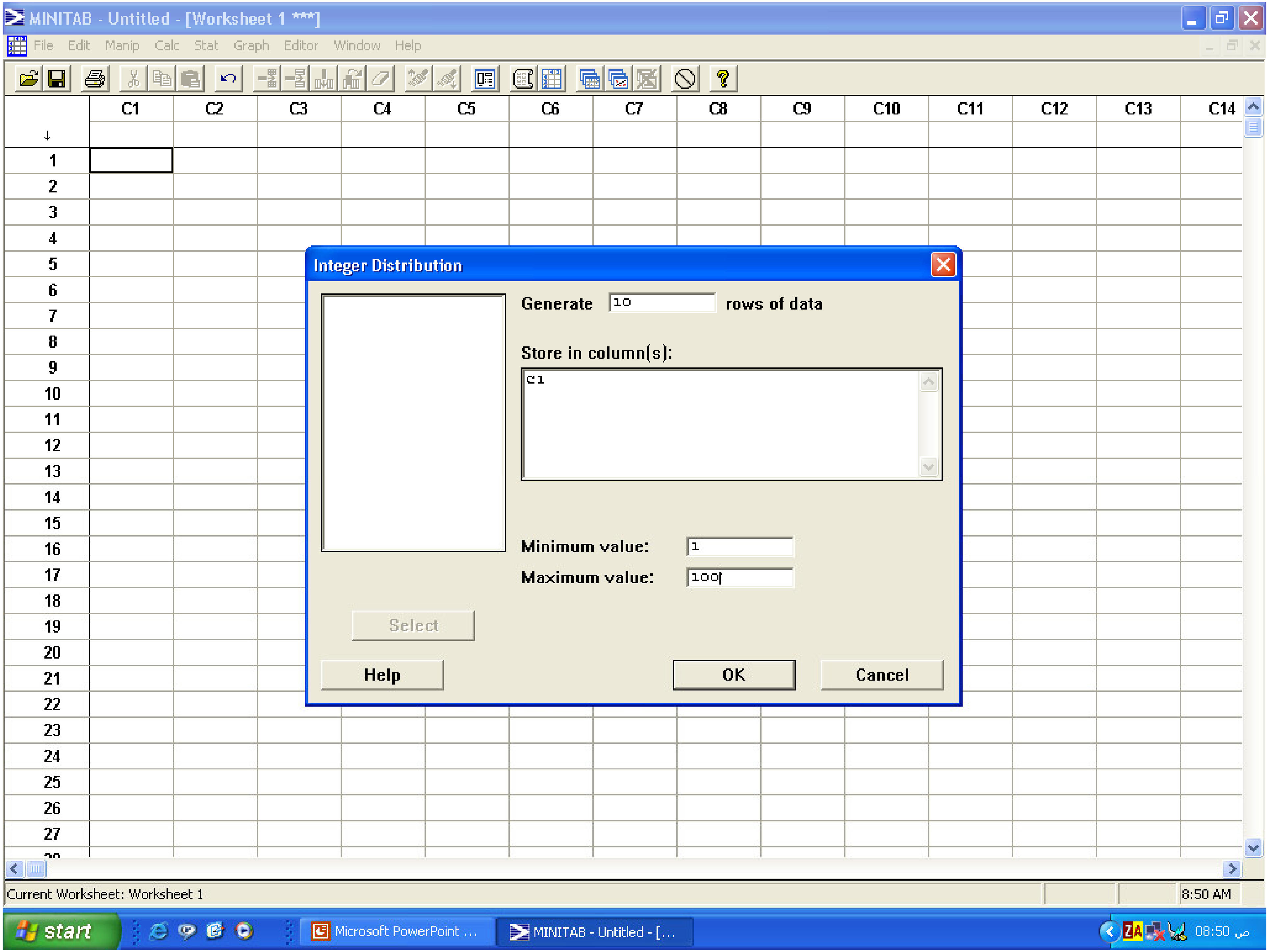
1. Randomization.
2. Replication.
3. Blocking.

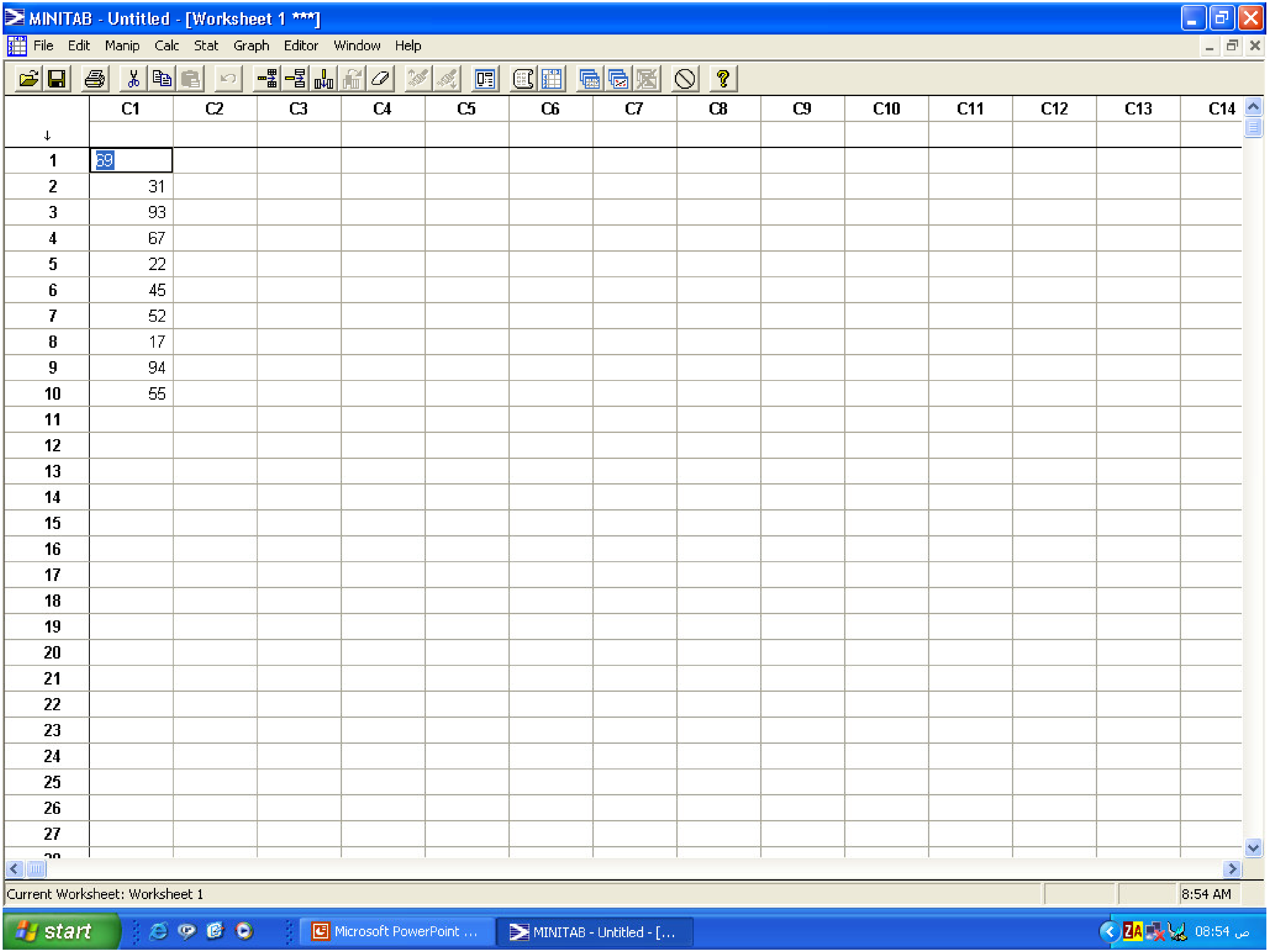
Randomization

Randomization is the cornerstone underlying the use of statistical methods in experimental design. By randomization we mean that both the allocation of the experimental material and the order in which individual runs or trials of the experiment are to be performed are randomly determined.




MINITAB - Untitled - [Worksheet 1 ***]
File Edit Manip Calc Stat Graph Editor Window Help
Calculator...
Column Statistics...
Row Statistics...
Standardize...
Extract from Date/Time to Numeric...
Extract from Date/Time to Text...
Make Patterned Data
Make Mesh Data...
Make Indicator Variables...
Set Base...
Random Data
Probability Distributions
Matrices
Sample From Columns...
Chi-Square...
Normal...
F...
t...
Uniform...
Bernoulli...
Binomial...
Discrete...
Integer...
Poisson...
Beta...
Cauchy...
Exponential...
Gamma...
Laplace...
Logistic...
Lognormal...
Weibull...
C1 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
Generate data from a discrete uniform distribution
start
Microsoft PowerPoint ...
MINITAB - Untitled - [...
08:45


MINITAB - Untitled - [Worksheet 1 ***]
File Edit Manip Calc Stat Graph Editor Window Help
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
Integer Distribution
Generate
10
rows of data
Store in column(s):
c1
Minimum value:
1
Maximum value:
100
Select
Help
OK
Cancel
Current Worksheet: Worksheet 1
8:50 AM
start
Microsoft PowerPoint ...
MINITAB - Untitled - [...
08:50



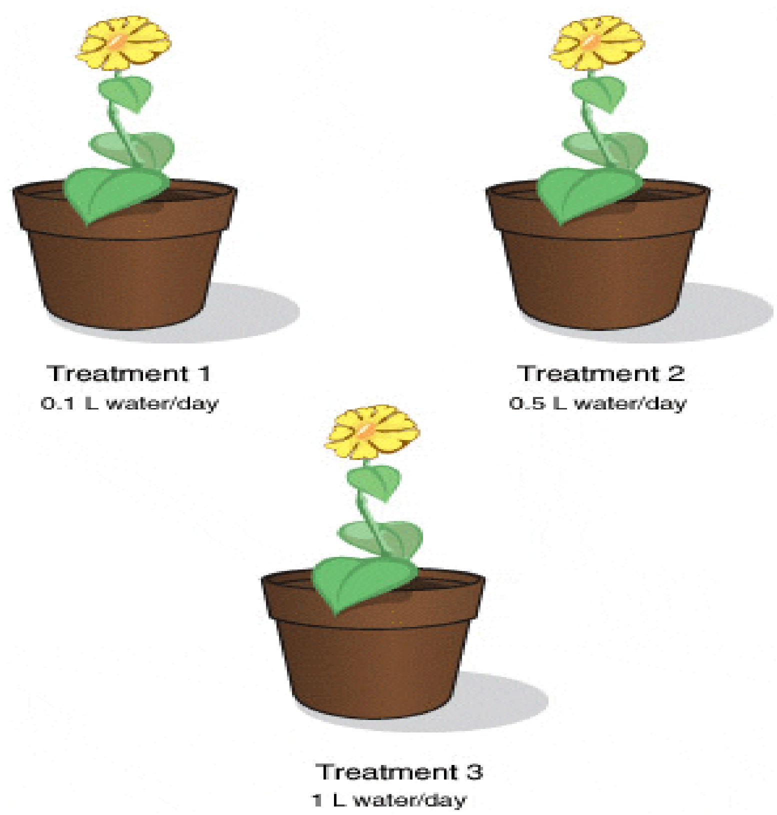
MINITAB - Untitled - [Worksheet 1 ***]
File Edit Manip Calc Stat Graph Editor Window Help
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14
1 69
2 31
3 93
4 67
5 22
6 45
7 52
8 17
9 94
10 55
Current Worksheet: Worksheet 1
8:54 AM
start
Microsoft PowerPoint ...
MINITAB - Untitled - [...
08:54 ص

Replication

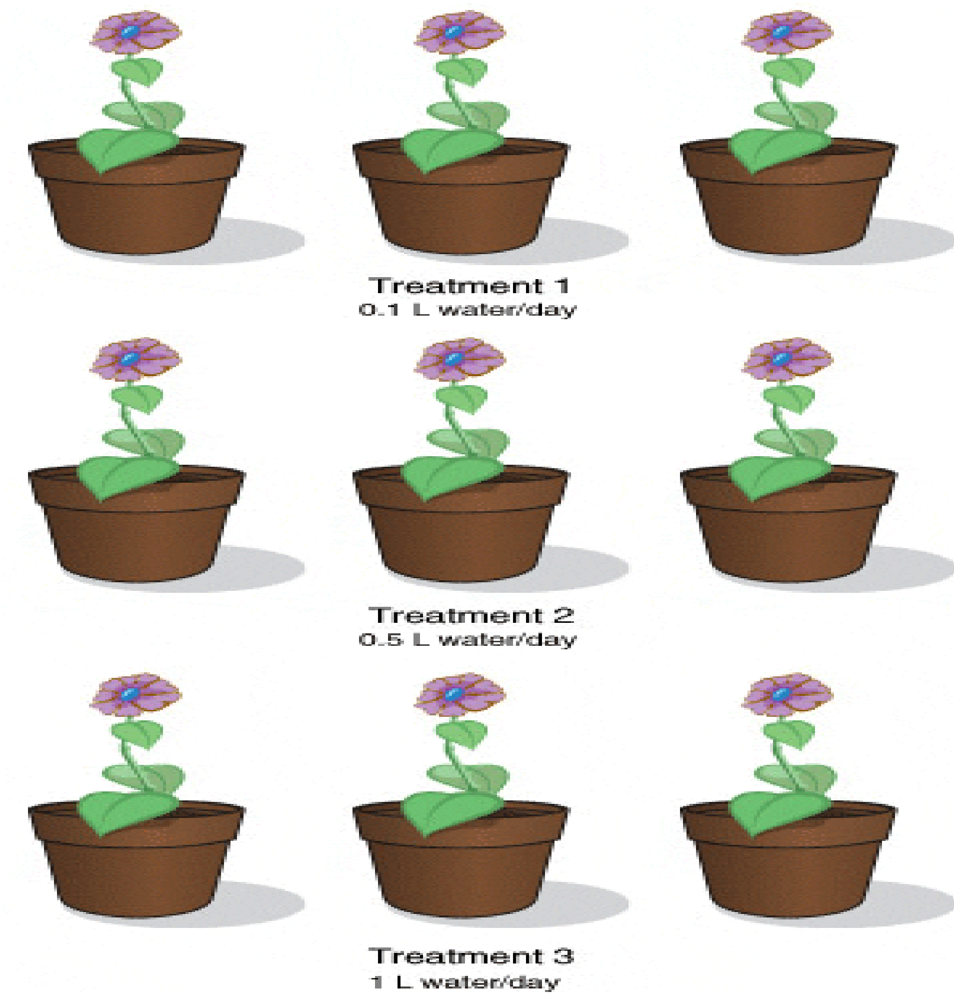
By replication we mean a repetition of the basic experiment. Replication has two important properties:

1. It allows the experimenter to obtain an estimate of the experimental error.
2. If the sample mean is used to estimate the effect of a factor in the experiment, then replication permits the experimenter to obtain a more precise estimate of this effect.

Without replication



With replication



Basic Statistics Terms

Null hypothesis Ho is a hypothesis that is presumed true until statistical evidence in the form of a hypothesis test indicates otherwise.

In formulating a particular null hypothesis, we are always also formulating an **alternative hypothesis Ha**, which we will accept if the observed data values are sufficiently improbable under the null hypothesis .

Definition of Type I and Type II errors

Sometimes our decisions will be correct and sometimes not. There are two possible errors, which we will call Type I and Type II errors, respectively.

A *Type I error* is the error of rejecting the null hypothesis when it is true. The probability of committing a Type I error is usually denoted by α.

A *Type II error* is the error of accepting the null hypothesis when it is false. The probability of making a Type II error is usually denoted by β.

Type I and Type II errors

HYPOTHESIS TESTING OUTCOMES	Reality: The Null Hypothesis Is True	Reality: The Alternative Hypothesis is True
Research: The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
Research: The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

P-value

P-value is a measure of how much evidence we have against the null hypotheses. The smaller the p-value, the more evidence we have against H0.

Traditionally, researchers will reject a hypothesis if the p-value is less than 0.05. Sometimes, though, researchers will use a stricter cut-off (e.g., 0.01) or a more liberal cut-off (e.g., 0.10). The general rule is that a small p-value is evidence against the null hypothesis while a large p-value means little or no evidence against the null hypothesis.

•**P-value**	•**Interpretation**
•$P < 0.01$	•very strong evidence against H0
•$0.01 \leq P < 0.05$	•moderate evidence against H0
•$0.05 \leq P < 0.10$	•suggestive evidence against H0
•$0.10 \leq P$	•little or no real evidence against H0

Choice of sample size

Why would we want to plan?

1. The larger the sample sizes are, the easier it is to detect or find differences in the means.
2. The larger the sample size is, the higher the "cost" and the more likely that practically unimportant differences are to be found statistically significant.

Planning to detect any important difference

Let Δ= smallest difference range considered important by the researcher.

Specify Δ, β, α, σ and r use table A.10 (Applied linear statistical models by Neter, Wasserman and Kunter) to determine the needed sample size n $(=n_1=n_2=\ldots=n_r)$.

Planning to detect any important difference

Example:

Let $\Delta=3$, $\beta=0.1$, $\alpha=0.05$, $\sigma=2$ and $r=4$

$\Delta/\sigma=1.5$, Power$=1-\beta=0.9$

⟹ Need $n=14$ observations at each factor level.

⟹ Need $14*4=56$ homogeneous units.

Planning sample size to find the best treatment

Let λ= important difference between any two adjacent means.

r=number of factor level.

σ= standard deviation.

Specify λ, α, σ and r use table A.11 (Applied linear statistical models by Neter, Wasserman and Kunter) to get d= $\lambda\sqrt{n} / \sigma$ and solve for n.

Planning sample size to find the best treatment

Example:

Let $\lambda=2$, $\alpha=0.05$, $\sigma=3$ and $r=5$

$1-\alpha=0.95$ → $d=3.0552$

$n=(3.0552*3/2)^2=21.002\approx 21$

We need 21 observations at each of 5 levels

→ we need 105 experimental units.

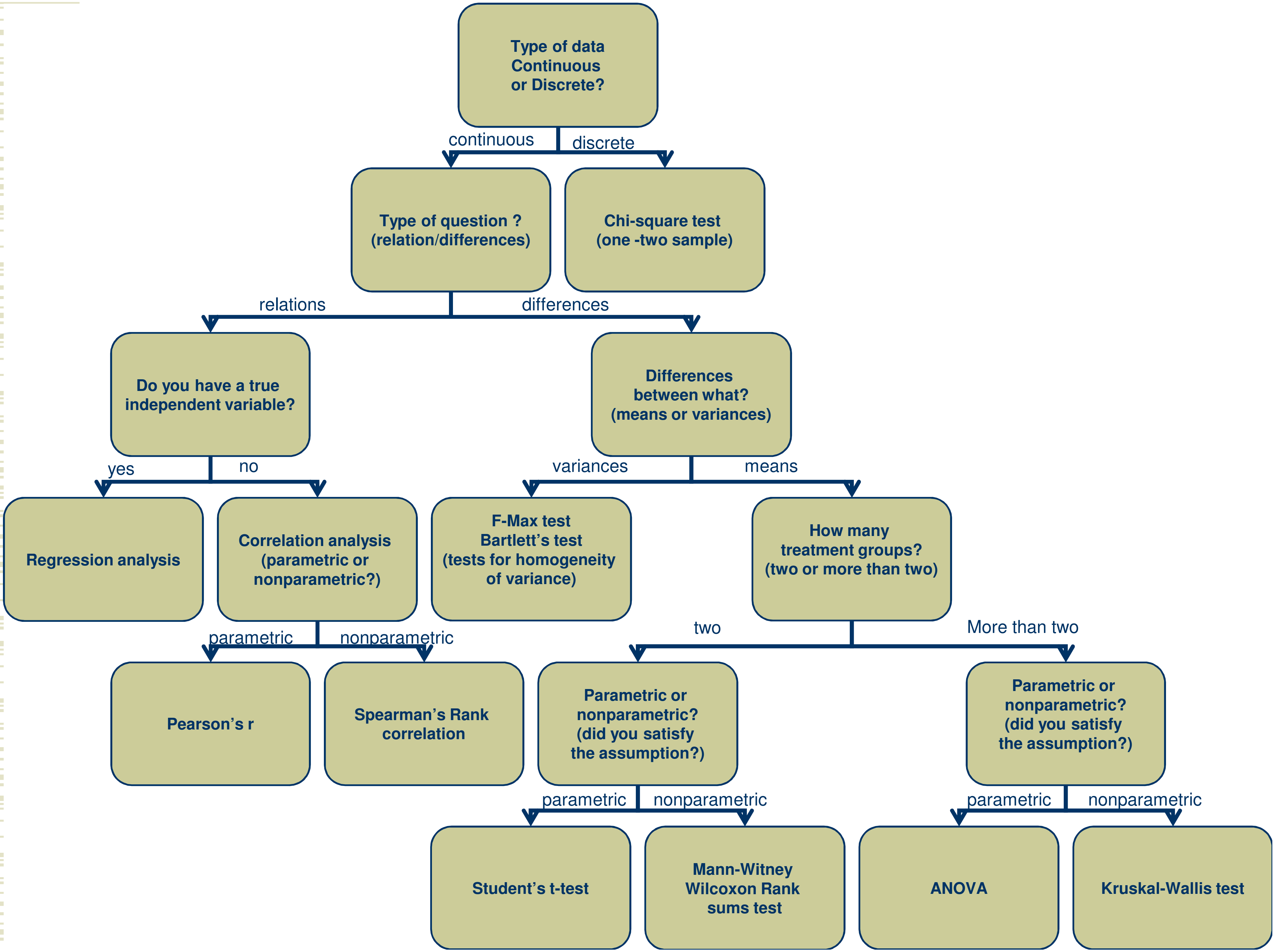
Blocking

If the experimental units are not homogeneous, considerable improvement can be achieved by blocking (grouping) together units that are homogeneous.

Example: Rats coming from the same litter.

Analyzing the data

Your choice of statistical analysis should be made prior to conducting an experiment. There is little sense in collecting data that you can’t analyze properly. Use the following flow chart to help you decide which statistic to use.



Type of data
Continuous
or Discrete?
continuous
discrete
Type of question ?
(relation/differences)
Chi-square test
(one -two sample)
relations
differences
Do you have a true
independent variable?
Differences
between what?
(means or variances)
yes
no
variances
means
Regression analysis
Correlation analysis
(parametric or
nonparametric?)
F-Max test
Bartlett's test
(tests for homogeneity
of variance)
How many
treatment groups?
(two or more than two)
parametric
nonparametric
two
More than two
Pearson's r
Spearman's Rank
correlation
Parametric or
nonparametric?
(did you satisfy
the assumption?)
Parametric or
nonparametric?
(did you satisfy
the assumption?)
parametric
nonparametric
parametric
nonparametric
Student's t-test
Mann-Witney
Wilcoxon Rank
sums test
ANOVA
Kruskal-Wallis test

Assumptions For ANOVA

1. *Normality*: assume that observations in each group are normally distributed.

2. *Homogeneity of the variance*: observations in each group have the same variance

3. *Independence of observations*: this means that knowing one observation in one experimental group tells us nothing about the other observations

Violation of assumptions

ANOVA is robust with respect to the violation of normality and homogeneity of variance.

But

large inequality of the sample sizes

and large heterogeneity of variances , are bad

The Logic of ANOVA

F statistic

$$F = \frac{\text{obtained variance between sample means}}{\text{variance expected by chance}}$$

Random variation + effect

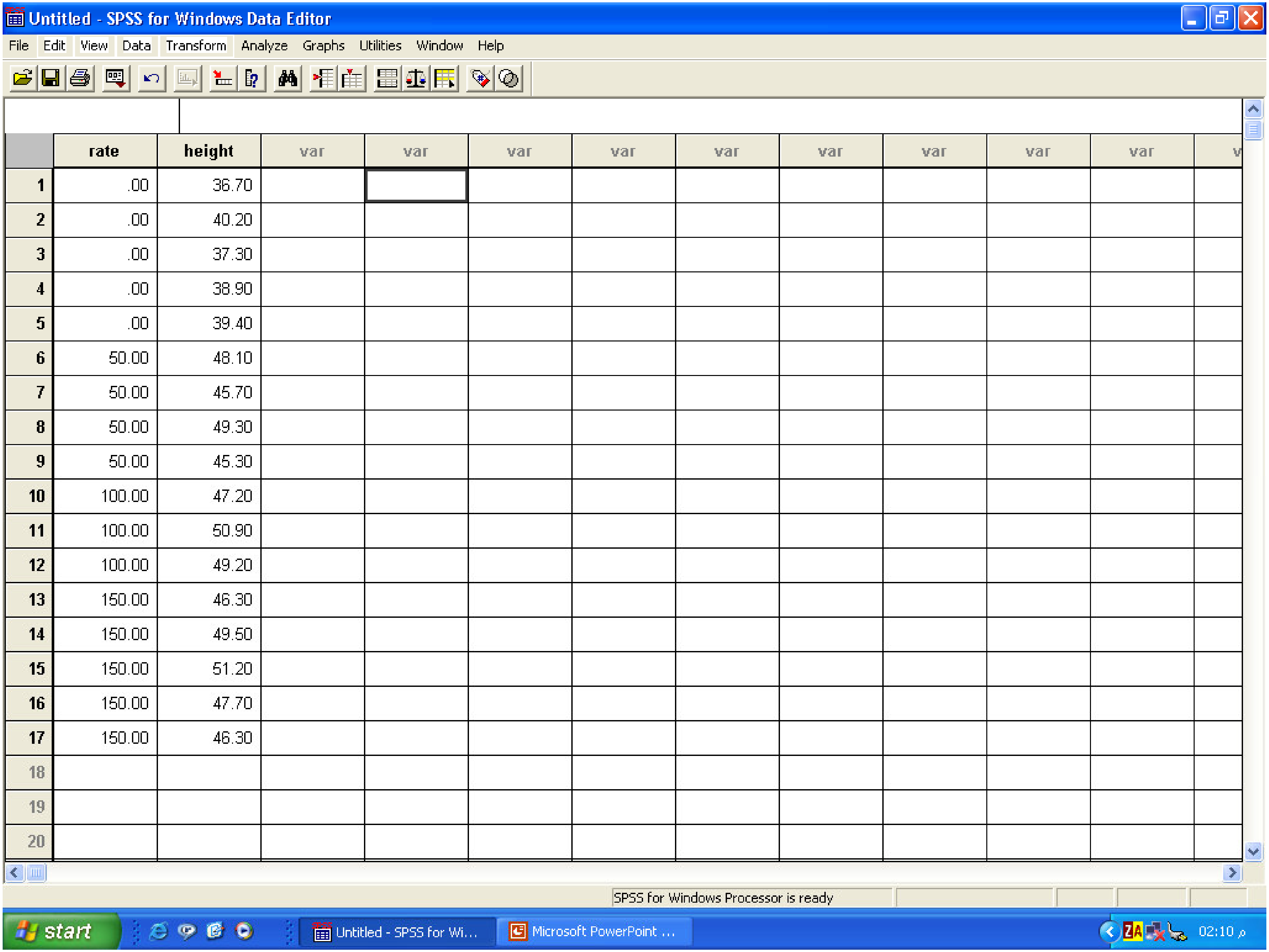
Random variation

ANOVA

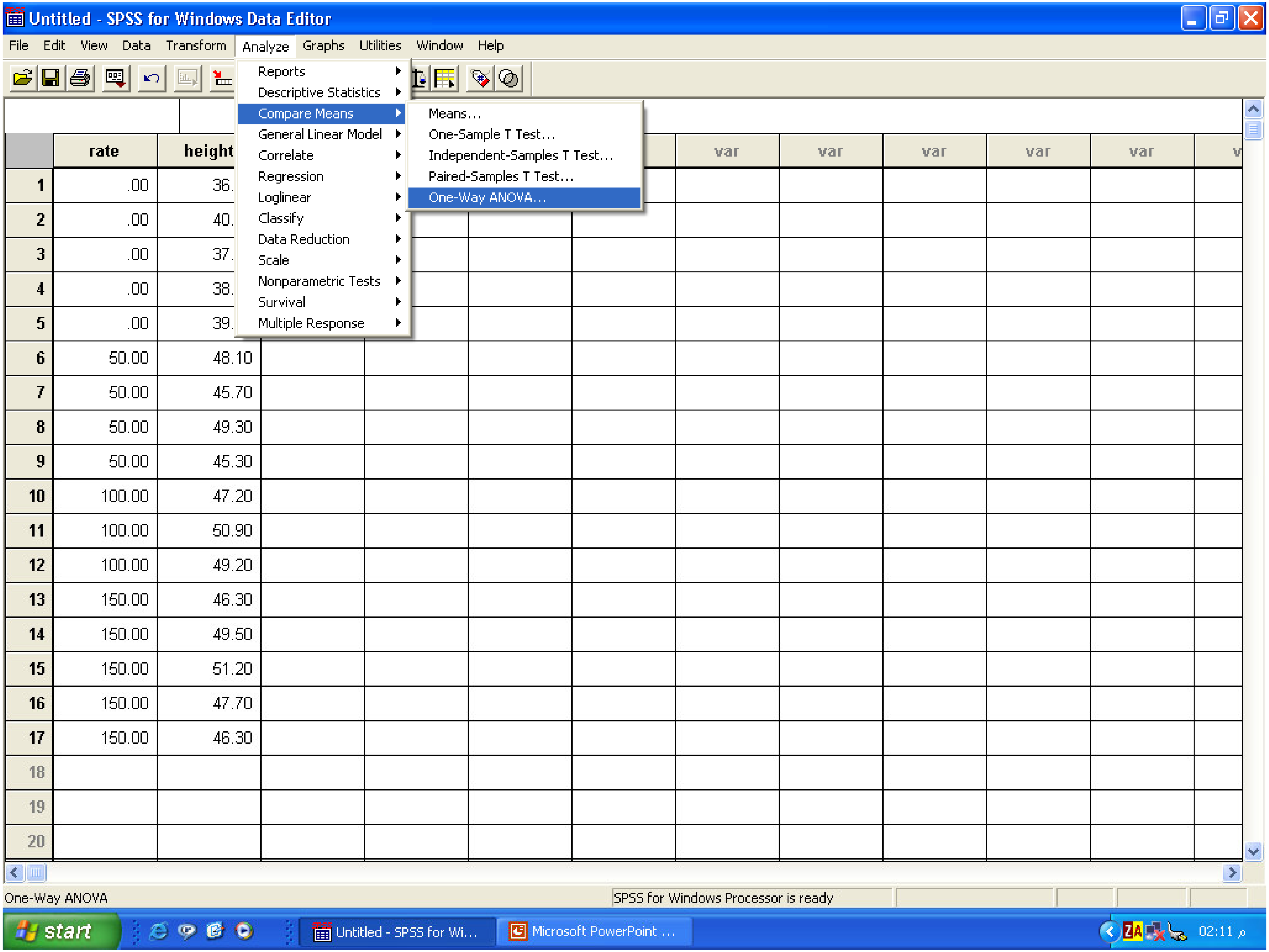
Example:

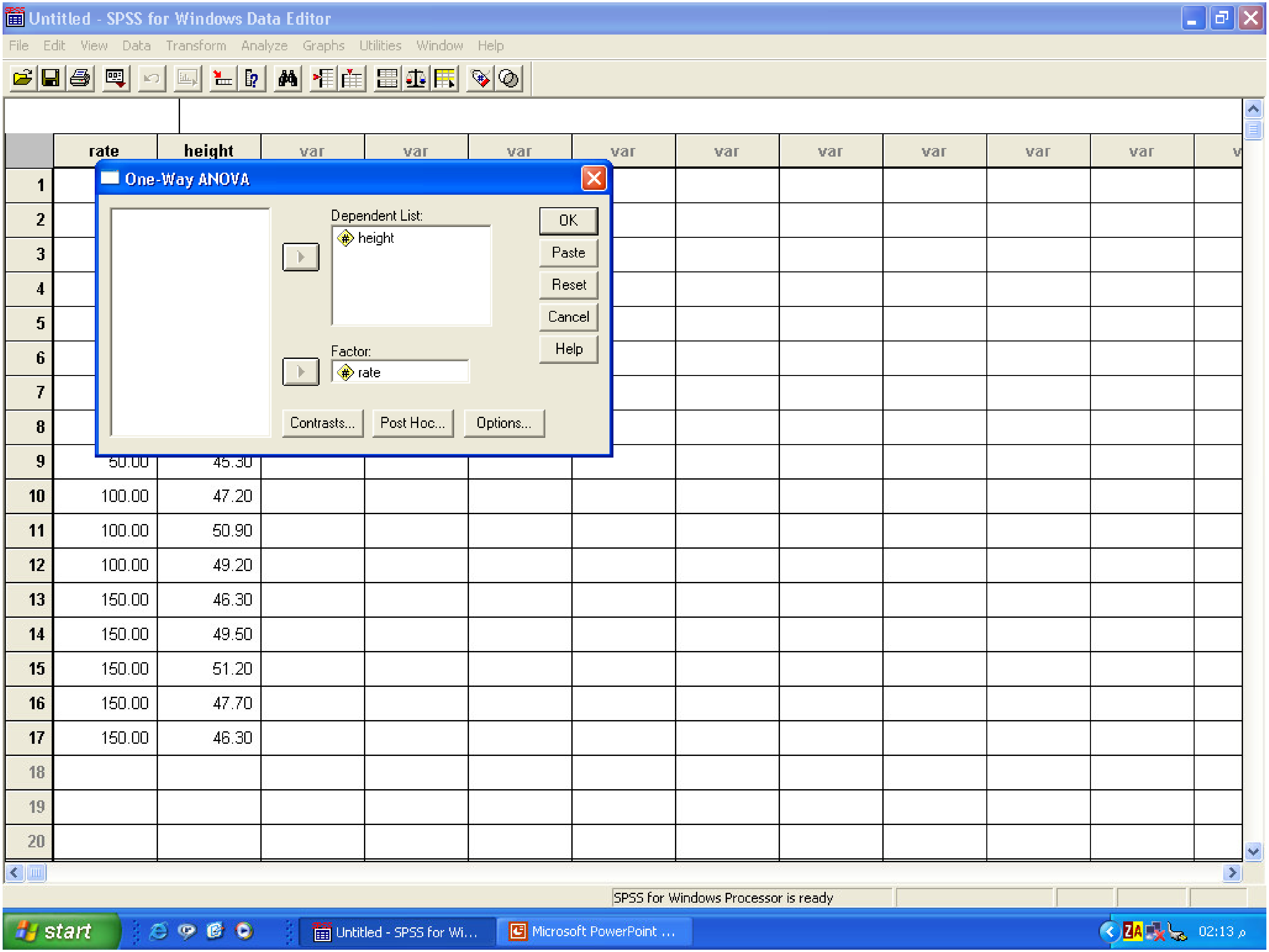
In a study on the effect of nitrogen fertilization on cereal crops, plots of a particular variety of wheat were randomly given fertilizer at one of four rates: 0, 50, 100, 150. At a certain date, plants were randomly selected from the plots and the plant height (in cm) was measured [based on Ghandorah(1985a)].

Can we conclude that all 4 fertilizer rates have equal effects on the average plant height?



	rate	height
1	.00	36.70
2	.00	40.20
3	.00	37.30
4	.00	38.90
5	.00	39.40
6	50.00	48.10
7	50.00	45.70
8	50.00	49.30
9	50.00	45.30
10	100.00	47.20
11	100.00	50.90
12	100.00	49.20
13	150.00	46.30
14	150.00	49.50
15	150.00	51.20
16	150.00	47.70
17	150.00	46.30


Untitled - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Survival
Multiple Response
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
rate
height
var
1 .00 36.
2 .00 40.
3 .00 37.
4 .00 38.
5 .00 39.
6 50.00 48.10
7 50.00 45.70
8 50.00 49.30
9 50.00 45.30
10 100.00 47.20
11 100.00 50.90
12 100.00 49.20
13 150.00 46.30
14 150.00 49.50
15 150.00 51.20
16 150.00 47.70
17 150.00 46.30
18
19
20
One-Way ANOVA
SPSS for Windows Processor is ready
start
Untitled - SPSS for Wi...
Microsoft PowerPoint ...
02:11


Untitled - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
rate
height
var
One-Way ANOVA
Dependent List:
height
Factor:
rate
OK
Paste
Reset
Cancel
Help
Contrasts...
Post Hoc...
Options...
9 50.00 45.30
10 100.00 47.20
11 100.00 50.90
12 100.00 49.20
13 150.00 46.30
14 150.00 49.50
15 150.00 51.20
16 150.00 47.70
17 150.00 46.30
SPSS for Windows Processor is ready
start
Untitled - SPSS for Wi...
Microsoft PowerPoint ...
02:13

ANOVA

ANOVA

HEIGHT

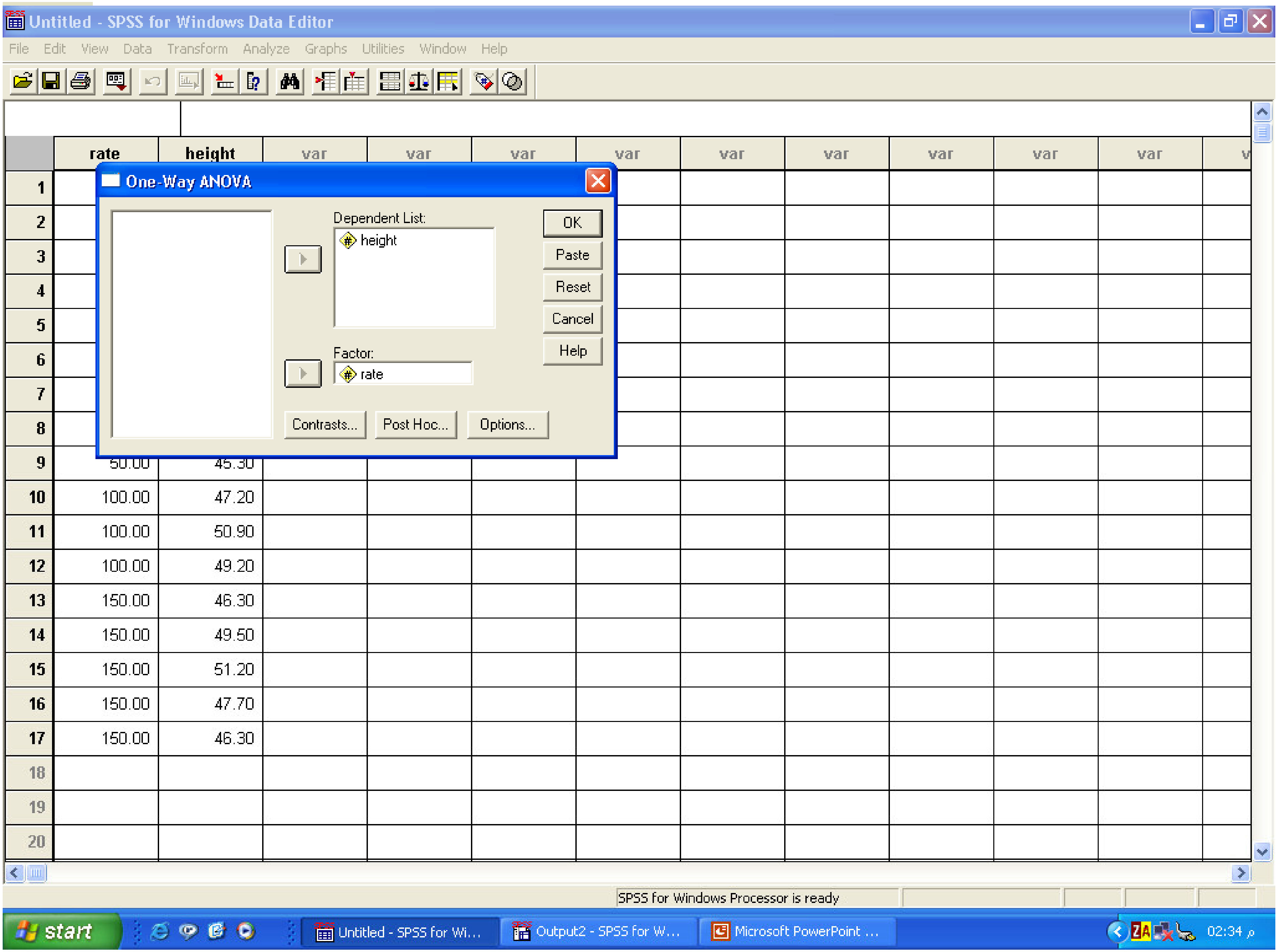
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	329.482	3	109.827	32.012	.000
Within Groups	44.600	13	3.431		
Total	374.082	16			

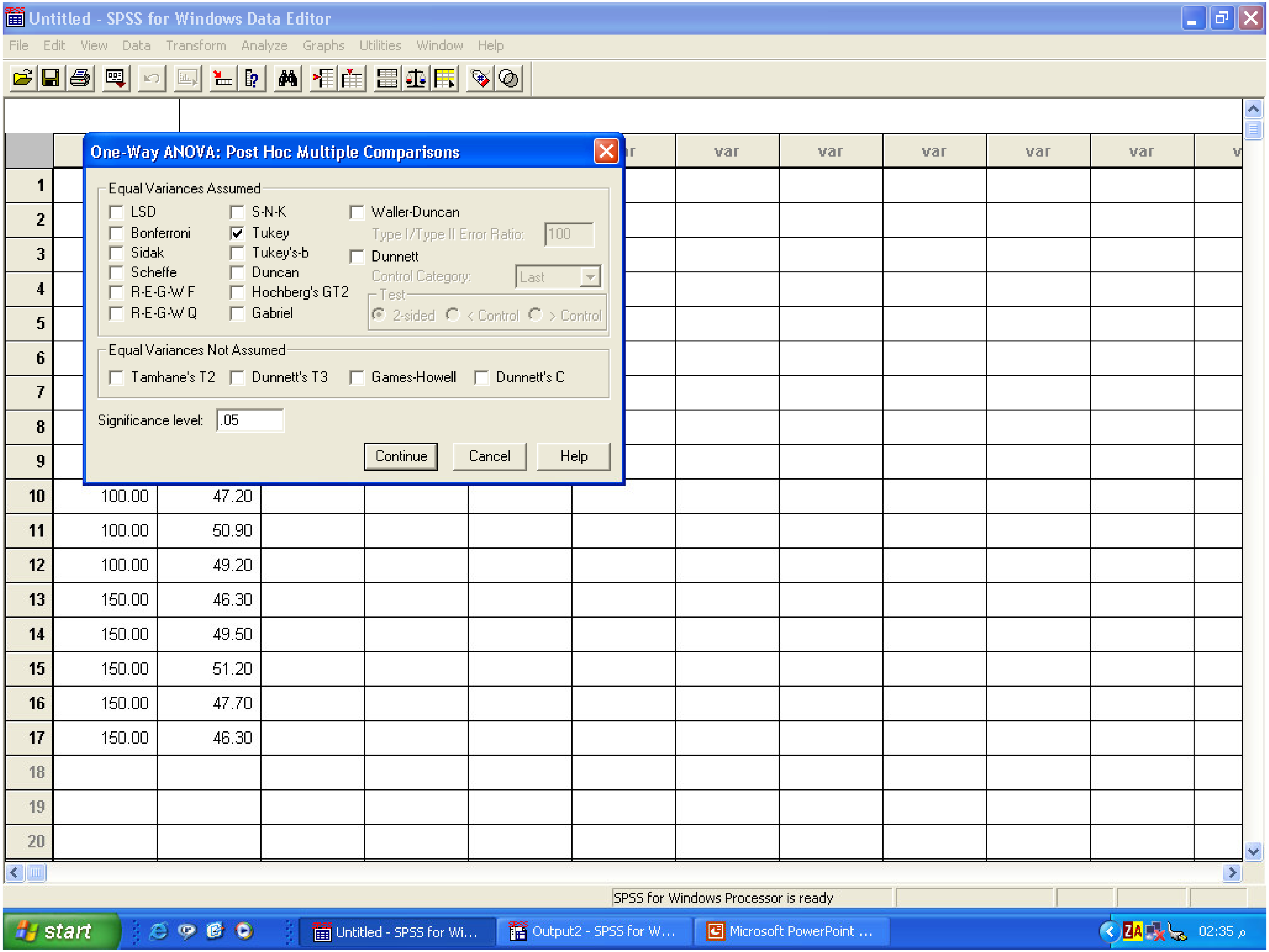
Ho: all 4 fertilizer rates have the same effect on plant height.

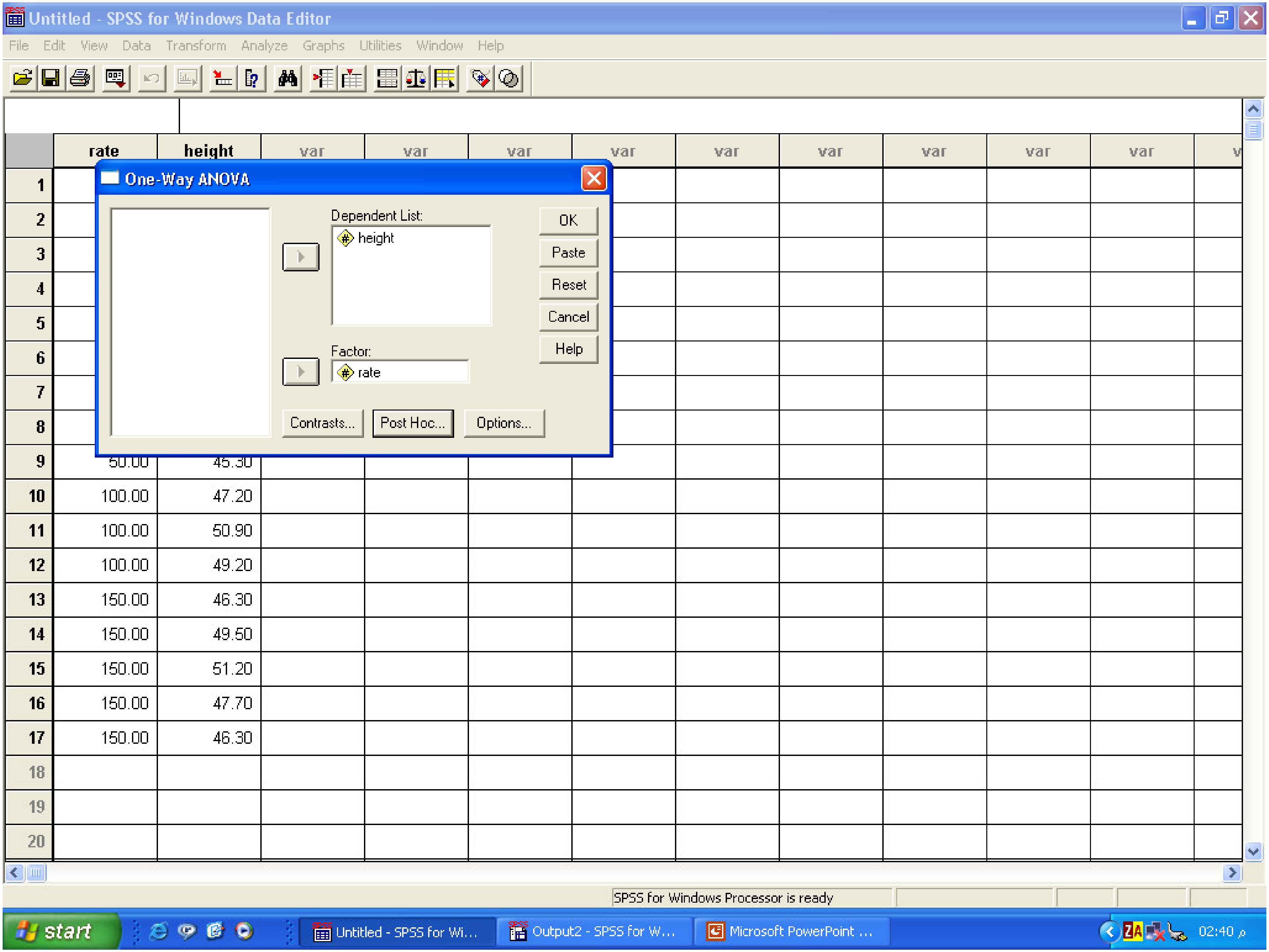
Ha: Some of the 4 fertilizer rates have different effects on plant height.

P-value = 0

We conclude that at least one of the 4 fertilizer rates have different effects on the average plant height.


Untitled - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
rate
height
var
One-Way ANOVA
Dependent List:
height
Factor:
rate
OK
Paste
Reset
Cancel
Help
Contrasts...
Post Hoc...
Options...
9 50.00 45.30
10 100.00 47.20
11 100.00 50.90
12 100.00 49.20
13 150.00 46.30
14 150.00 49.50
15 150.00 51.20
16 150.00 47.70
17 150.00 46.30
SPSS for Windows Processor is ready
start
Untitled - SPSS for Wi...
Output2 - SPSS for W...
Microsoft PowerPoint ...
02:34


Untitled - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
One-Way ANOVA: Post Hoc Multiple Comparisons
Equal Variances Assumed
LSD
Bonferroni
Sidak
Scheffe
R-E-G-W F
R-E-G-W Q
S-N-K
Tukey
Tukey's-b
Duncan
Hochberg's GT2
Gabriel
Waller-Duncan
Type I/Type II Error Ratio: 100
Dunnett
Control Category: Last
Test
2-sided
< Control
> Control
Equal Variances Not Assumed
Tamhane's T2
Dunnett's T3
Games-Howell
Dunnett's C
Significance level: .05
Continue
Cancel
Help
var
1
2
3
4
5
6
7
8
9
10 100.00 47.20
11 100.00 50.90
12 100.00 49.20
13 150.00 46.30
14 150.00 49.50
15 150.00 51.20
16 150.00 47.70
17 150.00 46.30
18
19
20
SPSS for Windows Processor is ready
start
Untitled - SPSS for Wi...
Output2 - SPSS for W...
Microsoft PowerPoint ...
02:35


Untitled - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
rate height var var var var var var var var var
One-Way ANOVA
Dependent List:
height
Factor:
rate
OK
Paste
Reset
Cancel
Help
Contrasts...
Post Hoc...
Options...
9 50.00 45.30
10 100.00 47.20
11 100.00 50.90
12 100.00 49.20
13 150.00 46.30
14 150.00 49.50
15 150.00 51.20
16 150.00 47.70
17 150.00 46.30
18
19
20
SPSS for Windows Processor is ready
start
Untitled - SPSS for Wi...
Output2 - SPSS for W...
Microsoft PowerPoint ...
02:40

Output2 - SPSS for Windows Viewer

Post Hoc Tests

Multiple Comparisons

Dependent Variable: HEIGHT
Tukey HSD

(I) RATE	(J) RATE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
.00	50.00	-8.6000*	1.2425	.000	-12.2470	-4.9530
	100.00	-10.6000*	1.3527	.000	-14.5703	-6.6297
	150.00	-9.7000*	1.1715	.000	-13.1384	-6.2616
50.00	.00	8.6000*	1.2425	.000	4.9530	12.2470
	100.00	-2.0000	1.4147	.513	-6.1522	2.1522
	150.00	-1.1000	1.2425	.812	-4.7470	2.5470
100.00	.00	10.6000*	1.3527	.000	6.6297	14.5703
	50.00	2.0000	1.4147	.513	-2.1522	6.1522
	150.00	.9000	1.3527	.908	-3.0703	4.8703
150.00	.00	9.7000*	1.1715	.000	6.2616	13.1384
	50.00	1.1000	1.2425	.812	-2.5470	4.7470
	100.00	-.9000	1.3527	.908	-4.8703	3.0703

*. The mean difference is significant at the .05 level.

Homogeneous Subsets

Output2 - SPSS for Windows Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

- Output
  - Oneway
    - Title
    - Notes
    - ANOVA
  - Oneway
    - Title
    - Notes
    - ANOVA
    - Post Hoc Tests
      - Title
      - Multiple Comparisons
      - Homogeneous Subsets
        - Title
        - HEIGHT

	100.00	-2.0000	1.4147	.513	-6.1522	2.1522
	150.00	-1.1000	1.2425	.812	-4.7470	2.5470
100.00	.00	10.6000*	1.3527	.000	6.6297	14.5703
	50.00	2.0000	1.4147	.513	-2.1522	6.1522
	150.00	.9000	1.3527	.908	-3.0703	4.8703
150.00	.00	9.7000*	1.1715	.000	6.2616	13.1384
	50.00	1.1000	1.2425	.812	-2.5470	4.7470
	100.00	-.9000	1.3527	.908	-4.8703	3.0703

*. The mean difference is significant at the .05 level.

Homogeneous Subsets

HEIGHT

Tukey HSD[a,b]

		Subset for alpha = .05	
RATE	N	1	2
.00	5	38.5000	
50.00	4		47.1000
150.00	5		48.2000
100.00	3		49.1000
Sig.		1.000	.444

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.068.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

SPSS for Windows Processor is ready

start | Untitled - SPSS for Wi... | Output2 - SPSS for W... | Microsoft PowerPoint ... | 02:38

ANOVA

0	50	100	150
38.5	47.1	49.1	48.2

From mean separation we can see that there is no significant difference between the effect of 50 , 100 and 150 fertilizer rates on plant height.

We can recommend to use the 50 fertilizer rate because it is coast effective.

Extra Example on Blocking

In a study it was desired to know the effect of water stress on the protein content of wheat.

Because the protein content of wheat is known to differ from one variety to another, six local varieties of Saudi wheat were chosen for the experiment and it was assumed that there is no interaction between the wheat varieties and the water stress levels on the protein content. Therefore, three plots of each type of wheat were chosen and randomly assigned to the three levels of water stress, namely three watering intervals of every 10, 16, and 22 days. After harvest, the wheat from each plot was separately ground into flour, and the protein content (as a percent of the dry weight) was measured [Based on Basahy (1990)].

glmdata - SPSS for Windows Data Editor

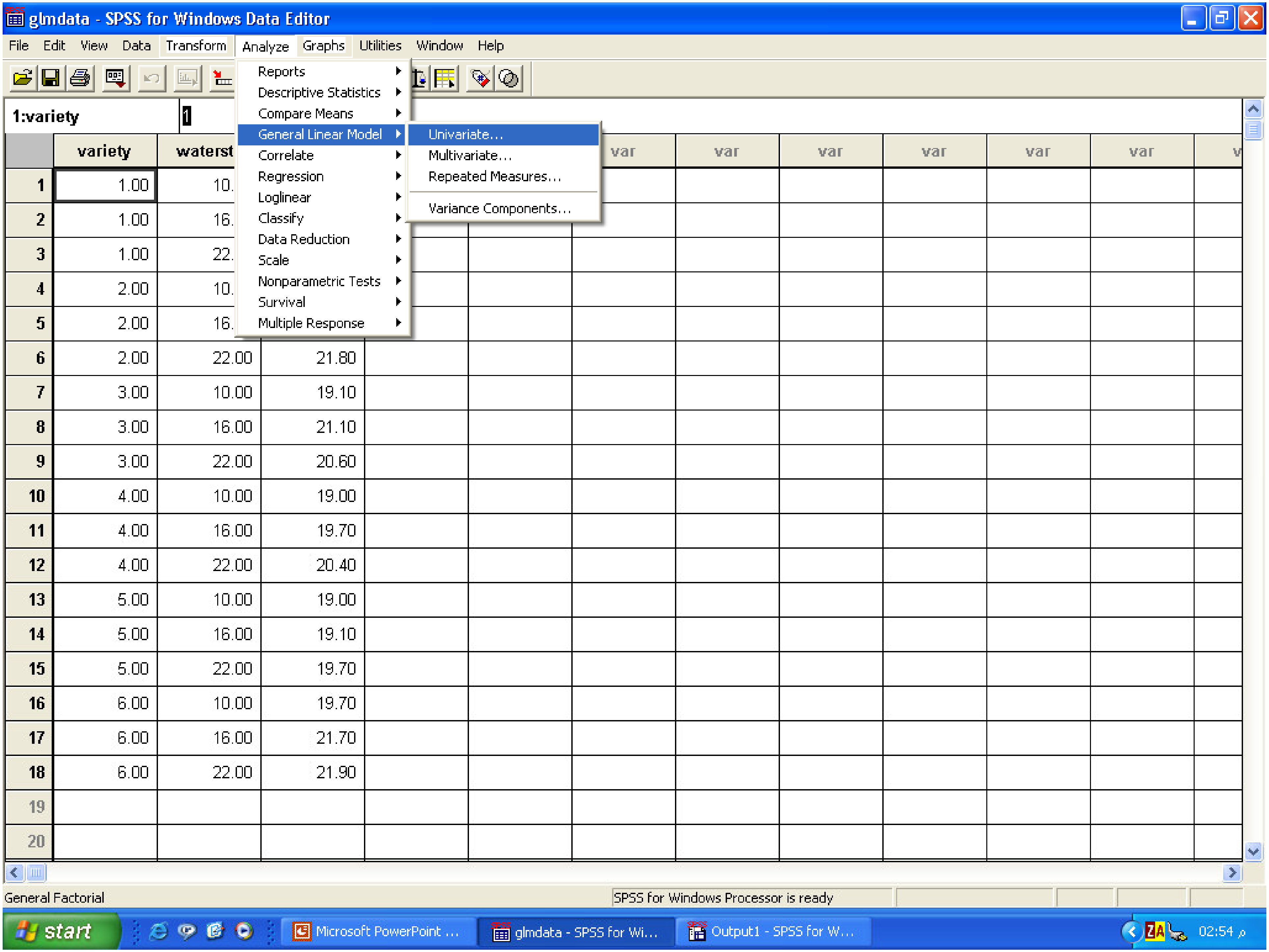
File Edit View Data Transform Analyze Graphs Utilities Window Help

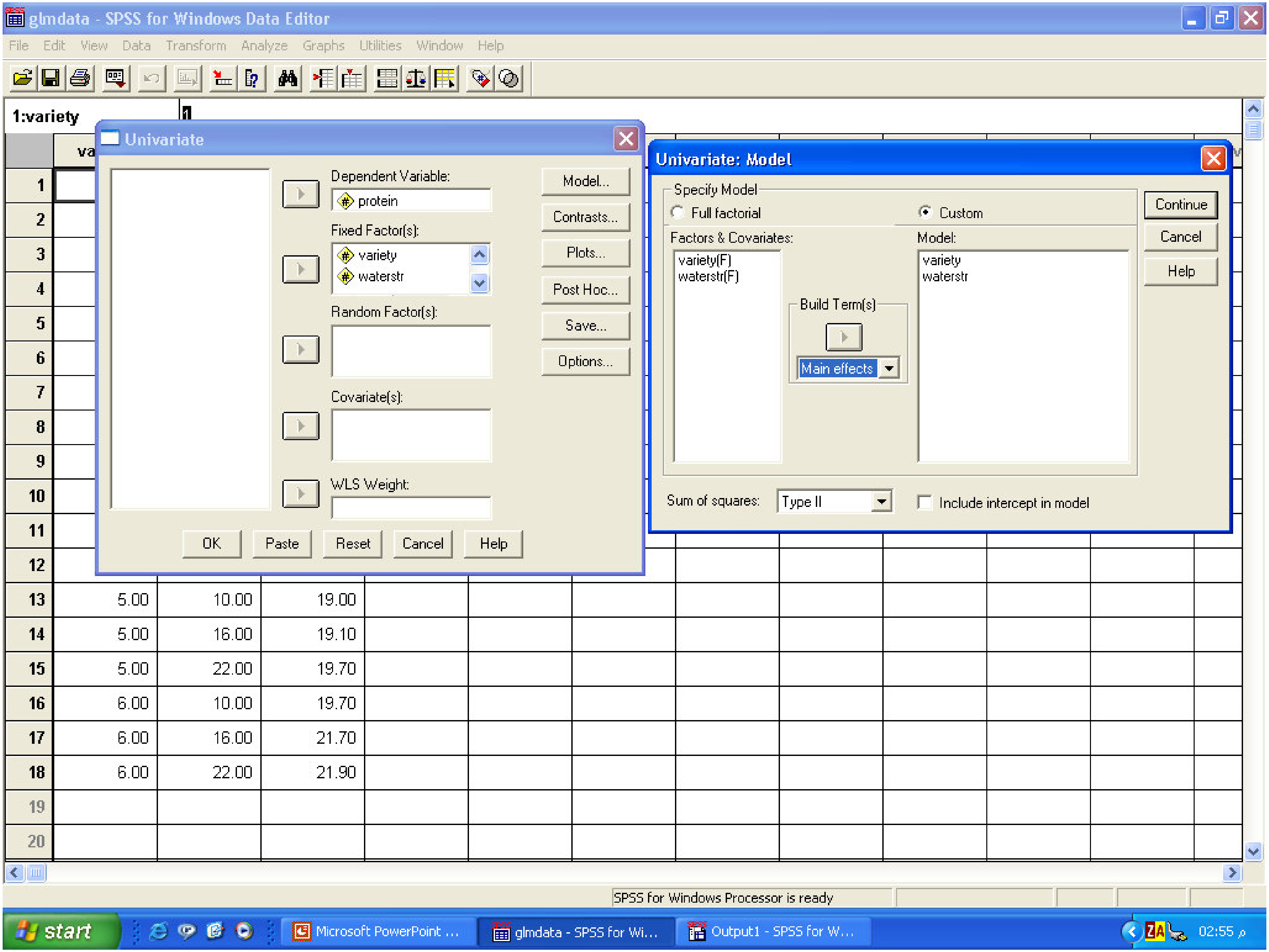
1:variety 1

	variety	waterstr	protein	var	var	var	var	var	var	var	var
1	1.00	10.00	19.20								
2	1.00	16.00	19.10								
3	1.00	22.00	20.60								
4	2.00	10.00	19.00								
5	2.00	16.00	21.40								
6	2.00	22.00	21.80								
7	3.00	10.00	19.10								
8	3.00	16.00	21.10								
9	3.00	22.00	20.60								
10	4.00	10.00	19.00								
11	4.00	16.00	19.70								
12	4.00	22.00	20.40								
13	5.00	10.00	19.00								
14	5.00	16.00	19.10								
15	5.00	22.00	19.70								
16	6.00	10.00	19.70								
17	6.00	16.00	21.70								
18	6.00	22.00	21.90								
19											
20											

SPSS for Windows Processor is ready

start | Microsoft PowerPoint ... | glmdata - SPSS for Wi... | Output1 - SPSS for W... | 02:53


glmdata - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1:variety 1
Reports
Descriptive Statistics
Compare Means
General Linear Model
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Survival
Multiple Response
Univariate...
Multivariate...
Repeated Measures...
Variance Components...
variety
waterst
var
1 1.00 10.
2 1.00 16.
3 1.00 22.
4 2.00 10.
5 2.00 16.
6 2.00 22.00 21.80
7 3.00 10.00 19.10
8 3.00 16.00 21.10
9 3.00 22.00 20.60
10 4.00 10.00 19.00
11 4.00 16.00 19.70
12 4.00 22.00 20.40
13 5.00 10.00 19.00
14 5.00 16.00 19.10
15 5.00 22.00 19.70
16 6.00 10.00 19.70
17 6.00 16.00 21.70
18 6.00 22.00 21.90
19
20
General Factorial
SPSS for Windows Processor is ready
start
Microsoft PowerPoint ...
glmdata - SPSS for Wi...
Output1 - SPSS for W...
02:54


glmdata - SPSS for Windows Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1:variety
Univariate
Dependent Variable:
protein
Fixed Factor(s):
variety
waterstr
Random Factor(s):
Covariate(s):
WLS Weight:
Model...
Contrasts...
Plots...
Post Hoc...
Save...
Options...
OK Paste Reset Cancel Help
Univariate: Model
Specify Model
Full factorial
Custom
Factors & Covariates:
variety(F)
waterstr(F)
Build Term(s)
Main effects
Model:
variety
waterstr
Continue
Cancel
Help
Sum of squares:
Type II
Include intercept in model
13 5.00 10.00 19.00
14 5.00 16.00 19.10
15 5.00 22.00 19.70
16 6.00 10.00 19.70
17 6.00 16.00 21.70
18 6.00 22.00 21.90
SPSS for Windows Processor is ready
start
Microsoft PowerPoint ...
glmdata - SPSS for Wi...
Output1 - SPSS for W...
02:55

Output2 - SPSS for Windows Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

- Output
  - Univariate Analysis of Variai
    - Title
    - Notes
    - Between-Subjects Facto
    - Tests of Between-Subje

Univariate Analysis of Variance

Between-Subjects Factors

		N
VARIETY	1.00	3
	2.00	3
	3.00	3
	4.00	3
	5.00	3
	6.00	3
WATERSTR	10.00	6
	16.00	6
	22.00	6

Tests of Between-Subjects Effects

Dependent Variable: PROTEIN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Model	7300.567[a]	8	912.571	2561.003	.000
VARIETY	7.498	5	1.500	4.209	.025
WATERSTR	8.823	2	4.412	12.381	.002
Error	3.563	10	.356		
Total	7304.130	18			

a. R Squared = 1.000 (Adjusted R Squared = .999)

SPSS for Windows Processor is ready

start | Microsoft PowerPoint ... | glmdata - SPSS for Wi... | Output2 - SPSS for W... | 02:57

Output2 - SPSS for Windows Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

Output
- Univariate Analysis of Variance
  - Title
  - Notes
  - Between-Subjects Factors
  - Tests of Between-Subjects Effects

Univariate Analysis of Variance

Between-Subjects Factors

		N
VARIETY	1.00	3
	2.00	3
	3.00	3
	4.00	3
	5.00	3
	6.00	3
WATERSTR	10.00	6
	16.00	6
	22.00	6

Tests of Between-Subjects Effects

Dependent Variable: PROTEIN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Model	7300.567[a]	8	912.571	2561.003	.000
VARIETY	7.498	5	1.500	4.209	.025
WATERSTR	8.823	2	4.412	12.381	.002
Error	3.563	10	.356		
Total	7304.130	18			

a. R Squared = 1.000 (Adjusted R Squared = .999)

SPSS for Windows Processor is ready

start | Microsoft PowerPoint ... | glmdata - SPSS for Wi... | Output2 - SPSS for W... | 02:57

Results

Using the randomized block design there is a significant difference in the protein content using different levels of water stress while using the simple one way anova there is no significant difference at $\alpha=0.01$.

Tests of Between-Subjects Effects

Dependent Variable: PROTEIN

Source	Type II Sum of Squares	df	Mean Square	F	Sig.
Model	7300.567[a]	8	912.571	2561.003	.000
VARIETY	7.498	5	1.500	4.209	.025
WATERSTR	8.823	2	4.412	12.381	.002
Error	3.563	10	.356		
Total	7304.130	18			

a. R Squared = 1.000 (Adjusted R Squared = .999)

ANOVA

PROTEIN

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.823	2	4.412	5.982	.012
Within Groups	11.062	15	.737		
Total	19.885	17			

Conclusions and recommendations

- Once the data has been analyzed, the experimenter may draw conclusions or inferences about the results. The statistical inference must be physically interpreted, and the practical significance of these findings evaluated. Then recommendations concerning these findings must be made.
- The use of graphical display is a very effective way to present experimental results.

Thank you for listening