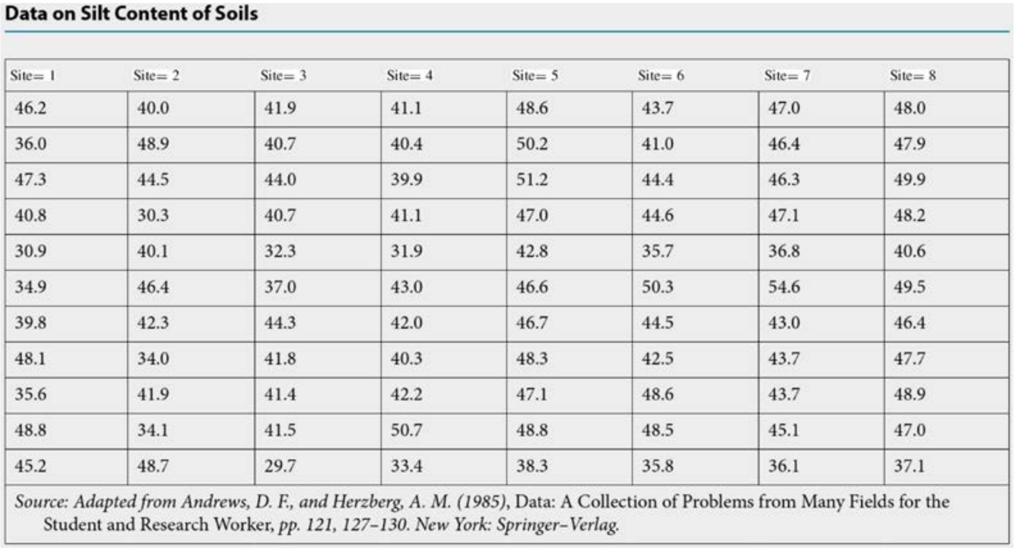
STAT 441/541 Statistical Methods II
Handout: Fun with a Statistical Dataset for One-Way ANOVA

Introduction

The data sets we have used so far in this course has been statistical datasets. The data in Excel files are arranged so that columns represent variables and rows represent observations. However, we may be given a table that contains data and we need to enter the data into an Excel file for statistical analysis. For example, the silt content of soils dataset is given in a data table:



Data on Silt Content of Soils

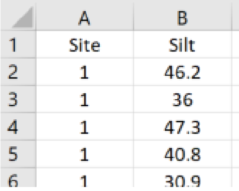
Site= 1	Site= 2	Site= 3	Site= 4	Site= 5	Site= 6	Site= 7	Site= 8
46.2	40.0	41.9	41.1	48.6	43.7	47.0	48.0
36.0	48.9	40.7	40.4	50.2	41.0	46.4	47.9
47.3	44.5	44.0	39.9	51.2	44.4	46.3	49.9
40.8	30.3	40.7	41.1	47.0	44.6	47.1	48.2
30.9	40.1	32.3	31.9	42.8	35.7	36.8	40.6
34.9	46.4	37.0	43.0	46.6	50.3	54.6	49.5
39.8	42.3	44.3	42.0	46.7	44.5	43.0	46.4
48.1	34.0	41.8	40.3	48.3	42.5	43.7	47.7
35.6	41.9	41.4	42.2	47.1	48.6	43.7	48.9
48.8	34.1	41.5	50.7	48.8	48.5	45.1	47.0
45.2	48.7	29.7	33.4	38.3	35.8	36.1	37.1

Source: Adapted from Andrews, D. F., and Herzberg, A. M. (1985), Data: A Collection of Problems from Many Fields for the Student and Research Worker, *pp. 121, 127–130. New York: Springer–Verlag.*

For One-Way Analysis of Variance, we will need to create two variables. One variable contains information on the different treatments (sites) and the other variable contains information for the measured response variable. The variable identifying treatments is called a factor and must be defined "as factors" when using RStudio for data analysis.

Let's create a variable called "Site" to denote the different treatments, or sites. Then let's create another variable called "Silt" for the silt content of soils. Note that there are 11 observations for each site and eight sites. Therefore, we have a total of 88 observations. So our data file will have 88 rows, one row for each observation. Each row will consist of a value denoting the treatment and a value for the response variable. We can use the first row of a statistical dataset for variable names.

An excerpt from Excel for the first few observations for Site 1:



	A	B
1	Site	Silt
2	1	46.2
3	1	36
4	1	47.3
5	1	40.8
6	1	30.9

Examine the complete dataset in Excel file: Silt Content of Soils.xlsx