



# Practical 1:

## Data summary and presentation



## Practical 1 - Data summary and presentation

1. The data for this practical exercise consists of 99 measurements of the silicon content of the hot metal drawn from a blast furnace. We are interested in ways of simply investigating the characteristics of the data before embarking on any more complicated analysis.

We can get a simple idea of the distribution of the data by drawing up a frequency table. This could be done by hand but it is obviously much more efficient to use a statistical package. We are going to use the statistical package GenStat.

The data are stored in **Blast** sheet, contained in the **practicals.xls** file. In GenStat, **click** on the **Excel Import Wizard** icon on the toolbar, go to the directory C:\USER, select **practicals.xls**, then the **Blast** sheet and **click Finish**. This opens a spreadsheet containing the data. By default GenStat will automatically update this data onto the server and give a brief summary of the data in the output window.

The data are stored in column 1, named **silicon**. The graphics menu may be used to give a visual summary of the data.

- Select **Graphics** → **Histogram**
- Specify **silicon** is to be plotted by putting it in the **Data** list
- Select **Next**. This allows various attributes to be set
- Select the **Options** tab and then **Fixed Bar Width** checkbox in the bottom left hand corner of the dialog box, so that any histogram produced has such bars. Also, select **Use data values** for defining **Boundaries**. This latter option specifies that GenStat is to decide what bars are used, etc
- **Click Finish**

The histogram produced is not unreasonable. A common problem is too few or many bars. You may alter this number in the dialog box.

The graphics window also offers tools that you may use to edit your plot interactively. By a simple **double-click**, you could label your axes and title your graph at this point if you had not done it earlier.

To produce a grouped frequency distribution, produce a histogram as before, but in the **Options** change the **Boundaries** from **Use data values** to **Limits**. In the latter box type **100 150 200**. **Click Finish**. Do you understand what has been produced and why?

Another simple descriptive plot which you might find useful is a boxplot of the data.

The sequence **Graphics** → **Boxplot** and selection of **silicon** will generate a boxplot. Can you interpret the plot?



2. Let us look at how GenStat will give you summary statistics for large sets of data which would be too big to input into your calculator.

Try the sequence **Stats** → **Summary Statistics** → **Summarize Contents of Variates** and select **silicon**, **check** the statistics you wish to have then **click Run**. Can you interpret the output?

(Note that GenStat also gives you the option of obtaining a histogram and a boxplot from this dialog box).

Look at your first histogram of silicon, again. Can you estimate what proportion of observations are outside the range  $\pm 2$  standard deviations from the mean?

How could you calculate the exact proportion?

3. Move to the spreadsheet window. Use **Spread** → **Sort ..** and select **silicon** into the **Sort on column** box, to sort the dataset in ascending order. You can then scroll through using the row numbers to read off the median value and the quartiles. Check your answers by looking back at the descriptive statistics produced earlier.

Close the spreadsheet containing the silicon data. Now clear all data from GenStat by the following sequence: → **Data** → **Clear All Data** and answer **Yes**. It is always advisable to do this if you intend to work on a different dataset in the next session.

#### IF YOU HAVE TIME, TRY ALSO THE FOLLOWING

4. Ten random samples of wire from manufacturers A, B, C and D, were assessed for strength.

The resulting dataset is stored in the sheet **Wire** of **practicals.xls**. Import this sheet in GenStat by **clicking** on the **Excel Import Wizard** icon on the standard toolbar. You will see data from the four manufacturers in columns a, b, c, and d respectively. Make the **which** column a factor.



Summary statistics for all 4 columns may be obtained using **Stats → Summary Statistics → Summarize Contents of Variates**. Select all four variables and the required statistics and **Click on Run** when done.

To compare the strengths of wire produced by the four manufacturers, you can also try a boxplot. Remember you could get the four boxplots for **a**, ..., **d** individually produced by ticking the graphics section of the **Summarize Contents of Variates** dialog box. These being separate implies the scales are not comparable. To use the same scales, the data need to be in a single column with an accompanying column of values indicating which manufacturer they belong to. These are provided as **abcd** and **which** respectively. In GenStat jargon **which** is a **factor**.

Now you can obtain the four boxplots together using **Graphics → Boxplot**. Select **abcd** as the data to be plotted. Choose the appropriate arrangement of the data, i.e. **Single variate with groups**, and enter the factor **which** in the **Groups** box. Finally **click Finish**.

What do you conclude?