## Creative Coding 2

## Introduction to Histograms and Frequency Distributions

This lab is concerned with the organisation and presentation of ‘numeric (univariate)’ data. It describes how numeric data can be organised into frequency distributions of various types and their graphical presentation (histograms).

Raw data is data obtained from a statistical survey or database which has not yet been worked upon. Generally raw data as they stand yield little or no information as it is impossible to glean information from a mass of data.

**Histogram:**

* A histogram is a plot of the number of observations in a particular interval.
* Histograms allow us to observe the general shape (or distribution) of the data and identify if any extreme points exist in the data set, which many otherwise remain hidden amongst the large amount of raw numbers.
* They are generally used to show continuous univariate data (data with one variable).

**Outliers:**

The extreme points mentioned above (also known as ***outliers***) may be the result of an error in recording a value, equipment malfunction etc., or may indicate the presence of special variation in our data set.

Whatever the reason, extreme or unusual data points should be investigated and a decision taken whether to include them or not in the calculation of summary statistics.

1. **Simple Frequency Distributions:**

Some sets of raw data contain a limited number of data values, even though there may be many occurrences of each value. In this type of situation, the standard form into which the data is organised is known as a ***simple frequency distribution***. A simple frequency distribution consists of a list of data values, each showing the number of items having that value (called the frequency). This type of structure is normally applicable to ***discrete*** ***raw data*** (i.e. where values have usually been obtained by counting), since data values are quite likely to be repeated many times. This simple distribution is not normally suitable for ***continuous data*** (i.e. where values have been measured), since the likelihood of repeated values is small.

**Example:**

Consider the following data on the number of days missed last month for each of the 42 employees in a company:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 3 | 2 | 0 | 2 | 0 |
| 5 | 2 | 4 | 0 | 0 | 2 | 4 |
| 3 | 0 | 0 | 2 | 1 | 3 | 6 |
| 2 | 2 | 2 | 1 | 0 | 0 | 1 |
| 2 | 1 | 2 | 3 | 2 | 2 | 1 |
| 1 | 2 | 0 | 4 | 3 | 0 | 3 |

**First Reaction:**

The first information you should look for is the smallest and largest values which are 0 days and 6 days. These values are calculated in Excel using the MAX and MIN functions.

**Forming a Simple Frequency Distribution:**

When the data contains a limited number of distinct data values with perhaps many occurrences of each value then we have already established that the data is arranged as a simple frequency distribution. A simple frequency distribution is a table listing the possible data values and the number of times each one occurred.

A ***simple frequency distribution*** consists of a list of data values, each showing the number of items having that value (called the frequency). This type of structure is normally applicable to ***discrete*** ***raw data*** (i.e. where values have usually been obtained by counting), since data values are quite likely to be repeated many times. This simple distribution is not normally suitable for ***continuous data*** (i.e. where values have been measured), since the likelihood of repeated values is small.

**Example:**

Consider again the table above. Here there are 7 different values present 0, 1, 2, 3, 4, 5 and 6. A simple frequency distribution involves counting the number of times each of the values occur. The resulting simple frequency distribution table is presented as follows:

**Number of Days missed last month for the 42 employees:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Number of Days missed | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Number of Employees | 10 | 8 | 13 | 6 | 3 | 1 | 1 |

In this case, ‘number of days missed’ is describing the data being investigated, and ‘number of employees’ is the frequency of occurrence of each value. When we add the frequencies together (the 2nd row of values in the table above), we should get the total number of employees (i.e. 42 in this case).

The main aim of a frequency distribution table is to summarise numeric data in a logical manner that enables an overall perspective of the data to be obtained quickly and easily.

What happens when the number of distinct values is large, greater than 20 say? We will return to this problem later.

**Using Excel:**

We will be using Excel to construct most frequency distribution tables. For simple frequency distributions, this will be relatively straightforward. It will involve the COUNTIF function. This is a function used to count the number of occurrences of a particular value. (**Exce**l **definition:** counts the number of cells within a range that meet the given condition)

The screen shot below (Figure 1) shows the output from the first question:

1. Type the heading:

**No of days missed**

Underneath, type in the values **0** to **6** – we want to count the occurrence of each of these values. These values are got by observing the particular table.

2. Then in the adjacent column, type:

**Frequency (of workers)**

In the cell underneath, type in the following:

=COUNTIF(‘choose all the data from the table’, click on the cell with 0 in it’). Press ‘Enter’.

3. Now we can use the autofill function, but before this we have to insert dollar signs into the range – see formula below (e.g. =COUNTIF($A$3:$G$8,A14)). This is because we want the range chosen to remain fixed. By default, autofil*l* uses relative, not absolute values, unless it sees the dollar signs. We leave the second parameter as it is (why?) and press ‘Enter’.

Now use autofill - select the cell with the formula, move the mouse to the bottom right hand corner. The cursor should change into a black cross. We now drag the mouse down the required number of cells, and choose ‘***Fill without formatting’***.

**Histograms:**

A histogram is a vertical bar chart (with no gaps between bars) used to display a simple or grouped frequency distribution (In Excel it is called a column chart).

As with all graphs the chart must have a title and the horizontal and vertical axes must be labelled.

**Example:**

Plot a histogram for the simple frequency distribution formed earlier based on the number of days missed of a sample of 42 workers.

**Construction of Histogram in Excel:**

1. Begin by showing the simple or grouped frequency distribution you wish to plot in a clear area of the sheet.
2. Select the range of data you want to use.
3. Select the ‘***column chart’*** option and the first option under ‘***2d column***’.
4. At this stage, you will probably have 2 series – shown in the legend.
5. Remove one of these series by right clicking on the chart and selecting the ‘***select*** ***data’*** option from the drop down menu (remember, if you have a 1-1 relationship between x axis and y axis data (i.e. only one series), then you do not need a legend).
6. A dialog box should appear – on the lhs, select the series you want to remove, and click the ***‘remove’*** button (in this example, it would be the ‘***No of days missed***’’ data). The series ‘***Freq (no of workers)’***’ should be left as is).
7. Now you have only one set of columns (one colour) as required – note: always check the data at this stage to see that the column heights correspond with the data in the original table.
8. While the dialog box is still open, check your horizontal axis labels – if they don’t correspond to the data in the table, click edit (on the rhs of dialog box), and select the correct data (all in one go). Click ok.
9. Now select the ‘***legend***’ option and get rid of it.
10. Enter the appropriate chart title, x-axis title and y-axis label. Display the axes and hide the gridlines. Do not show data table or data labels.
11. In the final step leave the default setting to have the chart placed as an object in the current sheet.
12. Finally the minor details; There will be gaps between the bars which are removed by selecting any bar and choosing ‘***Format Data Series’***, and then selecting the ‘***Series Options’*** option. Now reduce the gap width to zero. Click ok.

Graphical user interface, application, table, Excel

Description automatically generated

Figure 1

1. **Grouped Frequency Distributions:**

**Example:**

Consider the following data on the number of miles travelled recorded by 80 different sales people over a certain week.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 454 | 431 | 495 | 461 | 487 | 482 | 485 | 490 |
| 473 | 493 | 464 | 463 | 447 | 418 | 468 | 461 |
| 479 | 493 | 431 | 473 | 480 | 468 | 419 | 477 |
| 476 | 442 | 451 | 496 | 406 | 468 | 471 | 440 |
| 478 | 465 | 465 | 474 | 435 | 417 | 407 | 453 |
| 472 | 416 | 439 | 499 | 465 | 428 | 449 | 401 |
| 497 | 433 | 461 | 459 | 492 | 499 | 484 | 442 |
| 460 | 417 | 486 | 476 | 494 | 414 | 498 | 485 |
| 433 | 492 | 458 | 412 | 407 | 488 | 445 | 415 |
| 471 | 401 | 469 | 486 | 406 | 462 | 447 | 427 |

This data contains a large number of different values. A simple frequency distribution table would have too much detail. In this case a grouped frequency distribution table will be used where the data is organised into intervals, groups or classes of appropriate width, and the number of values in each group is counted.

**Note:** Each cell in the table above represents a sales person.

**For example:**

Sales person 1 (first cell) records 454 miles over this week

Sales person 2 (second cell across) records 431 miles over this week

Etc........

**Characteristics of Grouped Frequency Distributions**:

* The number of values in each interval is known as the frequency of the interval.
* The intervals are normally of equal width. The width is defined as the difference between the two end points of the group.
* All values must be represented within one and only one interval. ***Overlapping intervals are forbidden.***
* Ideally there should be between 6 and 10 intervals.
* Data are usually arranged in classes of the form 10 to 19, 20 to 29 etc.
* The end points of the intervals are known as the class limits.

**Note:**

Once a distribution adheres to the basic rules (i.e. no overlapping intervals etc.), there is no absolute right or wrong choice re number of groups/interval width. However too few intervals and you will not have a picture of the shape, too many and the display will be too fragmented to show an overall shape.

**Compiling Grouped Frequency Distributions from a set of raw data**

There are 4 steps:

1. Find the number of intervals.
2. Calculate the width of each interval.
3. Construct the intervals.
4. Count the number of values in each interval.
5. **Find the number of intervals:**

There are several ways of determining the number of intervals. We will look at one commonly used technique.

1. Calculate the end points of the interval 2i-1 to 2i for i ranging from 1 to the number of data points - which we will call n.
2. This calculation continues until the number of points in the data set is enclosed in the interval and the corresponding value of i gives the number of intervals required for the histogram.

Table 1

|  |  |  |
| --- | --- | --- |
| **i** | **2^(i-1)** | **2^i** |
| 1 | 1 | 2 |
| 2 | 2 | 4 |
| 3 | 4 | 8 |
| 4 | 8 | 16 |
| 5 | 16 | 32 |
| 6 | 32 | 64 |
| 7 | 64 | 128 |

Using the data from this Example, the number of data points is 80.

If we calculate the intervals 2i-1 to 2i for i = 1, 2, 3 etc. until 80 is enclosed within one of the intervals, we arrive at the interval 64 to 128 (see Table 1).

The corresponding value of i is 7 and this is the number of intervals required for the histogram.

For i = 7, 2i-1 =64 and 2i = 128.

1. **The width of each interval is calculated as follows:**
2. Calculate the range of the interval = Maximum - Minimum
3. Divide the range of the data set by the number of intervals.
4. The nearest whole number from this calculation is the interval width for each bar of the histogram.

For this example the range of the data set is 499 – 401 = 98 and the number of intervals is 7.

The interval width is therefore 98/7= 14.

The histogram will then consist of 7 intervals each of width 14 miles.

1. **Intervals are constructed as follows:**

Start the first interval with the lowest value, and make sure that the highest value is in the last group.

In this example, the first interval ranges from 401 miles (the lowest value) to 415 and so on.

Remember that the width is defined as the difference between the two end points of the interval.

Label the column accordingly. For example:

Table 2



1. **Count how many values are in each class using Excel.**

You can use the FREQUENCY function – for an explanation, click on the fx button on the toolbar. This function returns the number of values in different ranges of values.

**Steps:**

1. In the column adjacent to the ‘No of Miles’ column (Table 2), type the values 415 to 505. ***This column is only used in conjunction with the frequency function. It has no other purpose.***
2. In the next column, type the heading “Frequency (No. of sales people)”.
3. Underneath this heading, select ***all the blank cells*** in the column. This is because your result is going to be in the form of an array of values.
4. In the formula bar, type “=FREQUENCY(select the data, select the column with the values 415 to 505)”. These are the two parameters that this function requires.
5. Finally close the parentheses (bracket) ***BUT now instead of pressing enter, press “*Ctrl+Shift+Enter”**. This is to show Excel that you want the function to count an array of values. i.e. all the values up to and including 415 + (simultaneously) all the values up to and including 430 but not including 415 + (simultaneously) all the values up to and including 445 but not including 430 etc.

Table 3



The associated histogram is below – Figure 2:

Figure 2