Module 2: Introduction to Quantitative Data Analysis

Antony Fielding[1]
University of Birmingham & Centre for Multilevel Modelling

Rebecca Pillinger
Centre for Multilevel Modelling

Contents

[1] With contributed material from Kelvyn Jones and Fiona Steele and extensive comment by Harvey Goldstein.

Some of the sections within this module have online quizzes for you to test your understanding. To find the quizzes:

EXAMPLE

From within the LEMMA learning environment

- Go down to the Lesson for **Module 2: Introduction to Quantitative Data Analysis**
- Click"2.1 Univariate Data Summary" to open Lesson 2.1
- Click Q1(i) to open the first question

Introduction

The aim of this module is to give an overall view of some the principles of effective data analysis. The focus is on how we summarise data to uncover patterns and relationships between variables, and how these relationships can begin to explain the values of the variables that we observe. Some key statistical vocabulary is introduced and concepts are illustrated by example. You will learn about the following:

- Ways of summarising the shape and pattern of the values of a single variable, at different levels of measurement.
- Understanding the role of variability in comparative analysis and the study of relationships.
- The elaboration of relationships, the importance of control variables, and the concepts of confounding, suppression and interaction.
- The essential parts of a statistical model: pattern and residual variation.
- Key concepts in inference from samples.

C2.1 Univariate Data Summary

Before we decide on appropriate methods of analysis to address our research question we will usually want to 'get to know' our data. At the very minimum we will want to establish that the variables we plan to use do indeed vary. We will also want to establish an initial picture of the nature of that variation, since it is the study of such variation that will form the basis of further analysis. Thus the first step in any quantitative investigation is to separately summarise each particular variable in various informative ways. Since these initial summaries deal with only one variable they will be called *univariate analyses*. For instance in our research we might be contemplating using the variable Years of Education, from the European Social Survey data introduced in Module 1. It will be initially useful to know something about the *distribution* of respondents over the values of this variable. What is the general level of education? How widely dispersed are the values? Is there a concentration of a large number of units (individuals) at particular points? Are there units with extreme values in our data set? Are there curious or inexplicable values which may lead us to suspect errors? In the case of nominal variables, such as Marital Status or Ethnic Group, or of ordinal variables, such as Income measured by twelve ranges, how many units fall into each category? The answers to these questions will affect the analyses we perform.

We should also re-emphasise the major point discussed in Module 1 that the level of measurement of a variable may restrict what summaries are appropriate.

C2.1.1 Frequency distributions

*Frequency distribution*s are perhaps the most commonly used initial summaries. They count how many in the set of units under consideration have different values, or groups of values, of the variable. Frequency distributions can usually be employed whatever the level of measurement. However, there are a number of conventions which influence different ways of grouping values.

C2.1.1.1 Categorical data

Table 2.1 and Table 2.2 represent typical summaries for the categories of a nominal variable, Marital Status, and an ordered variable, Education Level (defined as the highest level of education attained).

Table 2.1. Frequency distribution of Marital Status

	Number	*% of all cases*	*% of valid cases*
Married	22974	54.2	54.5
Separated	627	1.5	1.5
Divorced	2758	6.5	6.5
Widowed	3836	9.1	9.1
Never married	11946	28.2	28.3
Refusal	102	0.2	-
Don't know	32	0.1	-
No answer	84	0.2	-
Total	42359	100.0	100.0

Table 2.2 Frequency distribution of Education Level

	Number	*% of all cases*	*% of valid cases*	*Cumulative % of valid cases*
Not completed primary education	1,684	4.0	4.0	4.0
Primary or first stage of basic	5,634	13.3	13.4	17.4
Lower secondary or second stage of basic	9,610	22.7	22.9	40.1
Upper secondary	13,730	32.4	32.7	72.8
Post secondary, non-tertiary	3,517	8.3	8.4	81.2
First stage of tertiary	5,448	12.9	13.0	94.2
Second stage of tertiary	2,419	5.7	5.8	100.0
Refusal	43	0.1	-	
Don't know	153	0.4	-	
No answer	121	0.3	-	
Total	42359	100.0	100.0	

These tables show the number of respondents in a subsample from the European Social Survey (ESS) data that fall into each of the categories of the variable.[2] There are also a number of categories relating to different sorts of 'missing data'. Handling such missing data is a complex problem beyond the scope of this module, but it is important to recognise their existence. When they are removed we have what are usually termed *valid cases* for further analysis. For the rest of this section we will focus only on summarising valid cases. The tables also show percentage distributions. Percentages are usually easier to interpret because of their familiar feel, and because they show us the relative frequency of the values of the variable in the data. This enables us to better judge the relative importance of the values. In Table 2.1 we can see, for instance, that over half of our subsample is married. In Table 2.2 we can see that there are very few cases in the first and last valid categories. Since the ordering of the categories has meaning here, we note that the extremes of Education Level contain relatively small numbers. The most frequently occurring categories, 'married' in the first case and 'upper secondary' in the second, are called the *modes* of the distributions. There is an extra column called the *cumulative frequency* in Table 2.2 that is appropriate for ordinal (or grouped ratio or interval level) variables, and which gives the percentage of units with a value of the variable less than or equal to the value of the row in question. So, for example, 72.8% achieve secondary education or lower or, to look at it another way, 27.2% go beyond secondary education. A cumulative summary in the opposite direction would produce figures like the latter[3], showing the percentage of units with a value of the variable greater than or equal to the value of the row in question.

It is often easier to pick out key features of the data by visualising them. Figure 2.1 is a *pie-chart* for the distribution of marital status where the angles (and thus the sizes) of the slices of the pie are in proportion to the relative frequencies in the last column of Table 2.1. The pie chart is perhaps the most common way of visualising distributions for categorical data but *bar charts* are sometimes also used.

[2] This is the subsample of cases made available on the ESS website for on-line analysis at http://essedunet.nsd.uib.no/cms/topics/1/.

[3] but shifted down a row.

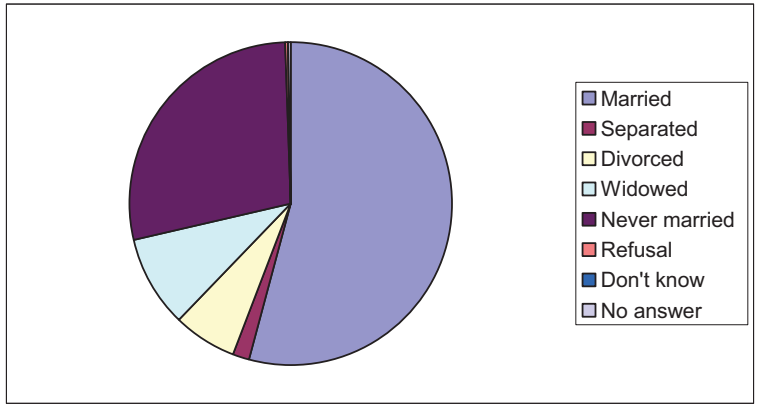

Figure 2.1: Pie chart of distribution of Marital Status

C2.1.1.2 Continuous data

There are a few new issues connected with frequency distributions for continuous variables since to display patterns we must often group values in a meaningful way. In the ESS there is a variable defined by answers to the question 'How long have you lived in this area?', which we will call Years of Residence. Data are recorded to the nearest whole year and the set of possible values are integers between zero and ninety eight, so only a small percentage of units will take each particular value. A full frequency distribution with ninety nine groups would be too detailed, patterns would be difficult to discern and the table would not be very informative. To avoid these problems the values must be grouped into classes or intervals. In determining the appropriate number of intervals and their widths we need to strike a balance between too many intervals (which leads to groups with relatively few cases in each) or too few intervals (which results in too great a loss of information). There are no hard and fast rules for this. One possible set of intervals for Years of Residence is shown in Table 2.3. Here the intervals are of an equal width of ten years, and are perhaps easier to handle when this is the case, but as will be seen in another example later in this section there is no necessity that this should be so.

Table 2.3 Grouped frequency distribution of approximate Years of Residence in area at interview (valid cases only)

Whole year range	Number of valid cases	Percentage	Cumulative Percentage
0-9	11606	27.8	27.8
10-19	8667	20.7	48.5
20-29	7470	17.9	66.4
30-39	5400	12.9	79.3
40-49	3567	8.5	87.8
50-59	2369	5.7	93.5
60-69	1381	3.3	96.8
70-79	987	2.4	99.2
80-89	247	0.6	99.8
90-99	80	0.2	100.0
Total	41774	100	

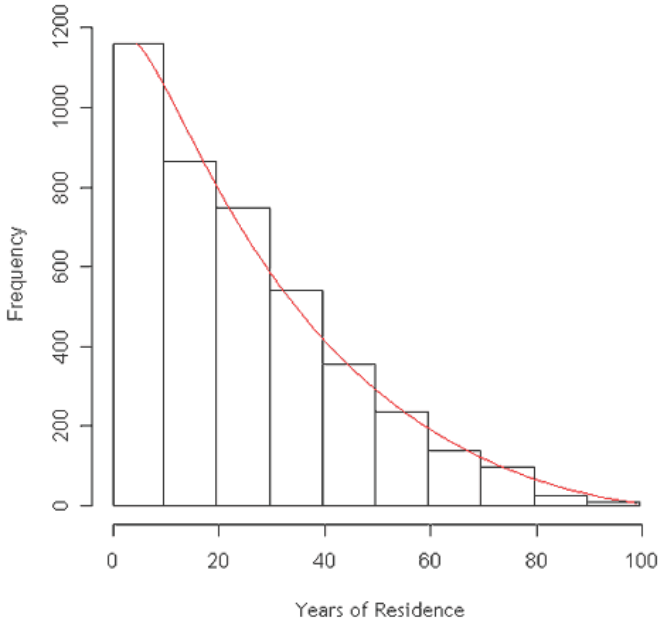

Figure 2.2 Histogram of approximate Years of Residence in area at interview (valid cases only)

There are also useful ways of presenting such frequency distributions of continuous

variables graphically. The most common are *histograms* although other devices such as *stem and leaf plots* (for small datasets) or *box and whisker plots* (also called boxplots) may be employed. Figure 2.2 depicts a histogram for the distribution of Years of Residence in Table 2.3. Each rectangle here corresponds to one of the intervals, with the horizontal width of the rectangle proportional to the width of that interval. The areas of the rectangles are proportional to the numbers (or percentages) in each range. It is important to note that the construction of a histogram is governed by the areas of the rectangles, not their heights. Only if the intervals are of equal width, as they are here, are the two equivalent. The areas give a proper visual interpretation of the relative sizes of these ranges. In the diagram an approximate smooth curve has also been drawn to connect the tops of the rectangles. This is called a *frequency curve* and gives us an impression of the shape of the distribution. Here we have an example of an asymmetric distribution with a long tail, called a *skewed distribution*. Almost 50% of lengths of residence are under 20 years and after that point the number of units in each range decreases gradually as the length of residence increases.[4] The long tail is therefore to the right and so this is an example of *positive skew*. If the long tail had been to the left, i.e. towards smaller values along the continuous range, we would refer to the distribution as *negatively skewed*.

Sometimes the raw data provided to the analyst is already grouped into intervals of values, so that there is less flexibility and the only decision then might be whether to summarise further by aggregating intervals. One such example is the Income variable, recorded from questionnaires in twelve ranges of values and coded 1-12 in our example dataset. The frequency distribution of valid cases is displayed in Table 2.4. The widths of the ranges here are not equal but are wider for ranges at the higher end of the income scale.

[4] More detailed examination of the ungrouped data would reveal that the 11606 cases in the first interval are almost uniformly spread over each of the separate single values in that group.

This document is only the first few pages of the full version.

To see the complete document please go to learning materials and register:

http://www.cmm.bris.ac.uk/lemma

The course is completely free. We ask for a few details about yourself for our research purposes only. We will not give any details to any other organisation unless it is with your express permission.