Chapter 4

Discrete data

4.1 Discrete data, barplots and histograms

Histograms are appropriate for the presentation of continuous data. Such data are usually contrasted with discrete data which, at their simplest are counted data in different and disjoint categories. An example would be counts of distinct vessels by type in an assemblage of pot or glass. Most archaeologists will be familiar with the presentation of such data in the form of pie-charts or bar-charts (barplots). These are among the most visible of statistical methods used in the archeological literature; they are sometimes over-used, used unnecessarily, or misused (despite their apparent simplicity).

Discrete data may be ordinal, in that categories have a natural ordering but the 'distance' between categories is not known. Common archaeological examples may involve chronology; for example, counts of a single artifact type ordered chronologically on the basis of stratigraphy or phasing, without knowing the absolute chronology. Whether or not data are ordinal has implications for graphical presentation, and for the choice of analytical method.

Barplots (and pie-charts) for a single set of counts, possibly expressed as percentages, are often pretty boring. Unless there are a large number of categories, looking at the numbers in a table is often all that is really needed. Things get more interesting when tables of counted data arise from data analysis. If, for example, counts of different artifact types are available for a single context a simple barplot will do. If such data are available for different contexts they can be expressed in tabular form (examples follow) and questions can then be asked about the similarity of contexts in terms of the artifact assemblages that characterize them, or the similariy of artifacts in terms of their distribution across contexts.

Data in such a form are variously referred to as *cross-tabulated*, *cross-classified*, or *contingency tables*. Compared to data matrices for continuous data, where rows

and columns (cases and variables) have a different 'status', in contingency tables the rows and columns have a similar status and a different notation is used here to reflect this. In general we refer to an $I \times J$ contingency table. Rows and columns can be interchanged, though in practice the emphasis may be on one or the other.

To herald later comment, it is convenient to focus on the difference between histograms and barplots. The examples in Figure 4.1 use the weights of loomweights data, from Tables B.3 and B.4. The lower-right plot shows the default `R` histogram using the `hist` function. Note that the bars associated with the bins 'touch' each other. Histograms are sometimes referred to as barplots or bar-charts in the archaeological literature and *vice-versa*. This is possibly understandable since the histogram is represented by bars whose *area* corresponds to the counts within bins, but should probably be regarded as incorrect.

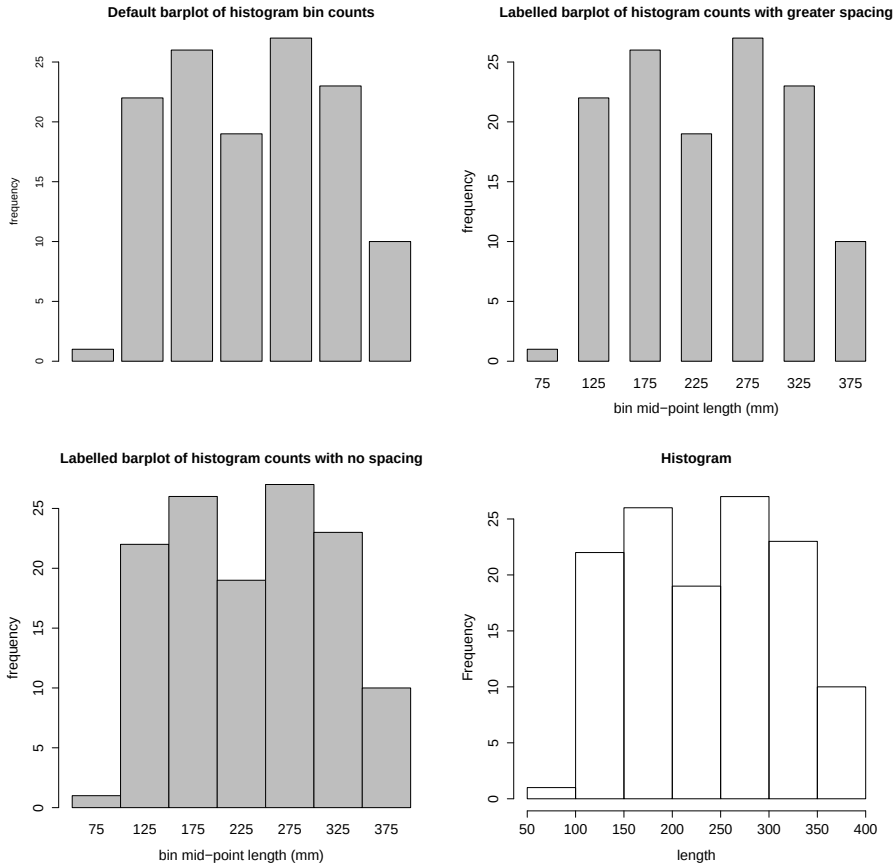

Figure 4.1: *Right and wrong ways of presenting a histogram, based on counts for bins in the default* `R` *histogram for the reduced loomweight weight data of Tables B.3 and B.4 – (1, 22, 26, 19, 27, 23, 10).*

It is easy enough to get the histogram if all the data are available. What if only counts within bins are given, in this instance (1, 22, 26, 19, 27, 23, 10)? It is impossible to tell by looking at the numbers alone whether they represent continuous or discrete data so the context, which would include the intervals, continuous or (usually) integers for discrete data, needs to be examined.

The upper plots incorrectly show the data as barplots; the gaps between bars indicate quite clearly (at least to my eye) that the data are discrete. This misuse is not uncommon. In `R` the `barplot` function allows control over the spacing used between bars and setting this to zero produces the histogram shown in the lower-left plot.

This kind of graphical/terminological misuse to be found in the literature is possibly not misinterpreted much, but may betray confusion about the distinctions involved, or a lack of adequate software. (The problem was common some years ago but may now be less prevalent because popular and widely used non-statistical software has belatedly caught up with the issue.)

Other issues concerning the use of barplots are illustrated later, but first a more interesting example is examined. The data relate to Roman pillar-moulded bowls found in excavations at Colchester. Eighteen periods, with a chronological sequence, were identified. Data are given in Table 4.1.

In the absence of knowledge of the date span of the periods it is legitimate to represent the data in the form of a barplot[1]. This is done in the top plot in Figure 4.2. On a technical note the width of a bar is arbitrary. What is shown is conventional, but vertical lines with heights corresponding to percentages would be as legitimate.

We can do better because information on the date span of the periods, in terms of actual dates and the width of the spans, is available. The earlier periods are more tightly defined than later ones. Pillar-moulded bowls are an early form and occur predominantly in the earlier periods. Knowing the dates and spans means that the data can be treated as continuous and represented in the form of a histogram. This is done in the right-hand plot in the figure, using period rather than mid-points to label the scale.

There is a complication. Most software for histograms produces equal bin-widths by default. Sarkar (2008: 39) goes so far as to say that unequal bin-widths are 'rarely used outside introductory statistics textbooks'. This, and others in Cool and Price (1995), is a counter-example.

How the histogram was produced is discussed in Section 4.4. Briefly, the `barplot` function was used, with zero spacing between bars, and widths of bars

[1]The date span is being ignored here; so the sequence is ordered but the date-span of categories is not known. This is an example of *ordinal* data; this is quite common with chronological data – the situation here where the date-span can be specified is less often seen.

Period	Date	Width	Midpoint	%
I	43-50	8	46	11
II	51-60	10	55	22
III	61-70	10	65	9
IV	71-80	10	75	8
V	81-90	10	85	6
VI	91-100	10	95	5
VII	101-125	25	112	9
VIII	126-150	25	137	5
IX	151-175	25	162	4
X	176-200	25	187	3
XI	201-225	25	212	2
XII	226-250	25	237	2
XIII	251-275	25	262	2
XIV	276-300	25	287	3
XV	301-325	25	312	2
XVI	326-350	25	337	1
XVII	351-375	25	362	1
XVIII	376-400	25	387	1

Table 4.1: *Chronology of Roman glass pillar-moulded bowls found during excavations at Colchester, 1971-1985. The data are ordinal, but the periods are of different lengths. See Cool and Price (1995: 15–19) and the text for discussion.*

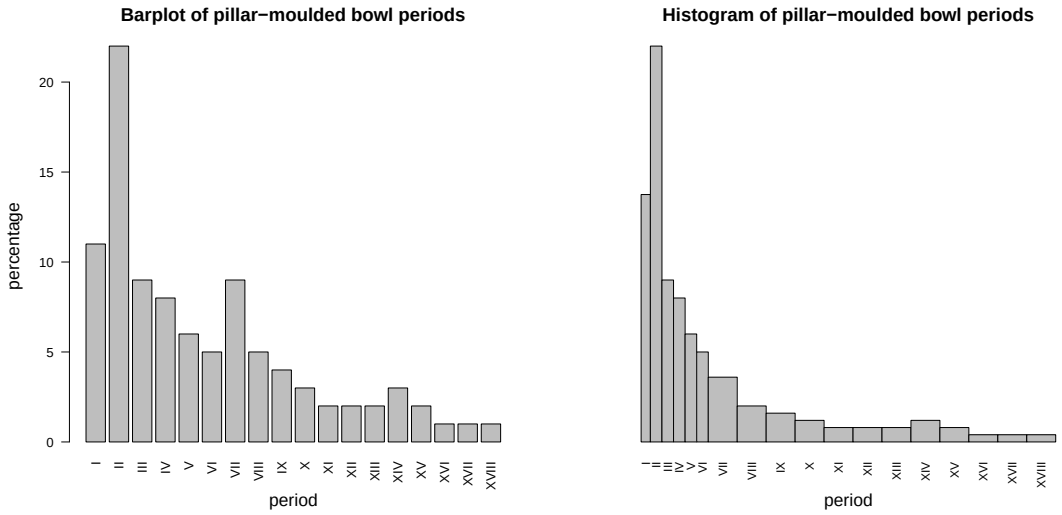

Figure 4.2: *Different ways of representing the data from Table 4.1. The barplot to the left respects the ordinal nature of the data, but not the fact that periods are of different lengths. The histogram to the right, constructed using the* `barplot` *function, does respect the differing lengths. Because of the unequal bin-widths a vertical scale is not appropriate.*

corresponding to the date span specified. Because of the unequal widths a sensible scale for the vertical axis is not available.

The advantage of the histogram compared to the barplot is that it emphasizes the decline in use over time more obviously. The quite sharp and steady decline in usage is readily apparent. The barplot does not do this, because it was constructed without use of the spans which are longer for the later periods. There is the suggestion in the barplot of a secondary peak in period VII that exists simply because the span is longer than earlier periods and, for the same reason, the sharp decline in usage over time is less apparent.

4.2 Barplots for two-way tables

The barplots in Figures 4.1 and 4.2 provide examples of what barplots for single sets of counts look like. This section deals with two-way tables of counts. For illustration, data adapted from Table 5 of Bailey *et al.* (1983) are used, showing the distribution by stratum of four classes of artifacts from excavations of the palaeolithic rockshelter of Kastritsa in north-west Greece. It is assumed that interest lies in a comparison of the distribution across strata of artifact types, expressed in Table 4.2 as percentages. The authors did not present analyses of the kind to follow – they are here to demonstrate 'technique' and coding.

Stratum	I	II	III	IV
1	13	7	21	59
3	12	9	19	60
5	17	14	16	53
7	16	14	19	52
9	11	9	8	72

Table 4.2: *Data on the distribution of artifact classes by strata (row percentages) from Table 5 of Bailey et al. (1983). The classes I–IV are cores, utilized flakes, unmodified flakes and waste.*

Figure 4.3 shows the types of plot available; to the left *stacked* barplots, and to the right *clustered* barplots. Choice of orientation is arbitrary. My impression is that stacked barplots are more prevalent than clustered barplots in the literature. Clearly Class IV, waste, dominates and is fairly similar in most strata, apart from Stratum 9 where there is a noticeable increase. As a generalization, with the exception of Stratum 9, the relative importance of other classes is mostly III, I and II. Re-ordering the first two columns of data would make this, and minor exceptions, even more clear. The sample sizes for Strata 7 and 9 are much smaller than for other strata.

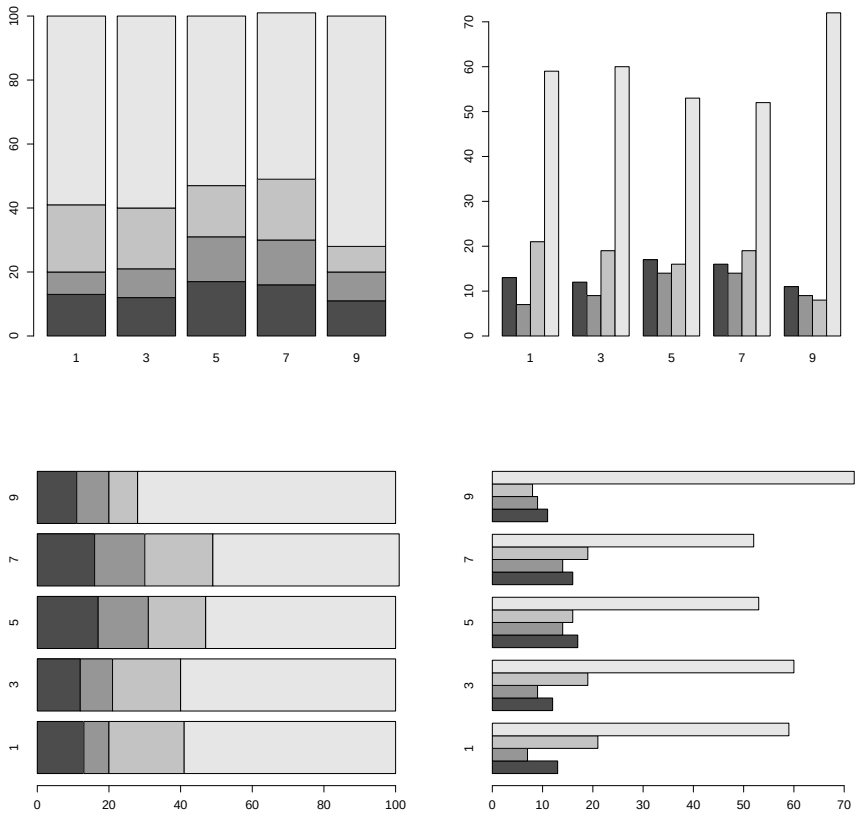

Figure 4.3: *Unannotated barplots for the data of Table 4.2; enhanced versions of the upper two charts are shown in Figure 4.4. Plots to the left are stacked and those to the right clustered; plots at the top and bottom are vertically and horizontally orientated respectively.*

Sarkar (2008: 61) suggests that the stacked barplot has limitations if one is interested in comparing proportions; if patterns are obvious a table will do; if not then making the necessary judgments may not be easy. In the barplots shown the fact that one class is dominant, and differentially so among strata, means that the less common classes are differentially 'squashed' at the bottom of the plot (in this example) so comparison across bars is not straightforward. The clustered barplot does a better job, since the dominant class can be mentally discounted and comparison among other classes is easier. If these are the main focus of interest the plot can always be produced omitting the dominant class.

By way of illustrating examples of barplot presentation that occur commonly but could be considered 'wanting', Figures 4.4 and 4.5 show, respectively, stacked

and clustered barplots from R and two versions of stacked barplots from Excel.

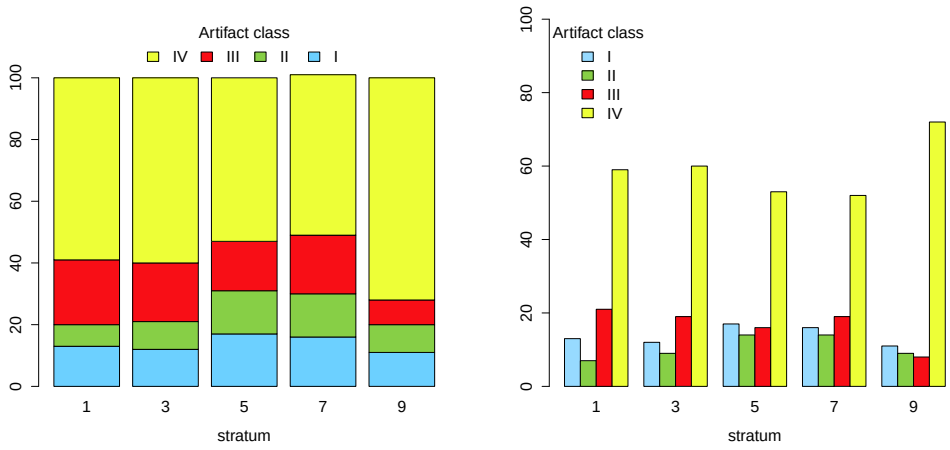

Figure 4.4: *Enhanced stacked and clustered barplots for the data of Table 4.2.*

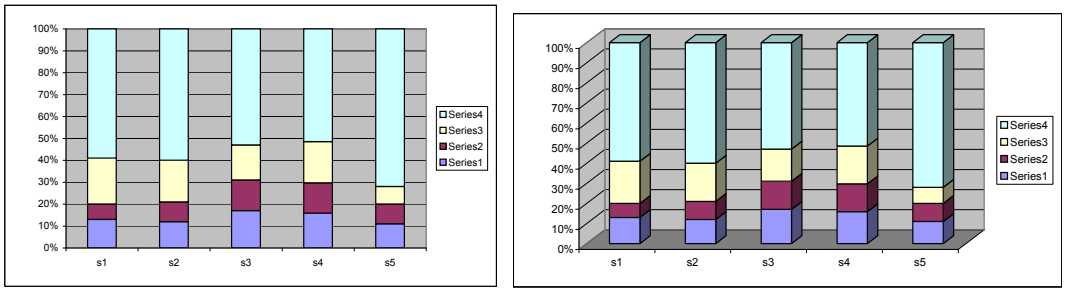

Figure 4.5: *Two- and three-dimensional* Excel *stacked barplots.*

There are several problems with the Excel barplots, which are defaults. In the two-dimensional plot the bars that should be there are innocent; the grid lines and background shading are the culprits. The lines are too prominent and there are too many of them. Coupled with the background the gridded parts also look like bars, and leap to the foreground if you stare long enough, distracting from the message of the plot. The three-dimensional plot is much worse. As with three-dimensional pie-charts, the third-dimension does not exist and should not be there. It adds, along with the grid and the 'three-dimensional' frame employed, to a variety of optical illusions. Even with the grid it is difficult to read off values from the scale. Such plots should not be allowed. If grids must be used they should be fewer than in the examples here and less obtrusive.

Barplots, of whatever variety, have probably been overused. Have users been seduced by myths such as 'every picture tells a story' or 'a picture paints a thousand

words'? Tables tell stories too; in the context of tables that might be presented as barplots they require nothing like a thousand words; typically occupy less space than a graphic with commentary; and retain more precise information about actual numbers and their difference. I have particularly in mind single sets of counts with few categories (I've seen a barplot presented with just one category) and fairly small contingency tables. To see patterns, reordering of categories for both tables and plots, if the data are not ordinal, may be useful. With large numbers of categories barplots may be difficult to read, and correspondence analysis (Chapter 9) represents an alternative method of presentation.

Other methods of analyzing discrete data exist, attracting different degrees of archaeological attention. For moderate to large tables correspondence analysis is widely used as a graphical method of presentation (Chapter 9). A common method of assessing whether there is a significant association between the rows and columns of a table is the chi-squared test (Section 12.3.3). More formal modeling methods, *log-linear models*, analogous to the use of regression models for continuous data (Chapter 5) are available. Their use was explored in the 1980s and 90s and Shennan (1997: 201–13) and Baxter (203: 131–36) have brief sections on them, but I do not think they have been widely used and other than a very brief notice in Section 12.5 are not accorded further discussion.

4.3 The iniquitous pie-chart

Pie-charts are circular graphs that are divided into segments or slices whose areas are proportional to the percentages for a single set of counts. The `R` help notes state that 'Pie-charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data'. There is nothing to disagree with here and we proceed mainly by illustration.

Wainwright (1984: 8) shows a pie-chart of the regional distribution of scheduled monuments in England, by period. It is stated that the chart shows 49% of prehistoric date, 43% medieval, 7% Roman and 3% post-medieval. It is noted that the 'percentages are crudely derived' and given this and just four categories all that is needed is to say that prehistoric and medieval dates predominate with similar percentages, and that the other two periods are much less common.

A graphic is not needed, and if you must have one a barplot is better. The pie-chart in the paper is emulated in the top-left of Figure 4.6. There is nothing wrong with it, except that it is unnecessary and occupies over a third of a page in the original publication.

It is possible to do worse, and often is. A pie-chart is a two-dimensional construct; properly done, the areas provide a true representation of the numbers

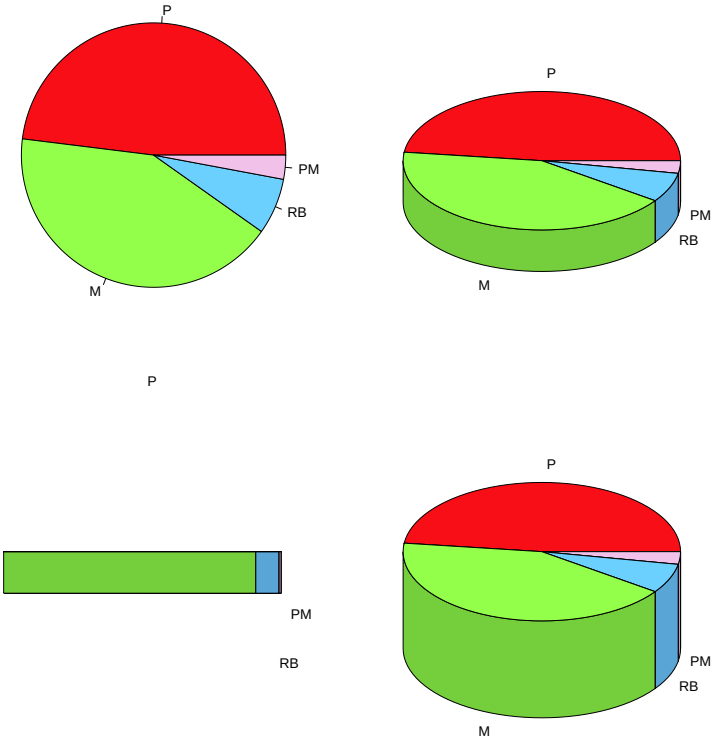

Figure 4.6: *Different 'pie-chart' presentations. The charts show the distribution of scheduled monuments by period; P = Prehistoric, M = Medieval, PM = Post-medieval, RB = Romano-British. (Source: Wainwright, 1984.)*

involved. The authors of too many publications have been seduced by the lure of 'business-type' presentations and seem to use default options in packages not designed for proper statistical use. These distort the 'truth' of data presentation and should not be allowed in academic publications. Similar comments apply to barplots, as illustrated in Figure 4.5. That the default options can be overridden is often neglected.

Some illustrations are provided in the rest of Figure 4.6. Plots like that in the upper-right are most commonly found. A meaningless third dimension, height, that is arbitrary, is added to the plot. To enable this to be seen it is necessary to tilt the plot, the degree of tilt also being arbitrary. Presumably this is done to make the chart look more 'exciting', but the price paid is to distort the information the chart is intended to display.

It might be claimed that this kind of misrepresentation does little harm, but take it to extremes! The bottom-left plot takes the tilt to its extreme and shows the chart from the side. It contains no useful information. The final plot uses a

height that begins to distract from such information as is there. In using three-dimensional effects, incidentally, it can become difficult to judge the size of the smaller categories. No-one, of course, would be stupid enough to use plots like the bottom two, but if you depart from the presentation given in the top-left figure you are on the slippery slope that leads you there.

Shennan (1997: 23) suggests that the 'pie-chart is a very helpful mode of data presentation when the aim is to illustrate relative proportions of unordered categories, especially when making comparisons'. This, particularly the comment about making comparisons, is highly questionable. His following comment that pie-charts can be confusing if there are numerous categories or categories with small or zero entries is to the point.

A good case can be made for abandoning the pie-chart as a method of archaeological data presentation.

4.4 R notes

Figure 4.1

Most of the labeling arguments in the code for this and the following example are omitted.

```
histbar <- function() {
wt <- loomweights$weight
wt <- wt[wt > 90 & wt < 400]
win.graph() # histogram
hist.wt <- hist(wt))

count <- hist.wt$counts #histogram counts
mid <- hist.wt$mids     #mid-points of bins

win.graph()
barplot(count))
win.graph()
barplot(count, names.arg = mid, space = 0.7)
win.graph()
barplot(count, names.arg = mid, space = 0)
}
histbar()
```

The loomweight data of Tables B.3 and B.4, from the file `loomweights`, are used. The weights are extracted, in `wt`, using the appropriate column heading from that file and outliers, determined by prior data exploration, are then omitted. This

is done by selecting weights greater than 90 g and less than 400 g, excluding six outliers outside this range[2].

As coded here the histogram needs to come first since an object `hist.wt` is created from which the bin counts and mid-points, used in the subsequent barplot constructions, are extracted. The first barplot is the default presentation; the subsequent barplots use the `names.arg` and `space` arguments to add the mid-points as axis labels and control the spacing between bars. The use of `space = 0` closes up the gaps and a histogram results. A more complicated example of this kind of usage follows.

Figure 4.2

```
hist.varbins <- function() {
z <- pillarmoulded

win.graph()     # barplot
barplot(z$Percentage, names.arg = z$Period, las = 2)

win.graph()     # histogram with unequal bin-widths
barplot(z$Percentage/z$Width, space = 0, width = z$Width, las = 2,
names.arg = z$Period, axes = FALSE)
}

hist.varbins()
```

The file `pillarmoulded` used is that based on Table 4.1 with the column headings as given there. Both plots use `names.arg = z$Period` to supply axis labels and this causes some problems in fitting all the labels on the plot, particularly with the small bar widths evident to the left of the histogram in the figure. Using `cex.names` to reduce the expansion factor is unsatisfactory because labels become too small to be easily legible. The 'solution' adopted here was to use the `las = 2` argument to produce labels perpendicular to the axis (other options can be found in the help for the `par` function).

Other than this, the first argument in the second plot 'adjusts' the percentages, dividing by width to compensate for the different duration of the periods; uses the `space = 0` argument (as illustrated in the previous example), in conjunction with the `width` argument that specifies the bin-widths to use, to produce the histogram

[2]Available logical operators include `<=` for 'less that or equal to' and `>=` for 'greater than or equal to'. Thus `wt <- wt[wt > 90 & wt < 400]` or `wt <- wt[wt >= 91 & wt <= 399]` could be used. Other operators include `==` for equality, and `!=` for lack of equality which can also be used with character variables.

desired; and uses the axes = FALSE argument to remove, in particular, the default y-axis which is meaningless given the different bin-widths.

Figure 4.3

```
Kastritsa.barplot <- function() {
z <- t(Kast) # Interchange rows and columns
win.graph(); barplot(z, beside = F)
win.graph(); barplot(z, beside = T)
win.graph(); barplot(z, beside = F, horiz = T)
win.graph(); barplot(z, beside = T, horiz = T)
}

Kastritsa.barplot()
```

The final four columns of Table 4.2 were imported into `R` as the object `Kast`. For the purpose of the analysis, which operates on the columns of the data table, the rows and columns need to be interchanged so that the columns correspond to strata. This is achieved by the `transpose` function `t()`. The default, `beside = F`, produces a vertically arrayed stacked barplot; the `beside = T` argument produces a clustered barplot; the argument `horiz = T` produces a horizontal array.

Figure 4.4

This is the code for the plot to the left. That for the plot to the right is identical except that the `beside = T` argument is used; the legend is displayed vertically at the top-left; and `ylim = c(0,100)` is used.

```
stacked <- function() {
barplot(t(Kast), names.arg = c("1", "3", "5", "7", "9"),
xlab = "stratum", legend.text = TRUE, args.legend = list(x = "top",
horiz = T, bty = "n", title = "Artifact class", cex = 1.3),
ylim = c(0,119), col = c("skyblue","yellowgreen","red","yellow"),
cex.lab = 1.3, cex.axis = 1.3, cex.names = 1.3)
}

stacked()
```

The `names.arg` supplies the names to be used for the x-axis. Note that quotation marks are used (e.g., `"1"`) so that the names are character rather than numeric variables. The legend is supplied as an argument to the `barplot` function, rather than externally, using `legend.text = TRUE`; the `list` function supplied to the

`args.legend` controls the placement and appearance of the legend. Most of the arguments are familar from previous uses of the `legend` function. Placement is more explicitly declared using `x = "top"` which puts it at the top center with a horizontal alignment using `horiz = T`. The `ylim = c(0,119)` argument expands the range of the y-axis slightly to accommodate the legend. The `R` help, `?barplot`, gives far more detail about the construction of barplots that could be used.

Figure 4.6

```
wainpie <- c(49,43,7,3)
wainpie.name <- c("P", "M", "RB", "PM")
Col <- c("red", "greenyellow", "skyblue", "pink")

Pie <- function() {}
library(plotrix)

win.graph()
pie(wainpie, wainpie.name, col = Col , radius = 1)
win.graph()
pie3D(wainpie, labels = wainpie.name, col = Col)
win.graph()
pie3D(wainpie, labels = wainpie.name, theta = pi/2, col = Col)
win.graph()
pie3D(wainpie, labels = wainpie.name, height = .7, col = Col)
}

Pie()
```

The `pie` function in `R` very properly does not encourage the use of three-dimensional pie-charts (or pie-charts at all, for that matter), for which the `pie3D` function from the `plotrix` package was used. The first three lines provide the four numbers from which the chart is constructed, shortened names indicating the period of construction of the monuments that are the focus of analysis, and colors to be used for plotting. The arguments `theta` and `height` control the angle of view and the 'depth' of the artificial third dimension.