Making Sense of Statistics for Family Practitioners

"Ready to Explore?"

Meticulous preparation and an intimate knowledge of available equipment and techniques, characterize most successful expeditions or adventures, whether they be the treacherous climb to the summit of Kilimanjaro, the arduous swim across the English Channel or nerve-wracking analysis of one's first data-set. Being confronted by a large number of measurements can be intimidating and attempts at analysis an utter waste of time unless a systematic approach is adopted. An important first exploratory step before launching into the production of florid graphics, is to convert data into a simple presentation form that allows insightful "eye-balling". We have already introduced the concept of discrete data with its distinct categories, e.g. male and female. These categories are useful for constructing tables and we will discuss the immense value of well-constructed tables for analysing and communicating data in the next article in this series.

*When exploring a continuous variable with a broad range of possible responses, such as time or age, it is essential to first group the measurements into a manageable number of categories (*class intervals*). In creating class intervals, the following guidelines must be kept in mind:*

- Class intervals must be mutually exclusive and include all data. For example, if the first interval is 0-4 years of age, then the next interval must begin with 5 years of age, and not 4. This is essential otherwise a child aged four will belong in two class intervals and be counted twice. This is particularly important when working with decimal measurements, e.g. a biochemical measure like serum glucose where measurement is to one decimal place
- It is preferable to start with a relatively large number of narrow class intervals for initial analysis, because if the data is collapsed into too few class intervals, important information may be concealed. Also, it is important to remember that after grouping data, all observations in the interval are treated as if they have the value of the center of the interval. In general, this should end up with 4 to 8 intervals.
- If rates are going to be calculated, then the intervals chosen for the numerator must be the same intervals as those used for the denominator.
- *It is often valuable to create a category for* "unknowns". *For example if the age of all study subjects is not available, then create a category "age unknown".*
- As a general rule all intervals should be of a constant size and it is preferable to avoid open-ended classes.
- Use natural or biologically meaningful intervals when possible e.g. use age groupings that are standard or are used most frequently (Table 1).[1]

If no natural or standard class intervals are apparent, several strategies are available for creating intervals and the most commonly used method is to divide the "range" *into equal* "class intervals". *This method is simple, readily adapted to graphs and to apply it, the following must be done:*

- Find the range of values in the data set i.e. the difference between the maximum value (or some slightly larger convenient value) and the minimum value (or zero)
- Decide how many class intervals are needed. The number will depend on what aspects of data are to be highlighted, but the rule of thumb is to have between 6 and 15 class intervals (a formula is available for guidance on the number of categories but should not be used rigidly). For tables, 4 to 8 class intervals are used and for graphs and maps, 3 to 6 class intervals will suffice.
- Determine the size of the class interval by dividing the range by the number of class intervals decided.
- Begin with the minimum value as the lower limit of the first interval and calculate class intervals of the size calculated until the maximum value in the data is reached.

As a simple example, let us consider the ages (in years) of the first 92 malaria patients presenting at Naas Clinic, Mpumalanga Province and recruited into an in vivo *study of first-line malaria drug effectiveness.*

21, 83, 34, 4, 25, 7, 8, 24, 33, 13, 54, 14, 25, 16, 72, 37, 18, 68, 18, 29, 29, 30,16, 20, 21, 22, 6, 23, 30, 23, 23, 5, 25, 25, 26, 46, 26, 26, 56, 27, 27, 28, 16, 28, 28, 19, 29, 29, 19, 30, 10, 6, 10, 30, 12, 30, 31, 31, 24, 31, 32, 12, 32, 33, 13, 4, 16, 34, 34, 36, 30, 20, 37, 17, 39, 43, 26, 47, 48, 23, 24, 15, 48, 50, 14, 6, 26, 59, 63, 18, 12, 3

If we wish to construct class intervals of equal length then the first step is to order, or rank, the data values from the smallest to largest value as follows: -

3, 4, 4, 5, 6, 6, 6, 7, 8, 10, 10, 12, 12, 12, 13, 13, 14, 14, 15, 16, 16, 16,16, 17, 18, 18, 18, 19, 19, 20, 20, 21, 21, 22, 23, 23, 23, 23, 24, 24, 24, 25, 25, 25, 25, 26, 26, 26, 26, 26, 27, 27, 28, 28, 28, 29, 29, 29, 29, 30, 30, 30, 30, 30, 30, 31, 31, 31, 32, 32, 33, 33, 34, 34, 34, 36, 37, 37, 39, 43, 46, 47, 48, 48, 50, 54, 56, 59, 63, 68, 72, 83

Then define the range, *by subtracting the lowest value from the highest value, i.e. 83 – 3 = 80. We decide on 8 class intervals by dividing the* range *(80) by 8, resulting in the length of the class interval*

being 10. It is then possible to divide all the observations into the new class intervals (Table II).

An under-utilised method that compliments the use of frequency distributions (class intervals) for initially exploring data is the "stem-and-leaf" display. This horticultural term describes a method where all leading values form the "stem", and each data value is represented by writing its trailing digit in the appropriate row next to the stem, forming the leaves. That is, each digit to the right of the stem is a leaf and the digit left of the stem is the stem label. This method is superior for initially looking at data as it retains all the original information. Unfortunately its value is limited to small sets of data. We will demonstrate the use of this method by summarising the platelet counts of a cohort of patients that were the focus of a confidential inquiry after they died with a diagnosis of malaria.[2] The platelet counts available for the 29 patients were:

10, 11, 12, 15, 16, 18, 20, 22, 23, 26, 31, 32, 35, 39, 39, 43, 45, 58, 59, 65, 67, 72, 72, 81, 81, 85, 103, 205, 218

Representing this data in a "stem-and-leaf" *display, the following display emerges:*

Stem	Leaf
1	0,1,2,5,6,8
2	0,2,3,6
3	1,2,5,9,9
4	3,5
5	8,9
6	5,7
7	2,2
8	1,1,5
9	
10	3
12	
13	
14	
15	
16	
17	
18	
19	
20	5
21	8

Table I: Malaria case fatality ratios by age group, Jan-Jun 1996, Mpumalanga Province.

Age Group (Years)	Case Fatality Ratios
0–9	0.74
10–19	0.13
20–29	0.88
30–39	1.08
40–49	1.47
50–59	2.01
60–69	1.80
≥70	9.68

Table II: Malaria patients recruited into a malaria drug effectiveness study, Naas Clinic, Mpumalanga Province, 2000.

Class Intervals Age Group in years	Number of Observations
3–13	16
14–23	22
24–33	34
34–43	8
44–53	5
54–63	4
64–73	2
74–83	1

At a glance this "stem-and-leaf" display demonstrates that almost all the patients had a platelet count below "90" and a large proportion of them had a very low platelet count below "40", something that is more difficult to derive from the list of platelet counts alone. In addition, the "stem-and-leaf" display allows any value that lies away from the other values (an outlier) to be detected easily, as can be seen with the platelet count of 218.

In conclusion, it is important to realize that when exploring a continuous variable with a broad range of possible responses, it is essential to group the measurements into a manageable number of categories i.e. class intervals. The "stem-and-leaf" display is an under-utilised method that compliments the use of frequency distributions (class intervals) for initially exploring data, provided the data set is small.

References:

1 *Dürrheim DN, Fieremans S. Profile of patients dying with* Plasmodium falciparum *malaria in Mpumalanga.* S Afr J Epidemiol Inf *1999; 14: 24-25.*

2 *Dürrheim DN, Fieremans S, Kruger P, Mabuza A, de Bruyn JC. Confidential inquiry into malaria deaths, Mpumalanga Province, South Africa.* Bull World Health Organ *1999; 77: 263-266.*

David N. Durrheim *MBChB, DTM&H, DCH, FACTM, MPH &TM*
Consultant: Communicable Disease Control, Department of Health, MPUMALANGA

Gboyega A. Ogunbanjo *MBBS, MFGP (SA), MFAM MED (MEDUNSA)*
Principal Family Physician & Senior Lecturer, Department of Family Medicine and Primary Health Care, MEDUNSA