

Essential statistics for the pharmaceutical sciences

Dr Phil Rowe BSc, MSc, PhD, FRSS

Publisher: John Wiley & Sons Ltd (<http://eu.wiley.com>)

Published: 2007

ISBN: 978-0-470-03468-2

Additional topics as supplements to the book:

Data presentation

Data presentation

A picture is worth a thousand words

The book focuses on statistical analyses that allow complex data and experimental outcomes to be summarised objectively and clearly in one or two numbers. However, used in isolation, statistical analyses can be horribly misleading and this document will emphasise the value of the pictorial representation of data. A clear graph or bar chart will often alert you to some important aspect of the data that would be missed if we relied solely on one or two dry, numerical statistics.



Data should be assessed both pictorially and statistically

- Pictures often reveal unsuspected aspects of the data
- Statistics provide an objective test of what the data really does (or does not) demonstrate.

1 Numerical tables

We have three different i.v. injectable chemotherapeutic agents. One is our current standard formulation and then we have two new candidate preparations. We want to look at the degree of nausea that they cause. Each is administered to 30 patients and they assess nausea on a 4 point ordinal scale (1 = none, 2 = slight, 3 = moderate, 4 = severe).

The results are shown in Table 1:

Table 1 Numbers of patients reporting varying degrees of nausea following i.v. injection of 3 different chemotherapeutic agents.

	Current standard	Candidate 1	Candidate 2
1 (None)	3	4	8
2 (Slight)	6	6	12
3 (Moderate)	19	17	9
4 (Severe)	2	3	1

Presenting the data as a numerical table has both good and bad aspects:

Good: The full details of the original data are available. No doubt we will have done our own analysis of the data, but others may want to analyse the same data in a different way or may wish to combine this data with that from other studies. Reporting the data as a numerical table, makes such re-analyses possible.

Bad: These tables are rather forbidding and lacking in immediacy. If your readership is highly numerate - e.g. colleagues at a scientific conference – they will not be fazed by this table. However, if you included the table in an article aimed at a lay audience, their eyes would glaze over and all higher intellectual functions would face imminent shutdown.

Even with a numerate audience there is still the problem of immediacy. About the only thing that emerges quickly is that most patients have suffered moderate levels of nausea, absence of nausea being quite rare. But, the more important issue is the comparison of one formulation with another. If you look carefully enough it is possible to see that Candidate 1 barely differs from the current standard and Candidate 2 may be a useful improvement, but such niceties certainly don't stand out immediately. The stacked bar chart discussed in the next section gets the message over much more dramatically.



Numerical tables

Good: Raw data available for further analysis.

Bad: Unfriendly and poor immediacy.

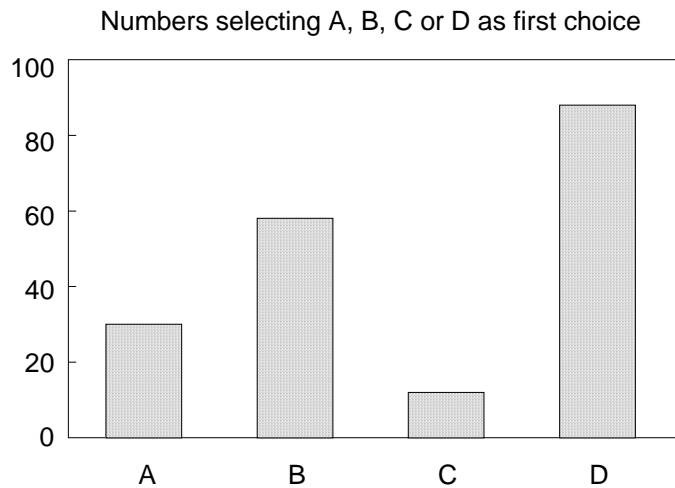
2 Bar charts and histograms

Any of the three types of data (Interval, ordinal or nominal; See [book](#) for details) can be reported as a bar chart. But the ease of doing so varies.

2.1 Simple bar charts

If we want to represent numbers of patients preferring product A, B, C or D we are presenting nominal scale data. Nominal data is always discontinuous and generally falls into a small number of natural and distinct categories. It is therefore ready made for presentation as a bar chart as in Figure 1:

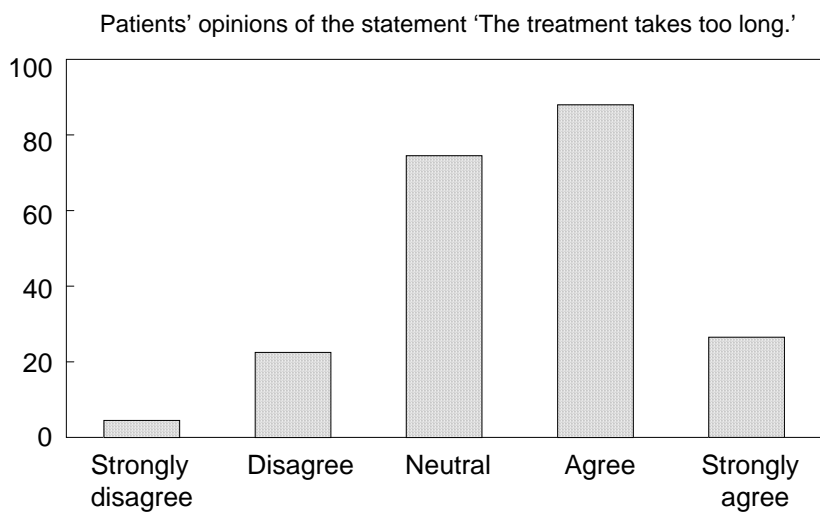
Fig 1 A bar chart based on nominal data - number of respondents preferring each product



Note that the horizontal scale represents nominal scale data. It does not represent a continuous scale of measurement. To emphasise the discrete nature of the categories, spaces are left between them.

Ordinal data is most commonly collected using a scale of measurement with a small number of possible values, so again it tends to be immediately appropriate for use in bar charts. The example below, (Figure 2) with a 5 point scale, forms a natural basis for a simple bar chart.

Fig 2 A bar chart based on ordinal data - opinion concerning a treatment



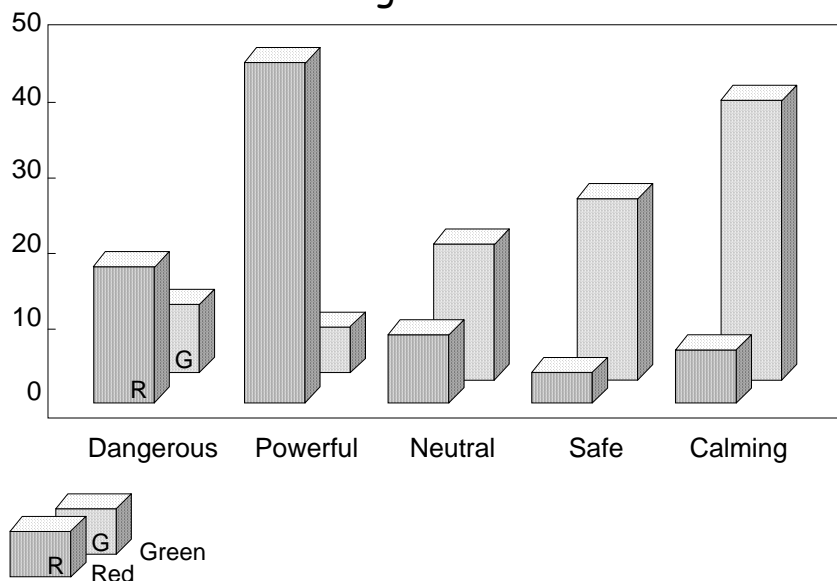
As with Figure 1, gaps have been left between the bars, to emphasise that the horizontal scale is not a continuous scale of measurement. There were no opinions recorded at any points in between the five points shown above.

A simple bar chart is adequate simply to describe a single outcome. However, if we want a visual representation that will allow us to compare two (or more) outcomes, we are going to need something fancier. There are a couple of ways to do it – Three dimensional and stacked bar charts.

2.2 Three dimensional bar chart:

A group of volunteers were each shown a single tablet and asked to choose one word that best expressed their opinion of it. They chose from a list of five ('Dangerous', 'Powerful', 'Neutral', 'Safe' or 'Calming'). All tablets were identical apart from their colour, which was either red or green. The results are presented as a three dimensional bar chart in Fig. 3 below. Patients are influenced by the colour, with red being seen as powerful or even dangerous while green is safe or calming. That difference stands out immediately in the bar chart.

Fig 3 Three dimensional bar chart - Reactions to red and green tablets

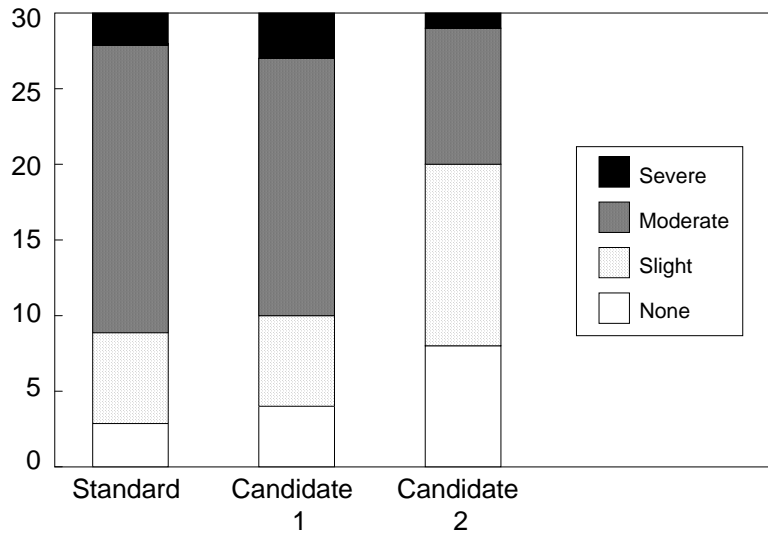


2.3 Stacked bar chart:

We might try to present the data from Table 1 concerning levels of nausea following chemotherapy as a three dimensional bar chart, but it doesn't actually work out too well, because some shorter bars are going to be hidden behind the taller ones. However, stacked bars can be used to portray the data quite effectively. It is immediately obvious from Figure 4 that candidate preparation number 1 has produced very little change relative to the existing

standard,. But candidate 2 is much more promising, giving us a majority with only Slight or No nausea.

Fig 4 A stacked bar chart - Nausea with three different formulations of chemotherapy agent



2.4 Histograms

Trying to present interval type data as a bar chart is less straight forward. This type of data is usually measured on a continuous scale and it does not generally fall into a small number of distinct categories. However we can artificially convert it to categories by breaking it up into bands. For example, we have some observations of patients' temperatures five days following surgery. We could classify each individual into one of the following bands based on their temperatures:-

36.8 – 37.0 °C
 37.1 – 37.3 °C
 ⋮
 etc
 ↓
 38.6 – 38.8 °C

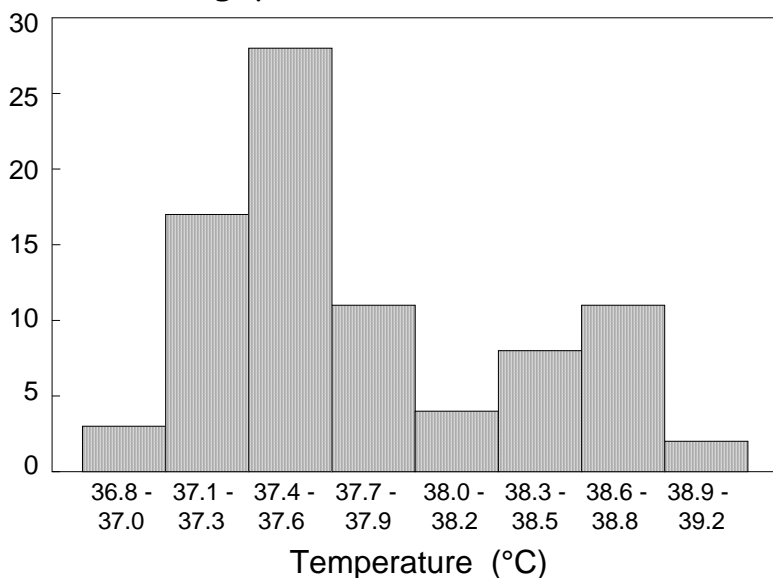
Note that these bands must fulfil three requirements:

- No gaps. If we had bands of 36.5 – 36.7 and then 36.9 – 37.1, we could not allocate a temperature of 36.8 to a category.

- No overlaps. If we had bands of 36.5 – 36.7 and then 36.7 – 36.9, an individual with a value of 36.7 could be allocated to two possible categories.
- All bands of equal width. If the first few bands covered a range of 0.3°C, but then we went to bands covering 0.6°C, the later categories would contain greater numbers of individuals. The increased heights of these bars would have nothing to do with these temperatures being commoner. It would just be an artefact of the way we had categorised the data.

The results are then presented as Figure 5:

Fig 5 A histogram - Patient's temperatures.
(Note - no gaps between the vertical bars)



The chart suggests that there is a distinct sub-population with elevated body temperatures – presumably they have become infected whereas the others have not.

Unlike all of the previous cases, the horizontal axis now represents an outcome measured on a continuous scale which in reality contains no sudden breaks. To emphasise the continuous nature of the scale, we do not leave gaps between the bars.

Where the scale of measurement is essentially continuous, but has been artificially broken into bands, the resultant chart is given the special name of a 'Histogram'.



Histograms

A histogram is a bar chart using data that was originally on a continuously varying scale, but which has been subdivided into ranges to render it in a classified format. No gaps are left between the bars.

2.5 A general assessment of bar charts and histograms:

Bar charts are probably somewhat less intimidating than numerical tables, for a non-numerate audience. However, they are still far from perfect. They are much better than numbers in terms of their immediacy. Not just patterns in single sets of data but also contrasts between sets of data are far more easily appreciated. The one loss is that we no longer have access to the exact data. It is possible to add numbers to the bars, but in many cases this is awkward. For example in Fig 3 several of the bars are partially hidden and we would have to write the numbers somewhere else on the chart and have arrows connecting them to the appropriate bars. This would greatly clutter the diagram and much of the simplicity and clarity would be lost.



Bar charts and histograms

Good: Excellent immediacy for all main messages and reasonably friendly.

Bad: Difficult to include exact values without loss of clarity.

3 Pie charts.

3.1 Simple pie charts

We would almost never use pie charts to present data that was interval or ordinal in nature. Such data falls on a scale with a low and a high end, which is naturally expressed in a bar chart. Pie charts are circular and simply don't match the needs of measurement data. They are useful for nominal type data where there is no logical sequence to the categories.

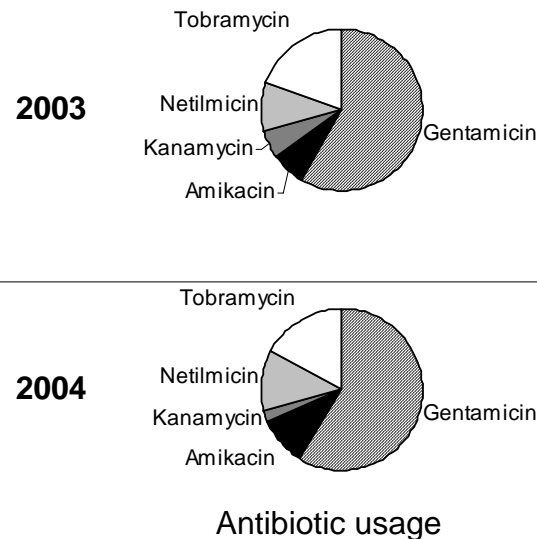
Figure 6 shows numbers of hospital patients treated with a variety of different aminoglycoside antibiotics in the years 2005 and 2006. Simple messages, such as which antibiotics are being used most frequently are conveyed with excellent immediacy. The medium is also pretty unthreatening to even a non-numerate audience – people quite like the mental image of a pie being sliced into larger or smaller portions.

Unfortunately, pie charts don't convey changes in patterns as effectively as bar charts. Changes are very easily seen in Figures 3 and 4, but in Figure 6 we have to check backwards and forwards between the two pie charts.

Eventually, you probably did notice that the use of Kanamycin declined markedly in 2004, but I bet it didn't hit your eye immediately.

As with barcharts, the original numerical data is lost, unless we are prepared to add a lot of clutter to what are currently nice clear figures.

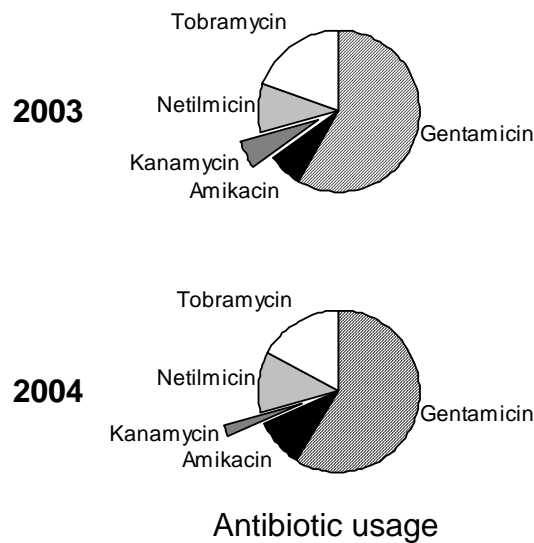
Fig 6 Pie chart - Survey of number of patients treated with various aminoglycosides 2003/4



3.2 Exploded pie charts

If one of the main points we want to convey is the reduction in the use of Kanamycin, then Figure 6 is a very poor approach. The Kamamycin slice is the last one we are likely to focus upon, with it being the smallest. Figure 7 is somewhat better. By exploding the relevant slice we can ensure that it gets noticed.

Figure 7 Exploded pie chart



Pie charts

Good: Excellent immediacy for conveying which categories occur most commonly. Friendly.

Bad: Only really appropriate for nominal type data. Less immediate identification of changes in patterns. (Exploded slices may help.)
Difficult to include exact values without loss of clarity.

4 Pictorial symbols

We have progressed from the least friendly mode of data presentation (Numerical tables to the much friendlier bar charts and pie charts. There is one more step we can take in our journey to nirvana – pictorial symbols. Figure 8 shows the utilisation of Wundadrug in large district hospital.

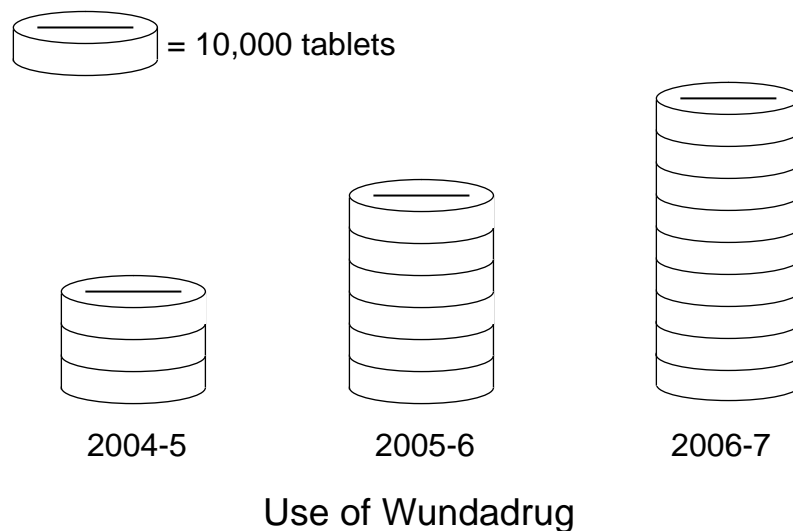
From a strictly objective stand point there is absolutely nothing wrong with this mode of representation. The meaning of the symbols is clearly defined (One tablet symbol equals 10,000 tablets dispensed) and the escalating use of the drug is immediately obvious and we get an accurate sense of the scale of increase.

So what's wrong with it? Why would you be laughed at if you included it in a presentation to a 'learned' society? The answer is almost certainly its utter clarity. Any member of the ordinary public could understand it and therein lies

the problem. The most important task for all academics is to convince the public we are much cleverer than them and in that respect, Figure 8 is a disaster. Joe Bloggs on the No 13 bus could understand it just as easily as Professor Halfmoons from the Institute of Advanced Obscurantism.

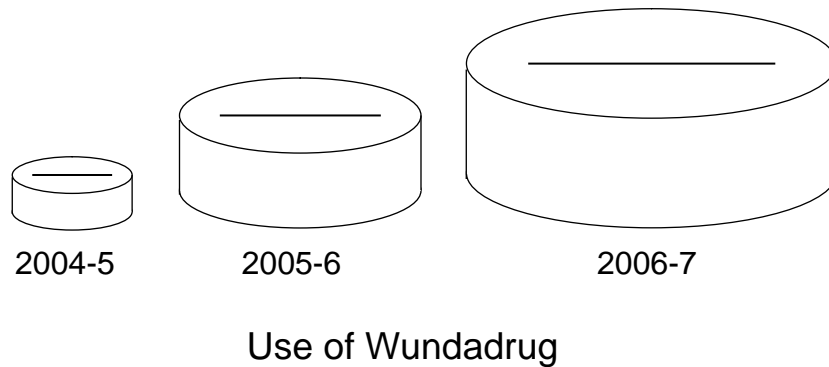
While you may never be able to use pictorial symbols in the academic world, they can be a valuable way to make a point to the general public. Using Figure 8 you could achieve what is normally impossible – convey quantitative data to an audience that would instinctively run a mile from anything with numbers in it.

Fig 8 Pictorial symbols



While Figure 8 is perfectly defensible, you may sometimes see an alternative use of pictorial symbols, as in Figure 9. Here, instead of replicating the same symbol an appropriate number of times, a single symbol has been rescaled to show change. Let's assume that Figure 9 is meant to convey the same information as Figure 8. If the reader focuses purely on the vertical heights of the symbols, the implied increase in use (three fold) will be correct. However, these are three dimensional objects and the width and depth are increased as well as the heights. The reader will tend to get an exaggerated impression in the extent of the increase. The tablet symbol for 2006-7 would be three times as high, wide and deep and its volume or weight would be 27 times greater than that for 2004-5.

Fig 9 Pictorial symbols - The wrong way!



5 Scatter plots

5.1 Dependent versus independent

All the data presentation methods we've looked at so far are appropriate for cases where there is just one measured value (parameter) being reported. Not uncommonly two parameters will have been determined and we want to look at the relationship between them. Here we normally use a scatter plot.

In statistics we frequently meet the distinction between a 'Dependent' and an 'Independent' variable. If we find that two parameters (A & B) are related then the question is how we would interpret that relationship. Is the value of A controlled by that of B or *vice versa*. For example in a pharmacokinetic trial, patients' body weights and their clearances of a drug would probably be related to one another. (Clearance describes the efficiency with which a drug is eliminated from the body.) We could reasonably assume that it was the clearance that was controlled by the body weight and not *vice versa*. In this case clearance is the dependent variable and body weight is the independent. We then always plot the dependent variable up the vertical axis and the independent along the horizontal. It is also customary to describe this as 'Plotting clearance against body weight.' (Note the order – it's dependent against independent, not the other way round.) Figure 10 shows data of this type as a scatter plot.

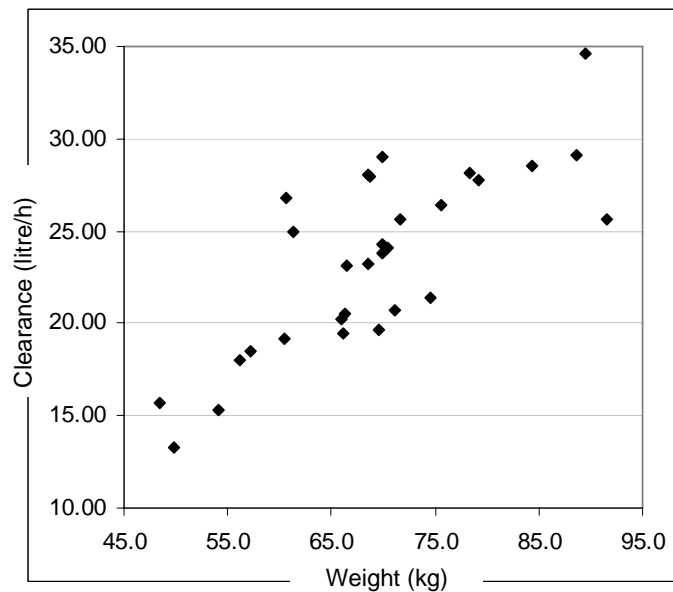


Dependent and independent variables

The dependent variable should be plotted up the vertical (Y) axis and the independent along the horizontal (X) axis.

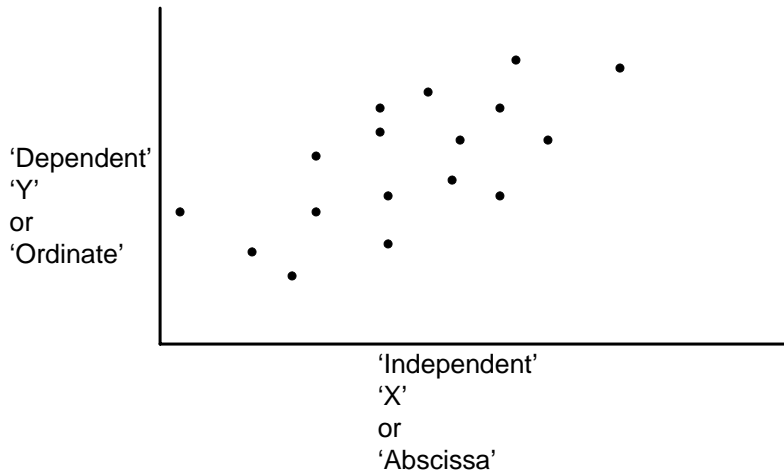
We say that 'The dependent variable is plotted versus the independent'.

Fig 10 Scatter plot of drug clearance against body weight of patients.



The horizontal and vertical axes may also be referred to as the 'X' & 'Y' axes. Other terms that are used (Albeit less frequently, since nobody can ever remember which is which) are the 'Abscissa' and the 'Ordinate'. See Figure 11.

Fig 11 Terms used to describe the horizontal and vertical axes of a scatter plot

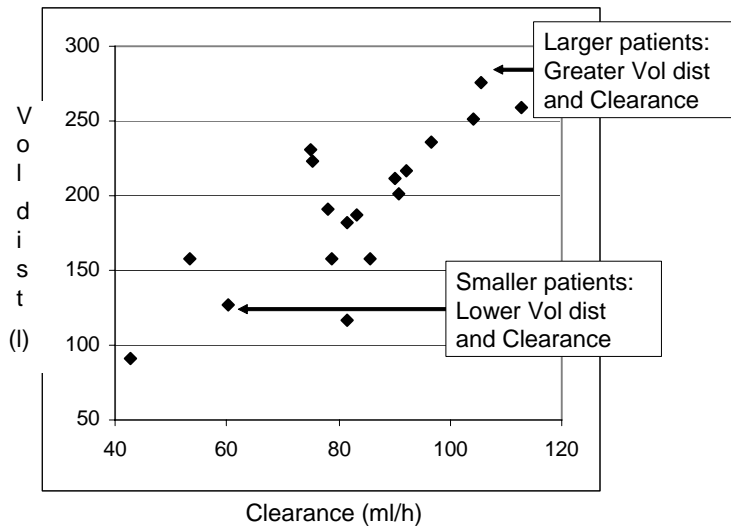


5.2 Scatter plots for data where there is no identifiable dependency within the data.

In some cases two parameters may show a clear relationship to one another, but neither can be identified as being dependent upon the other. This commonly arises when both parameters are dependent upon some third factor that causes them to vary together.

An example of the latter would be the volume of distribution and clearance of a drug. (Volume of distribution describes the extent to which a drug tends to distribute out of the blood into the tissues.) These two parameters are linked because both are dependent upon body weight. But there is no sense in which volume of distribution is dependent upon clearance or *vice versa*. In such a case we could equally well plot the data as volume versus clearance (Figure 12) or as clearance against volume.

Fig 12 Scatter plot for data with no identifiable dependency

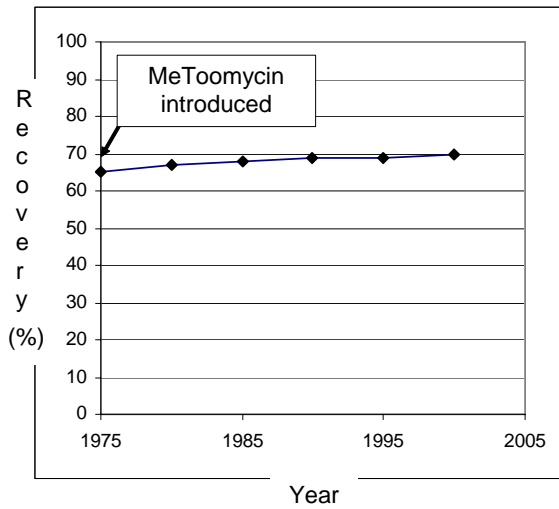


5.3 Darell Huff and the 'Gee Whiz graphs'

Back in the 1950's Darrel Huff drew attention to one of those tricks long beloved by presenters of misleading data. The basic idea is that you convert a disappointingly shallow graph into one that shoots up in a pleasingly dramatic way. The trick is to stretch the vertical scale and shrink the horizontal. Stretching the vertical axis, could of course lead to a graph that was excessively tall, but the real secret is to use only a small part of the available range of figures.

Consider some figures for improvement in cure rates for Fick-Tishus disease following the introduction of MeToomycin. Figures 13 and 14 describe the results. They look very different at first glance, but actually convey exactly the same results. Both could be subjected to criticism. Fig 13 is rigorously honest, but 90% of it is boring blank space. Linked to this, there is also the problem that if we wanted to read off what the recovery rates were in 1975 and 2000, it would be difficult to do so with any accuracy.

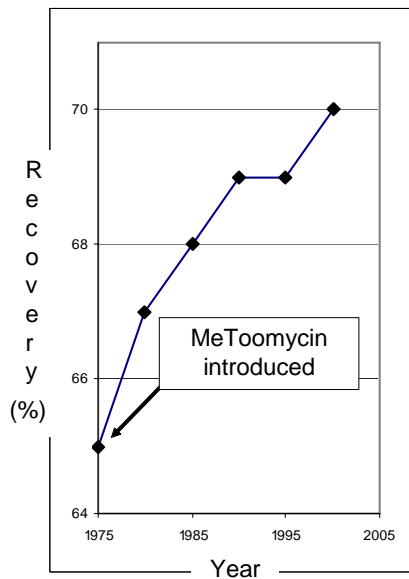
Fig 13 Graph using full vertical scale - Recovery rates from Fick-Tishus disease 1975 - 2000



MeToomycin – Modest improvements over 25 years in recovery from Fick-Tishus disease.

Fig 14 is far worse, being deliberately dishonest. A very small range of values has been stretched out to form the vertical axis, exaggerating the apparent increase in recoveries and this has been further emphasised by compressing the horizontal axis. The real crime is then the 'Gee whiz' headline to help the more gullible reader rush out and stock up on MeToomycin.

Fig 14 Gee whiz graph - same data as Fig 13



Fick-Tishus disease beaten by MeToomycin!!!

Apart from the abuse of the vertical axis in Fig 14, there is also the question that the graph only starts from 1975. We are given no idea what was going on before then. For all we know, general improvements in patient care may have been allowing a slow, steady improvement in recovery rates for the last 50 years and the introduction of the alleged wonder drug might have had no impact whatsoever.



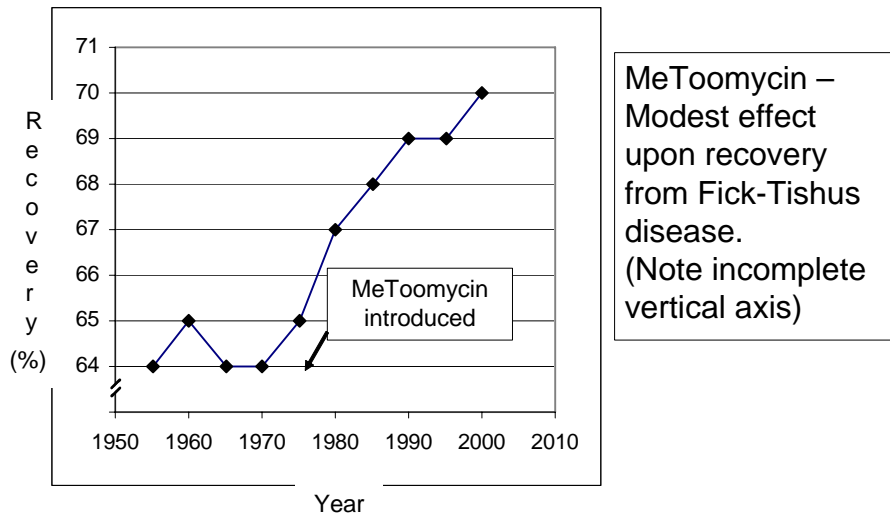
The pathetic becomes dramatic

A once brilliant scheme, now faded.

Even the most modest increase (or decrease) can be made to look impressive by quietly suppressing the zero on the vertical axis and expanding a small part of the scale. The problem is that Darrell Huff, in the best statistics book ever written ('How to lie with statistics') blew the cover on this one fifty years ago. There is no way you'll get away with it in any reputable journal. However, if you're writing a polemical article for a popular magazine or newspaper, you can still fool most of the people most of the time.

An acceptable presentation of the data would be something similar to Figure 14, but with a more modest headline and a clear indication that the vertical axis is incomplete. If data for the period prior to the introduction of MeToomycin is available, it would be useful to include it. (See Figure 15) From this figure we could read off exactly what did happen to recovery rates during the last 25 years and we can see that there had been no consistent progress with this disease prior to the introduction of MeToomycin. We should also gather that even after this drug was introduced, progress has been less than miraculous.

Fig 13 Fair and informative graph of the MeToomycin data.



6 Chapter summary

Data should always be explored graphically as well as statistically. A picture is worth a thousand words.

Numerical tables allow readers to access the primary data for re-analysis, but are unfriendly to less numerate readers and fail to convey the main features of the data with any great immediacy.

Bar charts can be used for any type of data. They are reasonably friendly and convey the main aspects of the data with excellent immediacy. They usually result in the loss of access to the primary data. If data from a continuously varying scale is rendered into classes based upon ranges, a bar chart of such data is then called a histogram.

Pie charts are mainly useful with nominal data. They are very friendly and convey which classes occur most commonly with great immediacy. Access to the primary data is generally lost.

Pictorial symbols offer a unique opportunity to smuggle quantitative information into public information, without scaring your readers. Compare the response you would get with Figure 8 to what you might expect from the same data presented as a numerical table.

Scatter plots are used to illustrate the relationship between two measured parameters. Where one parameter can be identified as being dependent upon the other, the dependent should be plotted up the vertical (Y) axis and

the independent along the horizontal (X). Beware of salespersons who make minor increases look disproportionately large by plotting only part of the Y axis.