Statistics in Question

SHEILA M GORE

ASSESSING METHODS—DESCRIPTIVE STATISTICS AND GRAPHS

In his 1981 presidential address to the Royal Statistical Society, Professor D R Cox said: "The setting out of conclusions in a way that is vivid, simple, accurate, and integrated with subject-matter considerations is a very important part of statistical analysis."

This series is about how to set out conclusions in medical papers and answers questions about when particular statistical methods are appropriate—and the hidden snags in using them. Types of problems that ought to be referred to a statistician are also discussed.

This article reviews descriptive statistics—mean, median, and mode, variance and interquartile range—explaining how a little detective work using these summary measures reveals a great deal about the distribution of observations even if a fully informative table or graph is not presented. It is usually helpful to use graphs to present data, and some reminders about how to do this are given—in particular, scattergrams are recommended.

(1) From the table find the most likely (modal) survival time and estimate median survival for the 347 patients with breast cancer who were referred to the department of radiotherapy, Edinburgh, in 1956.

—the most likely survival time is less than one year

—50% of patients survived for at least four years

—the difference (mean-median) is a crude measure of skewness

—measures of dispersion (variance, interquartile range)

are needed to qualify central measures such as mean and median

COMMENT

Authors are advised to summarise important aspects of their data—location, dispersion, skewness—by reporting descriptive statistics such as mean, median, mode, variance, range, percentiles, but too often the summary is presented at the expense of informative tables or graphs. The reader is left to infer from the summary the shape of the underlying distribution. Some guidelines are given for doing this.

Measures of location (or centre) are mean, median, and mode. Mean and median coincide for symmetric distributions. The sum of the observations divided by the number of observations estimates the true mean, and the value above which 50% of the observations lie estimates the median of the distribution. When the estimated mean and median are dissimilar and sample size—which should always be reported—is moderate the reader can deduce that the underlying distribution is asymmetric or skewed. From the table notice that 174 (50%) of the 347 patients survived for at least four years after diagnosis of breast cancer. Mean survival exceeded seven years. The difference (mean-

Survival of 347 patients with breast cancer

Survival time (years)*	Frequency	Cumulative frequency
<1	62	62
1-2	45	107
2-3	38	145
3-4	28	173
4-5	25	198
5-6	10	208
6-7	14	222
7-8	11	233
8-9	9	242
9-10	8	250
10-11	8	258
11-12	8	266
12-13	9	275
13-14	5	280
14-15	2	282
15-16	3	285
16-17	4	289
17-18	7	296
18-19	1	297
19-20	3	300
At least 20	47	347

*The interval 1-2 years includes survival times of one year and up to, but not including, 2 years. All patients were followed up to the 20th anniversary of first treatment; 300 patients died before 31 December 1976.

median) is a crude measure of skewness. For breast cancer it is positive—three years—and allows the reader to infer that the distribution of survival time is positively skewed—mean survival is exaggerated by a small but important proportion of long-term survivors.

Survival for 347 patients with breast cancer: summary

The most likely survival time, however, is less than one year. More generally there may be several maxima, in which case the distribution is multimodal. A typical example of there being two modes (bimodality) is when anthropometric observations for men and women are combined. Shoe size—a crude measure of foot length—is shown below for a class of 49 undergraduates. There are two modes: shoe sizes 6 and 9 correspond to the mode for women and for men respectively.

Scattergram of shoe size for a class of 49 undergraduates

Measures of dispersion are variance, range, and interquartile range. They are needed to supplement measures of location.¹ An appropriate measure of dispersion should be reported alongside mean, median, or mode, just as sample size is needed for interpreting proportions or percentages. Variance is the average squared deviation from the mean and as such it—or standard deviation, which is the square root of the variance and is measured thus in the same units as the observations—is the best measure of dispersion when the mean is the best measure of location. When the underlying distribution is skewed it is important to qualify the median by reporting also the quartiles—above and below which 25% of the observations lie—or percentiles, such as the 10th and 90th. From the table we estimate that 25% of the 347 patients had died within 18 months after diagnosis of breast cancer and that for 75% of patients survival time was less than 11 years. The quartiles are shown in the figure.

How much of the information in the table is conveyed by the figure alone? The answer is a great deal if you follow the clues: the most likely survival time was less than one year, 25% of patients had died within 18 months of breast cancer having been diagnosed, and half within four years. Because mean survival is three years longer than the median survival time, the distribution of survival is positively skewed and so there are long-term survivors. Another pointer to positive skewness is the fact that the time interval between the upper quartile and the median—seven years—is considerably longer than the corresponding interval between the median and the lower quartile. Long-term survival is confirmed because 25% of patients were still alive 11 years after being treated for breast cancer.

(2) Mean survival was approximately five years for the 300 patients (see table above) who died before 31 December 1976. Why is it wrong to conclude that patients in Edinburgh with breast cancer have a mean survival time of five years?

—because the calculation ignored long-term survivors

—five years is therefore a serious underestimate of mean survival time from diagnosis

—in any case, median survival is a more reliable measure

COMMENT

Five years is an underestimate of mean survival time from diagnosis because the 47 patients excluded from the calculation were the long-term survivors—that is, they survived for at least 20 years. Since follow-up was discontinued at the 20th anniversary information on these patients is right-censored—that is, we know only that their survival times exceed 20 years. A better estimator is derived by crediting these patients with 20 years' survival and so obtaining

$$(300 \times 5 \text{ years} + 47 \times 20 \text{ years})/347 = 7 \text{ years}$$

as a more realistic measure, but an underestimate still because the actual survival times from diagnosis for these 47 patients must be longer than 20 years.

In these circumstances median survival is a more reliable measure of location than the mean because the survival distribution is skewed and there are incomplete lifetimes. It is also easier to find. Complete reporting of survival data entails plotting life-tables and is the subject of a later article.

Another reason for being cautious is that if there is a time-trend in survival then patients referred in 1956 need not have the same survival pattern as patients referred more than two decades later. Patient care, or severity of disease, or treatment, or age at referral could have changed.

(3) Do figures A and B give the same information about tumour size?

—no, figure B is a misleading representation of the information given in figure A

—area represents frequency in a histogram

—neither figure mentions that tumour size was not recorded for 50 patients

COMMENT

Tumour size was measured clinically and rounded up to the nearest centimetre. Three patients had no measurable mass, and tumour size was not recorded for 50 out of 347 patients. Neither figure A nor figure B mentions these 50 missing observations. Figure A plots tumour size rounded up to the nearest centimetre, so that the 29 patients for whom recorded tumour size was 2 cm had tumours that measured more than 1 cm and up to 2 cm. Recorded tumour size is therefore on average $\frac{1}{2}$ cm more than actual size. Figure B purports to show the same data as figure A, but is misleading. The area of the bar over tumour size 10-15 cm represents a frequency of 40 patients—five times as many patients as really had a tumour that large. The bar over 15-20 cm is also in error by a factor of 5. Area, not height, represents frequency in a histogram. In figure B, moreover, it is not possible to separate the three patients who had no measurable mass from the 12 patients whose tumour measured up to 1 cm.

(4) What is the minimum information that should be given with a graph?

—title, labelled axes, scale, key to symbols

—scattergrams are recommended for exploring and reporting data

—distinct points should represent distinct measurements; in particular, bilateral observations (ocular tension in right and left eye) should not be represented as unrelated

COMMENT

Graphs should have a title, axes should be labelled, and the scales given. Breaks in scale should be clearly marked and attention drawn to outlying or rogue observations. Healy² has written a good article on informal graphical methods for detecting rogue observations. Different symbols that appear in the graph should be defined in a key, or as a footnote to the graph. Lowe *et al*³ studied the relation between the extent of coronary

artery disease and blood viscosity in 75 men aged 30 to 55 years. Fifty men were studied before they underwent selective coronary arteriography for assessment of chest pain; the other 25 had been admitted for elective minor surgery and served as controls. In a single vivid scattergram Lowe *et al* showed comprehensively the association between packed cell volume and blood viscosity for the asymptomatic controls, for the patients with no stenosis or stenosis of one vessel only, and for patients with stenosis of two or three major coronary vessels. Scattergrams are an

Blood viscosity and packed cell volume.

Conversion: SI to traditional units—Blood viscosity: 1 mPa s = 1 cP.

excellent device for exploring and reporting data, particularly for showing the association of one variable with another. Pocock *et al*⁴ showed the negative association of standardised mortality ratio and water hardness in 234 towns in this way and also used maps to good effect to illustrate the findings of the British Regional Heart Study.

Editors often advise authors that graphs should be interpretable without referring to the text. This is a good maxim. Another important point is that repeated observations on the same patient should be identified on the graph as being related, and should certainly not be represented as though they were measurements on several different patients. If several similar plots are reported together comparisons can be made most easily if the scale and the symbols are consistent from one graph to the next. The choice of scale is important. Transformation of data is discussed in the next article.

I thank G D O Lowe *et al* for permission to reproduce the scattergram under question 4.

References

- Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;i:85-7.
- Healy MJR. The disciplining of medical data. *Br Med Bull* 1968;24:210-4.
- Lowe GDO, Drummond MM, Lorimer AR, *et al*. Relation between extent of coronary artery disease and blood viscosity. *Br Med J* 1980;280:673-4.
- Pocock SJ, Shaper AG, Cook DG, *et al*. British Regional Heart Study: geographic variations in cardiovascular mortality, and the role of water quality. *Br Med J* 1980;280:1243-9.

Sheila M Gore, MA, is a statistician at the MRC Biostatistics Unit, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH.

No reprints will be available from the author.