7th March 2017

# Describing the participants in a study

**Pickering RM**

Associate Professor in Medical Statistics
Medical Statistics Group
University of Southampton
Tel 023 80796565
Fax 023 80794460
E-mail **rmp@soton.ac.uk**

*Address for correspondence*
Primary Care and Population Sciences (MP 805)
South Academic Block
Southampton General Hospital
Tremona Road
SOUTHAMPTON, SO16 6YD

**Word count Introduction to Acknowledgements** = 2828 words

**Abstract** (word count = 44)

This paper reviews the use of descriptive statistics to describe the participants included in a study. It discusses the practicalities of incorporating statistics in papers for publication in *Age and Aging*, concisely and in ways that are easy for readers to understand and interpret.

**Keywords**

descriptive statistics; study participants; attrition; generalizability

**Keypoints**

Descriptive statistics are used to describe the participants in a study so that readers can assess the generalizability of study findings to their own clinical practice.

They need to be appropriate to the variable or participant characteristic they aim to describe, and presented in a fashion that is easy for readers to understand.

When many patient characteristics are being described, the detail of the statistics used and number of participants contributing to analysis are best incorporated in tabular presentation.

## Introduction

Most papers reporting analysis of clinical data will at some point use statistics to describe the socio-demographic characteristics and medical history of the study participants. An important reason for doing this is to give the reader some idea of the extent to which study findings can be generalized to their own local situation. The production of descriptive statistics is a straightforward matter, most statistical packages producing all the statistics one could possibly desire, and a choice has to be made over which ones to present. These then have to be included in a paper in a manner that is easy for readers to assimilate. There may be constraints on the amount of space available, and it is in any case a good idea to make statistical display as concise as possible. This article reviews the statistics that might be used, and gives tips on how best to incorporate them in a paper for publication in *Age and Aging*.

## Describing the distribution of values

The values observed in a group of subjects when measurements of a quantitative characteristic are made, are called the distribution of values. Graphical displays can be used to show the detail of the distribution in a variety of ways, but they take up a considerable amount of space. A precis of two key features of the distribution, its centre and its spread, is usually presented using descriptive statistics. The centre of a distribution can be described by its mean or median, and the spread by its standard deviation (SD), range, or inter-quartile range (IQR). Definitions and properties of these statistics are given in statistical textbooks [1].

Figure 1a) shows an idealized symmetric distribution for a quantitative variable. The mean might be used here to describe where the centre of the distribution lies and the

SD to give an idea of how spread out values are around the centre. Standard deviations are particularly appropriate where a symmetric distribution approximately follows the bell-shaped pattern shown in Figure 1a) which is called the Normal distribution. For such a distribution the large majority, 95%, of values observed in a sample will fall between the values two SDs above and below the mean, called the Normal range. Presentation of the mean and SD invites the reader to calculate the Normal range and think of it as covering most of the distribution of values. Another reason for presenting the SD is that it is required in calculations of sample size for approximately Normally distributed outcomes, and can be used by readers in planning future studies. A graphical display of approximately Normally distributed real data (age at admission amongst 373 study participants) is shown in Figure 1c): with relatively small sample size a smooth distribution such as that shown in Figure 1a) cannot be achieved. The mean (82.9) and SD (6.8) of the age distribution leads to the Normal range 69.2 to 96.5 years, which can be seen in Figure 1c) to cover most of the ages in the sample: 14 subjects fall below 69.3 and 7 fall above 96.5, so that the range actually covers 352 (94.4%) of the 373 participants, close to the anticipated 95%. For familiar measurements, such as age, there is additional value in presenting the range, the minimum and maximum values attained. Knowing that the study included people aged between 65 and 101 years is immediately meaningful, whereas the value of the SD is more difficult to interpret.

When a distribution is skewed (Figure 1b) just one or two extreme values, ‘outliers’, in one of the tails of the distribution (to the right in Figure 1b)) pull the mean away from the obvious central value. An alternative statistic describing central location is the median, defined as the point with 50% of the sample falling above it and 50%

below. Figure 1d) shows the distribution of real data (hours in A&E amongst 348 study participants) following a skewed distribution. A few excessively long A&E stays pull the mean to the higher value of 4.9 hours compared to the median of 4.4 hours: the effect would be greater with a higher proportion of subjects having long stays. The median is often recommended as the preferred statistic to describe the centre of a skewed distribution, but the mean can be helpful. If the attribute being described takes only a limited number of values, the medians of two groups can take the same value in spite of substantial differences in the tails. In these circumstances the mean can be sensitive to an overall shift in distribution while the median is not. When a comparison of cost based on length of stay is to be made, presenting means of the skewed distributions facilitates calculation of cost savings *per* subject by applying unit cost to the difference in means. Figure 1b) suggests that the value with highest frequency might be a useful descriptor of the centre of a distribution. In practice this can prove awkward: depending on the precision of measurement there may be no value occurring more than once.

It is clear from Figure 1b), that no single number can adequately describe the spread of a skewed distribution because spread is greater in one direction than the other. The range (from 1.7 to 40.3 hours in A&E in our skewed example) could be used. Another possibility is the IQR (from 3.5 to 5.4 hours in A&E) covering the central 50% of the distribution. The SD may be presented even though a distribution is skewed, and could be useful to readers for approximate power calculations, but the Normal range derived from the mean and SD will be misleading. With mean(SD) = 4.9(3.2), the lower limit of the normal range of hours in A&E is the impossible negative value of

-1.5 hours, while the upper limit of 11.3 hours lies well below the extreme values exhibited in Figure 1d).

## Descriptive statistics in text

Descriptive statistics may be presented in text, for example:

> "*Participants' ages ranged from 50 to 87 years (M = 66.1, SD = 7.8) with 56% identified as female, 64% married or partnered, 23% reported being retired or not working, 55% had post-secondary and higher education, and <20% reported living alone. Over 60% of the participants identified as NZ European. The mean of net personal annual income was $34,615. The participants reported the diagnosis of an average of 2.63 (±2.07) chronic health conditions, with 50% reported having three or more chronic health conditions.*" [2]

There are perhaps too many attributes (age, gender, marital status, employment status, educational level, living arrangements, nationality, personal income, and number of chronic conditions) being described in the excerpt above: it would be easier to assimilate this information from a table.

## Descriptive statistics in tables

Where there are too many characteristics to be described in text, or several sub-groups of participants are being compared, tabular presentation becomes more convenient. An example summarizing the distribution of 11 categorical variables and two quantitative variables in the two phases of a before-after evaluation of the introduction of a care pathway for hip fracture [3] is shown in Table 1. The categorical variables (so called because they indicate which of several categories a participant falls in) are best described by the number (and percentage) in each category. Since categorical variables are in the majority in Table 1, the title indicates that the figures presented are "number (%) unless stated otherwise". It is best to give the number as well as the

percentage, unless a study is very large, to emphasise that percentages are estimated with imprecision. For example, the 90 males represent 23% of the 395 participants in the 1998/99 phase, but the percentage alone gives no indication of the appreciable imprecision in the estimate which has 95% confidence interval from 19% to 27%. Unless a very large sample is available, the information conveyed by the decimal places in a percentage is spurious accuracy. For example, the 66 participants whose operation was delayed for organisational reasons of the 172 with a reason stated in 1998/99, is displayed rounded to no decimal places as 38% in Table 1. Displayed with two decimal places it becomes 38.37%: had there been 67 participants delayed for an organisational reason the percentage would have been 38.95%. No other values between 38.00% and 39.00% are possible for a percentage calculated from a sample of 172. Even were a large enough sample available to distinguish between percentages of 38.37% and 38.95%, it would make no meaningful difference to interpretation here, but presentation as 38.37% with two decimal places clutters the display and makes the percentage difficult to assimilate. Rounding to no decimal places has resulted in the percentages for the 3 reasons summing to less than 100% (35%+38%+26%=99%). This artefact can occur in the final digit however many decimal places are presented. It is possible to describe the distribution of a binary characteristic with number (%) for both categories, as has been done for gender in Table 1, or for just one of them, as has been done for history of dementia, to save space. Where there are more than two categories it is better to present number (%) for all of them to clarify the options.

The distributions of the two quantitative variables in Table 1 are described by mean (SD) and range. The statistics being presented should be stated in the context of

the table, here in the left hand column, and could differ across variables. If the same statistics are presented for all the variables in a table they can be indicated in the column headings or title. From the mean (SD) and range in each phase, we can see that the age distribution is reasonably symmetrical because the mean falls close to the centre of the range, and the mean ± 2SD approach the limits of the range. The distribution of hours in A&E is skewed to the right but has been summarized with the same statistics. We can see that the distribution is skewed because the mean is much closer to the minimum than the maximum, and, if the Normal range is calculated, the upper limit does not approach the high values in either phase. For these reasons the Normal range should not be interpreted as covering 95% of values. These conclusions from descriptive statistics alone can be verified in Figures 1c) and 1d).

A choice arises when describing the distribution of an ordinal variable indicating ordered response categories, such as ambulation score in Table 1. If the variable takes many distinct values it can be treated as a quantitative variable and described in terms of centre and spread: ordinal variables often extend from the minimum to maximum possible values and in this case stating the range is not helpful. The meaning of the extremes should be stated in the context of the table to aid interpretation of results. Ordinal variables taking only a few distinct values are better treated as categorical variables and number (%) presented for each category. With only five categories this latter approach was adopted for ambulation score. Display as a categorical variable can be facilitated by combining infrequently occurring adjacent values.

## Describing loss of participants in a study

Readers will be better able to assess the generalizability of results if they can see how the participants contributing to analysis relate to the patient base from which they were drawn. Eligibility criteria and the approach to recruitment are detailed in the methods section of a paper, and their consequences in reducing the numbers available for analysis are shown at the start of the results section. This can be done in text, as in the excerpt below describing how total admissions were reduced to the sample from which rates of recovery from delirium after discharge were estimated:

> "*In the original study, 3,182 of 5,719 admissions were screened and 2,286 were eligible. Six hundred and ten patients were not available on the hospital units when the RA [Research Assistant] arrived to complete the CAM [Confusion Assessment Method]; 1,582 patients assented to complete the CAM and 94 patients did not assent; the CAM was not completed for 728 patients because an informant was not available to confirm an acute change and fluctuation in mental status prior to admission or enrolment. The CAM was completed for 854 patients; 375 had delirium; 278 were enrolled. Of the 278 enrolled patients, 172 were discharged before the follow-up assessment, 73 were still hospitalised, 8 withdrew from the study and 27 died. Of the 172 discharged patients, delirium recovery status was determined for 152, 16 withdrew from the study after discharge and 4 died."* [4]

The authors start with the 5,719 admissions and report the numbers lost at successive stages, to arrive at the analysis sample of 152. It may be easier to assimilate the detail of the process from tabular or graphical presentation. The CONSORT guidelines [5] concerning the reporting of randomized controlled trials (RCT) recommend that progress of participants through a trial be presented as a flow chart, and an example is

shown in Figure 2. These charts are unequivocally helpful and are now presented in studies other than RCTs.

In addition to loss of participants at each time point as shown in a flow chart, information on specific variables may be missing even though a participant was available at the study point in question. Taking Table 1 as an example, there were 395 and 373 admissions during the 1998/999 and 2000/1 phases respectively, as stated in the column headings, but the number of participants providing information varies considerably across the characteristics in the table. The reader should be able to establish how many cases contribute to each result, and to this end wherever the number available is lower than the total for the phase, it is stated below the descriptive statistics. For example, ambulation score was only available for 390 of the 395 participants in the 1998/99 phase. The percentages presented for ambulation score were calculated amongst cases where information was available, and this was done for all percentages in the table as indicated in the title. Alternatively, missing values in a categorical variable may be treated as a category in their own right. Where there is a large amount of missing information, this may be the best way of handling the situation with percentages calculated from the total sample size as denominator. Stating the numbers available allows the reader to check this point. Only participants whose operation was delayed by more than 48 hours, gave a "reason why operation was delayed" in the table, and from the stated numbers the reader can see that a reason was not given for all delayed cases.

## Comparing baseline characteristics in randomized controlled trials

In reports of RCTs, a table describing baseline characteristics in each trial arm demonstrates whether or not randomisation was successful in producing similar groups, as well as addressing the generalizability issue. If there are differences at baseline, comparisons of outcome may be confounded. Statistical tests of significance should not be used to decide whether any differences need to be taken into account [7, 8]. If the allocation was properly randomized we know that any differences at baseline must be due to chance. The question facing the researcher is whether or not the magnitude of a difference at baseline is sufficient to confound comparison of outcome, and this depends on the strength of the relationship between the potential confounder and the outcome, as well the baseline difference. A statistical test for baseline differences does not address this question, furthermore there may be insufficient numbers available to detect quite large baseline differences. Statistics describing baseline characteristics are used to judge whether any differences are large enough to be important. If they are, additional analyses of outcome controlled for characteristics that differ at baseline may be performed. On the other hand, in non-randomised studies, groups are likely to differ, and statistical significance tests can be used to evaluate the evidence that the selection process of patients to each intervention results in different groups. In this situation a primary analysis controlled for many predictors of outcome would probably have been planned, and should be carried out irrespective of any differences, or lack of them between study groups.

## Conclusions

Describing the main features of the distribution of important characteristics of the participants included in a study is the first step in most papers reporting statistical analysis. It is important in establishing the generalizability of research findings, and in the context of comparative studies, flags the need for controlled analysis. Usually space constraints limit the presentation of many descriptive statistics, and in any case, too many statistics can confuse rather than enhance insight. The attrition of subjects during a study should also be described, so that study subjects can be related to the patient base from which they were drawn.

## Acknowledgements

The author would like to thank Dr Helen Roberts for kindly granting permission to use data from the care pathway study [3] to produce Figures 1c) and 1d).

## References

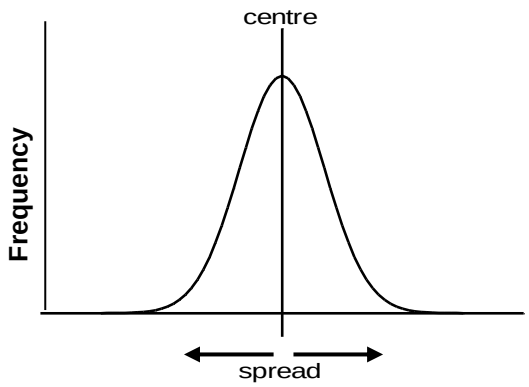

1. Altman DG. Practical Statistics for Medical research. London: Chapman & Hall, 1991.
2. Yeung P, Breheny M. Using the capability approach to understand the determinants of subjective well-being among community-dwelling older people in New Zealand. Age and Aging 2016; 45: 292-8.
3. Roberts HC, Pickering RM, Onslow E, Clancy M, Powell J, Roberts A, Hughes K, Coulson D, Bray J. The effectiveness of implementing a care pathway for femoral neck fracture in older people: a prospective controlled before and after study. Age and Aging 2004; 33: 178-84.
4. Cole MG, McCusker JM, Bailey R, Bonnycstle M, Fung S, Ciampi A, Bezile E. Partial and no recovery from delirium after hospital discharge predict increased adverse events. Age and Aging 2017; 46: 90-5.
5. Schulz KF, Altman DG, Moher D, for the CONSORT Group (2010) CONSORT 2010 statement: updated guidelines for reporting parallel-group randomised trials. BMJ, 340, 698-702.
6. Kwok BC, Pua YH. Effects of WiiActive exercises on fear of falling and functional outcomes in community-dwelling older adults: a randomised control trial. Age and Aging 2016; 45: 621-28.
7. Assman SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-9.
8. Altman DG. Comparability of randomized groups. Statistician 1985; 34: 125-36.

**Table 1.** Characteristics of subjects at admission and their operations before (1998/99) and after (2000/01) implementation of a care pathway. Figures are number (% of non-missing values) unless otherwise stated [3]

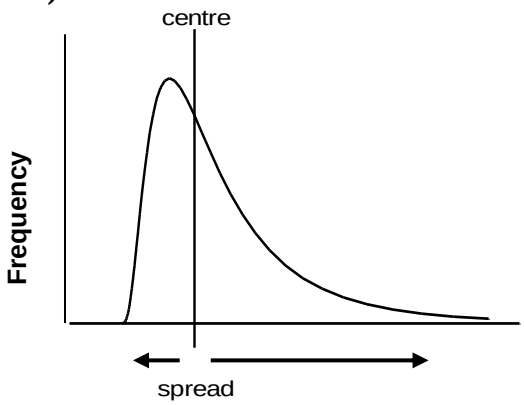
| | | 1998/99 (*n*=395) | 2000/01 (*n*=373) |
|---|---|---|---|
| Age on admission (yrs) | mean(SD) | 83 (7) | 83 (7) |
| | min-max | 65-101 | 65-101 |
| Gender | male | 90 (23%) | 90 (24%) |
| | female | 305 (77%) | 283 (76%) |
| Admission domicile | own home | 219 (55%) | 202 (54%) |
| | sheltered accommodation | 47 (12%) | 58 (16%) |
| | residential care | 90 (23%) | 83 (22%) |
| | nursing home | 18 (5%) | 15 (4%) |
| | other ward SUHT | 7 (2%) | 2 (1%) |
| | other trust | 14 (4%) | 13 (4%) |
| Ambulation score | bed/chair bound | 8 (2%) | 5 (1%) |
| | presence 1+ | 12 (3%) | 7 (2%) |
| | 1 person | 25 (6%) | 20 (5%) |
| | unable 50 metres | 145 (37%) | 138 (38%) |
| | able 50 metres | 200 (51%) | 197 (54%) |
| | | (*n*=390) | (*n*=367) |
| Time in A&E (hrs) | mean(SD) | 4.9 (3.2) | 5.6 (2.4) |
| | min-max | 1.7-40.3 | 0-21.4 |
| | | (*n*=348) | (*n*=328) |
| History of dementia | | 79 (20%) | 85 (23%) |
| | | (*n*=395) | (*n*=371) |
| Confused on admission | | 124 (32%) | 125 (34%) |
| | | (*n*=394) | (*n*=371) |
| Type of fracture | intra-capsular | 192 (54%) | 173 (52%) |
| | extra-capsular | 165 (46%) | 161 (48%) |
| | | (*n*=357) | (*n*=334) |
| Operation more than 48 hours after ward admission | | 183 (52%) | 205 (64%) |
| | | (*n*=354) | (*n*=323) |
| Reason for delayed operation | medical | 61 (35%) | 74 (43%) |
| | organisational | 66 (38%) | 72 (42%) |
| | both | 45 (26%) | 27 (16%) |
| | | (*n*=172) | (*n*=173) |
| Type of operation | Thompson's hemiarthroplasty | 101 (27%) | 87 (24%) |
| | Austin-Moore hemiarthroplasty | 69 (19%) | 18 (5%) |
| | dynamic screw | 162 (43%) | 165 (46%) |
| | asnis screws | 38 (11%) | 38 (11%) |
| | bipolar hemiartroplasty | 3 (1%) | 48 (14%) |
| | | (*n*=373) | (*n*=356) |
| Grade of surgeon | consultant | 46 (12%) | 110 (32%) |
| | SPR | 318 (86%) | 220 (63%) |
| | SHO | 6 (2%) | 18 (5%) |
| | | (*n*=355) | (*n*=348) |
| Grade of anaesthetist | consultant | 1206 (34%) | 175 (55%) |
| | SPR | 99 (28%) | 52 (16%) |
| | SHO | 133(38%) | 81 29%) |
| | | (*n*=352) | (*n*=318) |

**Figure 1.** Idealized and real data distributions

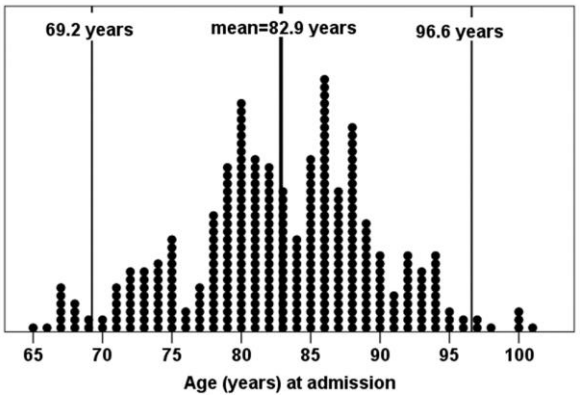
**a) symmetrical distribution**




**b) skewed distribution**




**c) dotplot (each dot representing one value) of an approximate symmetrical distribution indicating the Normal range: age in years at admission (n=373)**




**d) dotplot (each dot representing one value) of a skewed distribution with outliers emphasised and indicating mean and median: hours in A&E (n=348)**

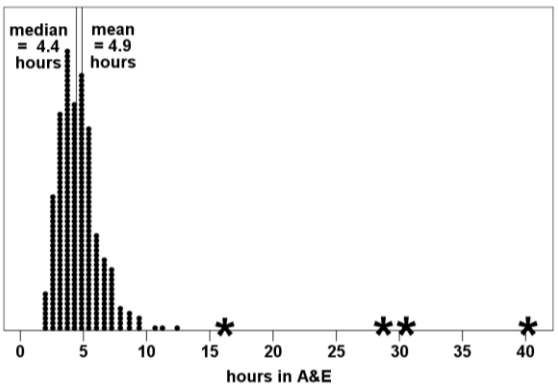


**Figure 2.** Recruitment and attrition rates in an RCT of WiiActive exercises in community dwelling older adults [6]