Early Human Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

WASP (Write a Scientific Paper) using Excel –5: Quartiles and standard deviation

ARTICLE INFO

Keywords:
Software
Computers, statistics
Biostatistics

ABSTRACT

The almost inevitable descriptive statistics exercise that is undergone once data collection is complete, prior to inferential statistics, requires the acquisition of basic descriptors which may include standard deviation and quartiles. This paper provides pointers as to how to do this in Microsoft Excel™ and explains the relationship between the two.

1. Introduction

A population is usually too large to study in-toto so a representative sample is utilised. Naturally, care must be taken so that the sample truly represents the population of interest, i.e. the one that is actually being studied, so that results from the sample comprise a meaningful representation of the target population, and useful extrapolations may be derived about the population in question.

Data comprising a sample can be shown pictorially as a graph, such as a boxplot and whisker or a histogram. In the former, standard and relatively simple mathematical notation is also essential and includes terms such as range (upper and lower levels), mean, median, standard deviation and confidence interval. Indeed, “the purpose of statistical methods is to put numerical data into a context by which their meaning can be better judged… No mathematical knowledge beyond simple arithmetic is required” [1]. This permits the construction of meaningful tables that can be understood by all (Table 1) [2]. It is often standard deviation and confidence interval that confuse novices, but their meaning and derivation will be described.

2. Quartiles

For data to be compared, it must first be understood in its spread and ramifications. There are several measures that help to define and therefore comprehend the pattern of quantitative datasets. The range is simple: the smallest and largest values. However, these values do not give much indication of the spread of observations about the mean. Quartiles are helpful and the first to fourth quartile (Q1 through to Q4) are defined by three boundary values: the points which comprise the 25%, 50% and 75% of the ranked dataset (i.e. values listed in sorted order from smallest to largest). The middle ranked value is the median. The interquartile range is the distance between the first and third quartiles (IQR = Q3 − Q1 i.e. the difference between 75th and 25th percentiles) and is one of the most significant and basic robust measures of spread.

Excel function quartile calculations for a urinary lead sample are shown Table 2. The fourth column contains the exact equations as implemented in the third column. Note that the median function result is identical to that which calculates the second quartile.

The uppermost and lowermost values, as well as the quartiles can be used to construct a boxplot and whisker graph from Excel 2016 onward. First select the data to be plotted, then in the Insert tab, in the Illustrations group, click Chart, and from the All Charts tab, click Box & Whisker.

Boxplots may be plotted in one of two ways. The first method is to plot the uppermost and the lowermost datapoints as the highest and lower whisker respectively, thus encompassing all data within the boxplot and whisker. Alternatively, in order not to let outliers distort the shape of the whiskers, the whiskers may be used to incorporate values that are not outliers. Outliers are defined as datapoints that fall below Q1–1.5xIQR and datapoints that fall above Q3 + 1.5xIQR. Thus, in the latter type of plot, whiskers would denote the upper and lower ranges that do not include outlier values while outliers would be indicated as individual points. These calculations are also shown in Table 2 (fourth column middle and bottom).

3. Standard deviation

Quantitative datasets, such as many biological characteristics, tend to be normally distributed. Based on this precept, the calculation of standard deviation provides a foundation for data interpretation using probabilities. This is because the range covered by one standard deviation, above and

https://doi.org/10.1016/j.earlhumdev.2018.01.012

Table 1
Paediatric visceral leishmaniasis in Malta (1980-1998) - haematological and biochemical indices at presentation [2].

	Median	*Mean*	*SD*	*Minimum*	*Maximum*	*Normal values*
Haemoglobin (g/l)	76	79	18	47	179	105–145
Mean cell volume (fl)	70	68.8	13.2	7.1	88	78–94
Mean cell haemoglobin (pg)	23	23.8	3.5	17	34	23–30
Total white count ($\times 10^9$/l)	4	4.8	2.9	1.1	14.5	6–17.5
Neutrophil count (%)	24	25.6	11.3	9	66	54–62
Lymphocyte count (%)	66	64	13.1	21	87	25–33
Monocyte count (%)	7.5	9.5	6.4	1	40	3–7
Platelet count ($\times 10^9$/l)	111	121.1	53.2	25	292	150–400
Erythrocyte sedimentation rate (mm/h)	95	89	34	10	150	0–13
Bilirubin (μmol/l)	15	15	7	5	50	<10
Alkaline phosphatase (U/l)	227	263	138	26	683	100–350
γ-glutamyltranspeptidase (U/l)	10	15	22	1	124	<40
Alanine aminotransferase (U/l)	27	49	100	2	680	0–35
Total protein (g/l)	74	74.2	10.5	56	95	57–80
Albumin (g/l)	31	30.3	7.7	2.7	45	33–47
γ globulin (g/l)	22.2	23.6	17.1	1.8	70	17–38

Table 2
Urinary lead quartiles.

	A	B	C	D
1	Urinary lead (umol/24 h)			
2	0.1	1st quartile	0.95	=QUARTILE(A3:A17,1)
3	0.4	2nd quartile	1.5	=QUARTILE(A3:A17,2)
4	0.6	3rd quartile	1.95	=QUARTILE(A3:A17,3)
5	0.8	Median	1.5	=MEDIAN(A3:A17)
6	1.1	Interquartile range	1	=C5-C3
7	1.2			
8	1.3			Outliers
9	1.5			A. Above
10	1.7			Q3 + 1.5xIQR
11	1.9			=C5 + (1.5*C7)
12	1.9			3.45
13	2			B. Below
14	2.2			Q1–1.5xIQR
15	2.6			=C3-(1.5*C7)
16	3.2			−0.55
17				
18				
19				
20				

below the mean, incorporates circa 68% of the data, a range of two standard deviations above and two below the mean incorporates just over 95% of the data, and a range of three standard deviations above and two below the mean incorporates just over 99% of the data.

The standard deviation of a sample is an estimate of the population's standard deviation. The variability or spread of the data pattern will therefore be approximately the same for both sample and population. The sample standard deviation will not decrease with increasing sample size, but will more accurately reflect the population standard deviation.

Standard deviation is conventionally represented by the symbol σ (sigma), and its calculation requires the subtraction of each datapoint from the mean (Table 3, third column) and the conversion of values below the mean (negative relative to the mean) to positive values. This can be done by squaring the difference of each datapoint from the mean (Table 3, fourth column). The sum of the squared values divided by the number of datapoints minus 1 provides the variance. However, this is in units squared. The standard deviation is obtained by taking the square root of the variance (see formulas at on left of table). The other calculations will be discussed later.

4. Utility of standard deviation

Standard deviation permits the calculation of certain probabilities. E.g. 95% of the values in the dataset (and by extension, in the population) fall within the range of 1.96 multiplied by the standard deviation (0.84 in this example), above and below the mean (1.5 in this example). Thus, in this case 95% of values in this sample fall within the range $1.5 \pm (1.96 \times 0.84)$. This is useful, for example, in the calculation of reference ranges.

Indeed, the following applies:

- The mean ± 1 SD incorporates 68.2% of the data.
- The mean ± 2 SD incorporates 95.4% of the data.
- The mean ± 3 SD incorporates 99.7% of the data.

This is often referred to as the 68–95–99.7 rule.