VISUALISING THE EFFECT OF SAMPLE SIZE ON DESCRIPTIVE STATISTICS: WITH APPLICATION TO A COUNTERMOVEMENT JUMP TEST

Nicholas P. Linthorne

Division of Sport, Health and Exercise Sciences, Department of Life Sciences, Brunel University London, Uxbridge, United Kingdom

Students usually develop an intuitive feel for the interpretation of experimental data relatively slowly through analysing different data sets. The aim of this paper is to encourage educators to develop their student's expertise more rapidly by using visual representations of statistical concepts and by having the student visually examine many sets of simulated data. The jump height achieved in a countermovement jump is used to illustrate this approach. Some students report that viewing many sets of simulated data in quick succession turbo-charged their expertise as a researcher. It is also shown that calculating the mean of 5–8 jump trials is a good compromise between reducing the uncertainty in the mean and the time spent testing.

KEYWORDS: confidence interval, confidence limits, mean, standard deviation, statistics.

INTRODUCTION: Most university sports biomechanics courses introduce students to the practice of quantitative research. This is usually done through classes on research methods, through laboratory practical sessions, and through conducting small research projects under the supervision of an experienced researcher. One of the aims is for the student to develop an intuitive feel for the analysis and interpretation of experimental data. In particular, students are expected to become familiar with the statistics that describe a data set, especially measures of central tendency (mode, median, mean) and measures of variability (range, quartiles, variance, standard deviation). Students are also expected to become familiar with the confidence limits for a statistical parameter, appreciate the need to include error bars on a plot of the data, and be able to report data with an appropriate number of significant figures.

Usually, new researchers develop an intuitive feel for statistical concepts relatively slowly through analysing different data sets. In this paper I encourage educators to enhance their students' understanding through examining visual representations of statistical parameters. I also encourage educators to accelerate their students' intuitive feel through visually examining a large number of simulated data sets rather than relying on a mathematical analysis of a small number of real data sets. My messages echo those of researchers in mathematics and statistics education. Studies have demonstrated the potential of visual and graphical representations to clarify concepts and to reveal previously inaccessible concepts (Arcavi, 2003), and studies have demonstrated that viewing simulated data sets can draw a student's attention to aspects of a situation that might not be evident under normal conditions (Mills, 2002).

This paper presents visual representations of the effect of sample size on selected descriptive statistical parameters. Many studies in sports biomechanics have a relatively small sample size; often less than 50, and sometimes less than 20. Therefore, it is important for the student to appreciate that the uncertainty in a statistical parameter can be large if the sample size is very small. A student's intuitive feel for the effect of sample size and randomness in data can be accelerated through viewing a large number of simulated data sets. A useful feature of *Microsoft Excel* is that sets of simulated data can be easily generated, and plots of these data sets can be viewed in quick succession by repeatedly saving the data file.

To illustrate the main concepts I have used a specific example from sports biomechanics; the jump height achieved in a countermovement jump on a force platform. This performance test is familiar to most sports biomechanics students, and the plots generated for this example are qualitatively similar to other examples with different numerical values of the statistical parameters.

METHODS: Consider a set of samples of the quantity *X* from a population that has a normal distribution with a mean, $\bar{X}$, and a standard deviation, σ. The uncertainties in the statistical parameters $\bar{X}$ and σ depend on the sample size, *n*. The upper and lower confidence limits for the mean can be calculated using (Motulsky, 2018)

$$\bar{X} - \frac{t_{1-\alpha,n}\,\sigma}{\sqrt{n}} \leq \bar{X} \leq \bar{X} + \frac{t_{1-\alpha,n}\,\sigma}{\sqrt{n}}, \qquad (1)$$

where $t_{1-\alpha,n}$ is the critical value of the *t*-distribution for that level of significance (α) and sample size (*n*). The upper and lower confidence limits for the standard deviation can be calculated using (Sheskin, 2011)

$$\frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(\alpha/2)}}, \qquad (2)$$

where *s* is the sample standard deviation and $\chi^2_{(\alpha/2)}$ is the critical two-tailed value in the chi-square distribution below which a proportion equal to 1–α/2 of the cases falls.

The 90%, 95%, and 99% confidence limits of the mean and standard deviation of a countermovement jump were calculated using equations 1 and 2, and plotted as a function of the sample size. The $t_{1-\alpha,n}$ and $\chi^2_{(\alpha/2)}$ values were calculated using the 'two-tailed inverse of the Student's *t*-distribution' function and the 'inverse of the chi-squared distribution' function in *Microsoft Excel* (with *df*=*n*–1).

Microsoft Excel was also used to generate many simulated data sets of a countermovement jump that could be rapidly viewed by the student. Again, the 'random' function and the 'inverse cumulative standardised normal distribution' function were used to generate a normally distributed data set with a known mean and standard deviation, and a plot of the data for various sample sizes was produced. A useful feature of *Excel* is that the plot is updated with a new data set each time the file is saved. Thus, the student was able to visually examine many data sets in quick succession by repeatedly saving the file. Also, the cumulative mean and cumulative standard deviation of the simulated data set were calculated and plotted as a function of sample size. Again, the student visually examined many data sets in quick succession by repeatedly saving the file.

RESULTS: Visual representations of how the confidence limits for the mean and standard deviation of a countermovement jump depend on the sample size are shown in Figure 1. As the sample size increases, the uncertainty in the mean decreases toward zero (as expected). The uncertainty steadily decreases with more than about 10 samples but increases rapidly when there are fewer samples. The confidence interval is symmetrical about the mean value. Likewise, as the sample size increases the uncertainty in the standard deviation decreases toward zero (as expected). However, the confidence interval is not symmetrical about the standard deviation. Again, the uncertainty steadily decreases with more than about 10 samples and increases rapidly when there are fewer samples.

Figure 2 shows an example of a simulated data set generated using *Microsoft Excel*, and Figure 3 shows a plot of cumulative mean and cumulative standard deviation as a function of sample size. When a large number of data sets are generated and viewed, most of the data sets lie within the confidence limits (as expected).

DISCUSSION: Anecdotal reports from my students suggest that the visual representations helped them develop a deeper understanding of statistical concepts. Many research studies in education have shown that learning can be increased when the student is given an artefact (such as a diagram) to examine, which can then stimulate a discussion among a group of students. Also, most learners prefer to consider concrete examples which they can link to previous experience before proceeding to abstract concepts or mathematical concepts. In most science courses the student performs statistical calculations on only a limited number of data sets. When this is the only avenue for developing expertise, the student does not always develop a broad feel for the statistical parameters. Viewing plots of

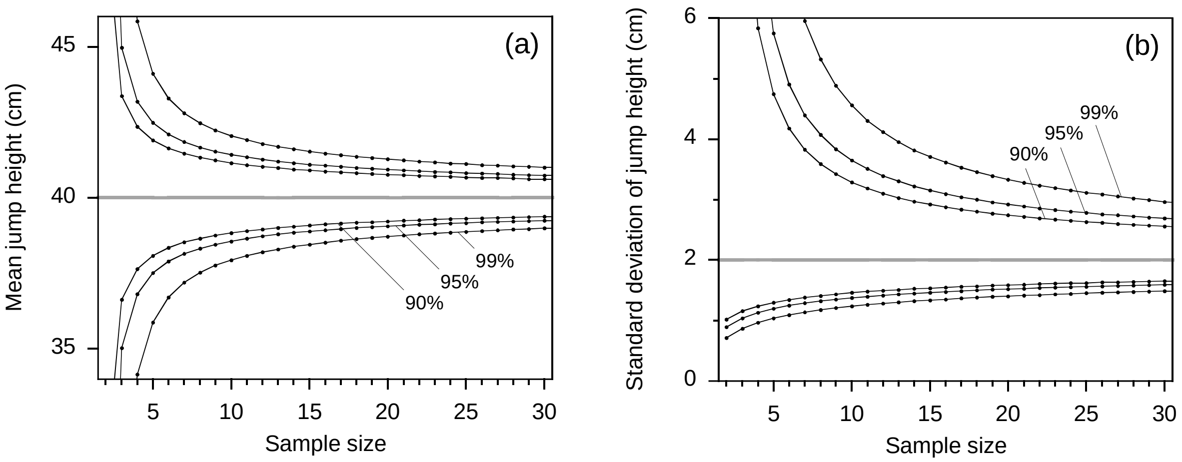

Figure 1: Plot of the statistical parameter (grey line) and the upper and lower confidence limits (black lines) as function of sample size: (a) mean; (b) standard deviation. Results generated using equations 1 and 2. Calculations are for a jump height with a mean $\bar{X}$ = 40 cm and standard deviation σ = 2 cm.

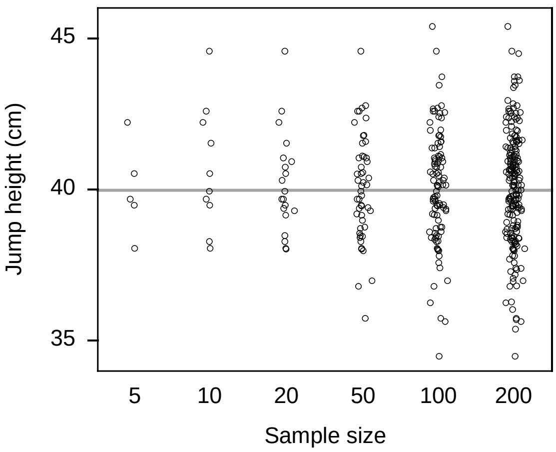

Figure 2: Plot of a simulated data set for selected sample sizes. When the data are calculated using *Microsoft Excel* the student can visually examine a sequence of 50 different data sets in only a few minutes by repeatedly saving the file.

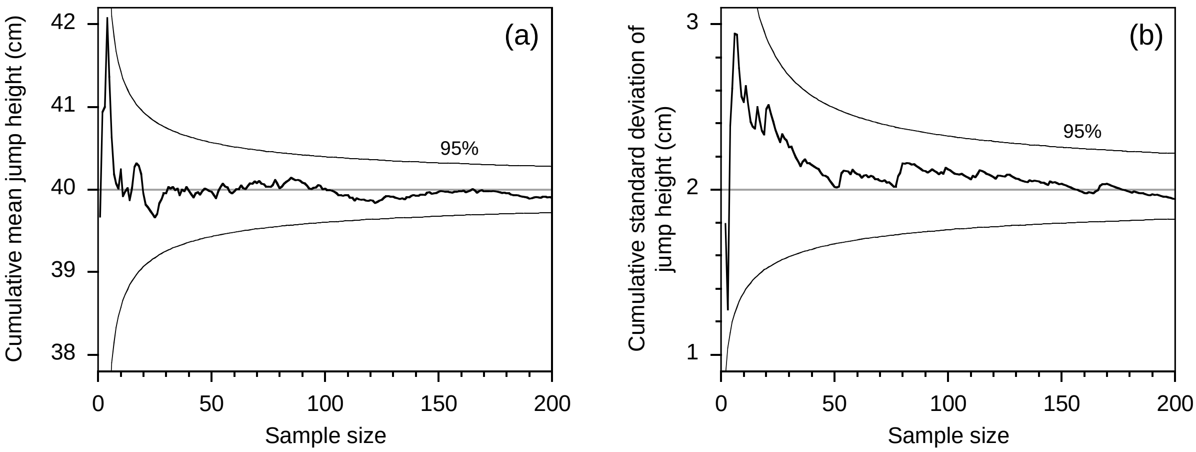

Figure 3: Plot of (a) the cumulative mean and (b) the cumulative standard deviation as function of sample size (thick black line) for a simulated data set. Also shown are the upper and lower 95% confidence limits. The student can visually examine a large number of different data sets in a short time by repeatedly saving the file.

statistical parameters in a concrete context, and the ensuing discussions, can allow the student to grasp the bigger picture more quickly. In the example presented in this paper the students examined the mean and standard deviation. However, other statistical parameters such as the confidence limits in a correlation coefficient can be represented visually.
Some of my students reported that viewing many sets of simulated data in quick succession turbo-charged their expertise. This method was particularly effective in generating a feel for randomness in data and highlighting that what initially might look like a pattern in the data can actually be just random noise. In the example presented in this paper the student examined the mean and standard deviation. However, simulated data sets can also be used to examine the effect of sample size and randomness on the normality of the data by viewing a histogram, a box and whisker plot, and a normal Q-Q plot. Likewise, the effect of sample size and randomness can be visually investigated in a correlation plot, a regression plot (with confidence limits), and a Bland-Altman plot.
The diagrams presented here have some useful applications in sports biomechanics research. The uncertainty in the standard deviation is a key statistical result and should be used in deciding how to report the results with an appropriate number of significant figures. Even experienced researchers do not always appreciate that a standard deviation has an uncertainty and that this uncertainty depends on the sample size. With sample sizes of n = 20, 50, 100, 200, and 20,000 the uncertainty (±95% CI) in the standard deviation is 35%, 21%, 14%, 10%, and 1% respectively. Therefore, with a small sample size (<200) we are usually justified in reporting a standard deviation to only one significant figure (e.g., 5 cm, rather than 5.3 cm or 5.31 cm). When reporting a mean value and its standard deviation we should round the standard deviation to an appropriate number of significant figures, then round the mean to the same number of decimal places as the standard deviation (e.g., 61 ± 5 cm, rather than 61.3 ± 5 cm). I conducted a quick review of sports biomechanics studies reported in scientific journals. In many studies the data is reported with too many significant figures and hence implies an accuracy in the data that is unwarranted.
In this paper I have used the jump height achieved in a countermovement jump to illustrate statistical concepts. We can use the diagrams presented in this paper to decide upon the number of trials that should be performed in a countermovement jump test. When monitoring jump performance in an athlete it is better to calculate the mean of a number of trials rather than to record just the best trial (Al Haddad, Simpson & Buchheit, 2015; Kennedy & Drake, 2018). However, in a jump testing session the athlete cannot perform many trials (more than about 10) without inducing fatigue or taking a long time to complete the test (more than about 10 minutes). From Figure 1a we see that the confidence interval in the mean is relatively large when there are fewer than five trials. Therefore, calculating the mean of 5–8 jump trials is a good compromise between reducing the uncertainty in the mean and the time spent testing. This result (5–8 samples) is a useful guideline that can be applied to many other instances of measurement in sport and exercise science.

CONCLUSION: Examining visual representations can help the student to develop a deeper understanding of statistical concepts. Visually examining a large number of simulated data sets can turbo-charge the student's expertise as a researcher.

REFERENCES

Al Haddad, H., Simpson, B.M. & Buchheit, M. (2015). Monitoring changes in jump and sprint performance: Best or average values? *International Journal of Sports Physiology & Performance,* 10, 931-934.
Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics*, 52, 215-241.
Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, 10, http://www.amstat.org/publications/jse/v10n1/mills.html.
Motulsky, H. (2018). Intuitive biostatistics, 4th ed. Oxford: Oxford University Press.
Sheskin, D.J. (2011). Handbook of parametric and nonparametric statistical procedures, 5th ed. New York: Chapman and Hall/CRC.