The StatPlay Software for Statistical Understanding: Confidence Intervals and Hypothesis Testing

Geoff Cumming[1], Neil Thomason[2], Andrew Howard[1], John Les[1] and Mark Zangari[1]
1 - School of Psychology, La Trobe University, Bundoora, Australia 3083
psygdc@latrobe.edu.au
2 - Department of History and Philosophy of Science,
University of Melbourne, Parkville, Australia 3052

Abstract

StatPlay is a collection of computer-based demonstrations and interactive simulations intended to promote cognitive change and good understanding of central aspects of statistics and experimental design. StatPlay is intended for use in a variety of ways, including lecture demonstrations, by groups and by individuals. In addition to free exploration, challenging tasks—some in game formats—offer structure and guidance to the learner's activities. With support from the Committee for the Advancement of University Teaching (CAUT) StatPlay is being developed in Visual C++ for Windows. It currently comprises microworlds, or Playgrounds, for working with: discrete datasets; continuous distributions; sampling; confidence intervals; and hypothesis testing. Recent development has focussed on representations and activities that help students acquire good understanding of confidence intervals, simple hypothesis testing, and the relation between these. We will demonstrate the software, describe use of StatPlay by students and report an experiment evaluating students' conceptual change.

Keywords

conceptual change, cognition, statistics learning, microworlds, interactive simulations, evaluation

1. StatPlay: Outline of rationale and design

The rationale and basic design of StatPlay were described by Thomason, Cumming and Zangari (1994). We give a brief outline of this rationale then describe the software and discuss several design issues closely tied to choice of how some fundamental statistical issues are presented pedagogically.

1.1 Statistical misconceptions: The Law of Small Numbers and statistical significance

Misconceptions of statistics and probability are widespread, resistant to conventional education and have damaging consequences. Tversky and Kahneman (1971) presented evidence of a fundamental misconception about randomness: people generally underestimate the amount of variability from sample to sample, overestimate the similarity of a sample to the population and under-estimate the width of confidence intervals. They fail to realise the crucial role of *N*, the sample size. Only when *N* is very large are samples in general closely similar to each other and to the population, and experimental results repeat: this is the Law of Large Numbers, a theorem of mathematical statistics. By analogy, Tversky and Kahneman concluded that people often behave as if they believe in a Law of Small Numbers (LSN), a 'law' of human behaviour easily shown to be discrepant with the world.

Misunderstanding of hypothesis testing has been discussed by Gigerenzer (1993), Oakes (1986) and others. One problem is confusion between statistical significance and size or importance; another is interpretation of non-significance as meaning the null hypothesis is true; another is the belief that repeating an experiment with a significant result is likely again to give significance. More basically, significance is often wrongly interpreted as the probability the null hypothesis is true rather than the *conditional* probability of obtaining extreme results *if the null hypothesis is true*.

We suggest two core issues underlie statistics misconceptions, and identify these as teaching targets:

- LSN misconceptions about sampling variability and the way this depends on *N*, and
- problems with probability, especially conditional probability.

1.2 Naive Statistics: The analogy with Naive Physics

Naive physics refers to intuitive beliefs people use to guide their expectations about everyday events (McCloskey, 1983). Some naive physics beliefs are persistent, but wrong. For example, many people believe a force is necessary for motion to continue, or that a ball flying around on a string will take a curved path after the string breaks. Traditional education diminishes the influence of naive beliefs, but even physics graduates often have everyday expectations that accord with naive physics.

By analogy we introduced the term 'naive statistics' for everyday beliefs about probability and statistics (Thomason et al., 1994). As with physics, these beliefs seem to have arisen through normal life experience. In both cases the naive beliefs are used, usually unconsciously, to guide everyday expectations about the world: in one case about moving objects, in the other perhaps about reaching conclusions after seeing a small sample—perhaps even just one or two cases. Because there is little research on teaching statistical concepts (Shaughnessy, 1992) we look for guidance to the analogy with naive physics, and to science education generally.

1.3 Lessons from science education and naive physics

White (1993) described ThinkerTools, a collection of microworlds for 11-12 year olds learning elementary physics, in which learners undertook modelling and game-like activities about forces and moving objects. Children formulated and tested hypotheses, and confronted any mistaken beliefs. White presented impressive evidence for the effectiveness of her approach. White expressed several recommendations succinctly: 'Employ manipulable, linked representations for key abstractions. … Make the phenomena easy to see and interpret. … Reify the knowledge to be acquired.' (pp. 49-50) The most general lesson is the notion of conceptual change (West and Pines, 1985): learning will be more successful if we take account of initial naive beliefs.

1.4 StatPlay: Design issues

StatPlay is intended to help learners replace naive statistics beliefs with correct conceptions. It is not intended to provide a complete curriculum, but for use alongside textbooks and other software. StatPlay should be usable for lecture demonstrations, by a single learner and by a pair or small group. Sampling variability is the first concept addressed, as the basis for work on confidence intervals, power and statistical significance testing. Later we will address conditional and other probabilities.

We adopt as some of the central design principles for StatPlay:

- Abstract symbolic representations should be used, and tied to graphic, concrete representations the user manipulates. (This is White's 'Reify the knowledge…' principle.)

- Multiple representations are needed to integrate theoretical and practical understanding: multiple representations yoked together offer a powerful strategy, and facilitate transfer to the real world.
- Interest and engagement should be maximised; game formats can be valuable.
- Guidance for structured learning activities and on-line advice should always be available.
- The software should feel open, and encourage learner choice, initiative and exploration.

2. StatPlay: The current implementation

StatPlay is being developed in Visual C++ under Windows. Four microworlds, or playgrounds, have been implemented. These are described briefly below and illustrated in the Figures.

2.1 The Data Playground

A simple list of data values is shown (see Figure 1) also as an ordered list, a frequency histogram and a dot plot. These four representations or views are linked: select by mouse any data values or values in one view and see the corresponding values in all four representations highlighted. A learner can enter their own dataset, load a set from disk or generate a set from one of several distributions. A wide variety of display options and activities are available. Learning aims include understanding of frequency histograms, dot plots, a range of descriptive statistics, percentiles and *z*-scores.

2.2 The Continuous Distribution Playground

Figure 2 shows a normal distribution, but a distribution of any shape may be investigated. Numeric values of mean, SD, skew and tail probability can be shown. Dragging handles (small squares on the X and Y axes) causes the display to change smoothly: it 'feels plastic under your mouse fingers'. The numeric values change dynamically also; multiple representations (curve and numeric values) are thus coupled. Typing in new values causes the curve to change, so the linking is symmetric.

In the game formats you compete against the clock, and possibly a second learner, to estimate the mean and SD of any continuous distribution, or areas and *z*-scores. You are confronted with feedback from mistaken estimates and can guess again.

2.3 The Sampling Playground

The Sampling Playground is the first part of our attack on LSN. It shows in the upper part of the screen the population and in the lower part some representation of samples taken (Figures 3-5). Data panels give information about the population, the sampling process, and the current state. The rate slider allows choices ranging from slow step-by-step sampling to rapid taking of a series of samples.

In Figure 3 each line in the scrolling window shows a dot plot of one sample and its mean (inverted triangle). Means vary greatly, as expected with a small sample size ($N = 4$). Figure 4 shows confidence intervals (shaded bars) for the population mean (vertical line). Figure 5 gives an alternative view, for $N = 20$: two frequency distributions are shown—that of all 2000 values and that of the 100 means. This sampling distribution has approximately a normal distribution, illustrating the Central Limit Theorem. The scrolling window or the frequency distribution view can be watched as it changes dynamically during repeated sampling, with rate under control of the slider set by the user.

If the number of samples taken is set to infinite, the frequency distribution representation shows that the distribution of individual values matches the population shape exactly, while (for most continuous population distributions) the sampling distribution appears smooth, and approaching normality.

2.4 The Hypothesis Testing Playground

If the data come from a distribution that exists in the computer, this distribution can remain hidden (the upper dark panel in Figure 6) or can be displayed. Meanwhile the user conjectures a distribution (the Null Hypothesis for example) and the sampling distribution derived from this is shown in the middle panel, allowing the sample—shown as a dot plot at the bottom—to be used to make a hypothesis test of the conjectured distribution.

3. Key design issues for statistics learning

3.1 Probabilistic simulations and the Law of Small Numbers

Demonstrations and simulations in physics and most other sciences are deterministic: on repetition exactly the same thing should occur. In striking contrast, simulations of sampling and other probabilistic processes must vary from occasion to occasion, showing regularity only in the long run. So, if a learner simply takes one or two samples, appreciation of vital aspects of sampling, such as the likely discrepancy between sample and population, and the extent of variation from sample to sample, may well not be illustrated in any dramatic way.

In fact the learner's LSN misconception is likely to impede understanding: the learner will, under the influence of LSN itself, jump to a conclusion on the basis of the first, or the first few samples seen, even though these are likely by chance to give an erroneous impression. Belief in LSN thus seems to be self-perpetuating, as Tversky and Kahneman (1971, p. 109) noted.

There is fundamental conflict between our desire to work with single cases and single samples to keep things simple and concrete for the learner, and the necessity to consider rarefied concepts such as 'the set of infinitely many samples'. Difficult entities such as this underlie the idea of variability from sample to sample, which is the key concept that LSN severely under-estimates, or fails to recognise.

The difficulty can be stated simply: the appeal and potency of learning environments lie in the use of simple, concrete events. Many fundamental statistical ideas, however, are broad properties of *sets* of events. Furthermore, they are uncertain or variable properties. Our example so far has been the variation from sample to sample but the point applies also to conditional probabilities and even to simple probabilities, as in Shute and Gawlick-Grendell's (1993) Stat Lady: appreciating a probability of 0.70 for example requires consideration of a sequence of trials. The first few trials by themselves are likely to give a misleading idea of the probability and, crucially, if further trials are then observed the initial concept will probably not be revised sufficiently to take account of the later observations (the 'anchoring effect'; Nisbett and Ross, 1980). How can we retain the concreteness of the single trials, present a powerful representation of the higher-order concept, and tie these two together?

Our approach to overcoming the LSN trap has been to provide in the Sampling Playground the two representations of sets of samples shown in Figures 3 and 5, and to give the user great control over the sampling process. Using the dot plot representation of Figure 3, and the Sampling Rate slider set to the left, taking a single sample can be observed as a multi-step process, emphasising concretely the individual values obtained and the variation or 'clumpiness' within this one sample.

Further single samples can be taken, then the rate can be increased further and a sequence of hundreds of samples taken quickly, with the dot plot or cumulating distributions (Figure 5) representations

developing as you watch. We thus aim to offer representations of sets of many samples that are (i) easily understood in terms of the single samples they are made of, and yet (ii) exhibit the higher-order features—such as great sample-to-sample variability—that learners need to grasp.

3.2 Confidence Intervals

A series of samples is sometimes shown in statistics textbooks and teaching software (e.g. Models'N'Data, Stirling 1993), with the confidence interval for each sample indicated, and the extent of variation from sample to sample emphasised. This is a valuable approach but we are concerned that it is insufficiently concrete, and insufficiently tied to the sample values themselves and to the sampling process. Figure 4 shows our representation of confidence intervals. Our hope is that by presenting confidence intervals in the Sampling Playground they can be seen as concrete in relation to any single sample, yet also as forming a probabilistic pattern over a long series of samples. In other words, single-concrete-to-overall-patterning understanding of confidence intervals should follow corresponding development of understanding of sampling, as discussed above.

3.3 Hypothesis Testing

One key difficulty is to realise that a hypothesis is precisely that—the user's conjecture, and all tail areas or significance levels are probabilities conditional on that hypothesis. In developing our Hypothesis Testing Playground (Figure 6) we are aiming to separate clearly the true state of the world—the distribution from which the sample actually came—from user hypotheses. The user can derive a sampling distribution from any *conjectured* distribution, and use this as the basis for hypothesis testing with the single sample. These are the principles on which further development of this playground is being based.

4. StatPlay in use

A class of second year psychology undergraduates saw a brief lecture demonstration using StatPlay, then in a tutorial used it in pairs, with brief guidance from worksheets. The exercise was arranged as a single-group, pretest-posttest quasi-experiment (Cook and Campbell, 1979), with pencil-and-paper assessment of the accuracy of several student intuitions.

One goal was for students to improve their conceptions of sampling variability, the standard error, and how these vary with *N*. A sensitivity index, where 1.0 means no sensitivity to *N* and 1.78 means appropriate sensitivity, increased from 1.07 before the lecture to 1.61 after the tutorial, giving initial encouragement that our approach is fruitful.

5. References

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In C. Keron and C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*, Hillsdale, N.J.: Erlbaum, pp. 311-339.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, Vol. 248, pp. 114-122.

Nisbett, R. and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning*, New York: Macmillan, pp. 465-494.

Shute, V. and Gawlick-Grendell, L. (1993). An experiential approach to teaching and learning probability: *Stat Lady*. In P. Brna, S. Ohlsson and H. Pain (Eds.), *Artificial intelligence in education, Proceedings of AI-ED 93, World Conference on Artificial Intelligence in Education,* Edinburgh, August, 1993. Charlottesville, VA: AACE, pp. 177-184.

Stirling, D. (1993) *Models'N'Data.* (Software) New Zealand: Massey University.

Thomason, N. R., Cumming, G. and Zangari, M. (1994). Understanding central concepts of statistics and experimental design in the social sciences. In K Beattie, C. McNaught ans S. Wills (Eds.), *Interactive multimedia in university education: Designing for change in teaching & learning* Amsterdam: Elsevier, pp. 59-81.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, Vol. 92, pp. 105-110.

West, L. H. T. and Pines, A. L. (1985). *Cognitive structure and conceptual change*. Orlando, FL: Academic Press.

White, B. Y. (1993). ThinkerTools: Causal models conceptual change and science education. *Cognition and Instruction*, Vol.10, No. 1.

6. Acknowledgement

The development of StatPlay is supported by the Australian Committee for the Advancement of University Teaching (CAUT).

Figure 1. The Data Playground, showing four views of the same dataset: list of values, ordered list, frequency histogram and dot plot. The percentile cursor is positioned at 26.1%, with values below the cursor highlighted in all four views. The triangles and arrows show mean, median and s.d.

Figure 2. A normal distribution in the Continuous Distribution Playground: *z*-scores are indicated by vertical lines and symmetric tails are shaded. Probabilities, *z*, and X values are shown for the tails.

Figure 3. The Sampling Playground. Some of a series of samples from the upper population distribution are shown, as dot plots, one sample per line in the lower scrolling window. Sample means (inverted triangles) vary widely because the sample size, $N = 4$, is so small.

Figure 4. The Sampling Playground. The series of samples shown in Figure 3, now with the 95% Confidence Interval for the population mean (vertical line) shown as a shaded bar for each sample. There are button click options to show confidence intervals based on *z* (known population s.d.) or *t* (unknown population s.d.), and for various levels of confidence from 80% to 99%.

Figure 5. The Sampling Playground. A series of 100 samples of size $N = 20$ was taken. The dark distribution is the cumulation of all 2000 data points; the light distribution is the sampling distribution of the 100 sample means, approximately normal as expected by the Central Limit Theorem.

Figure 6. The Hypothesis Testing Playground, currently at an early stage of development. The true population may be hidden or shown in the upper panel; the sampling distribution under a hypothesised distribution appears in the middle, and the sample is shown as a dot plot below.