How Significant Is A Boxplot Outlier?

Robert Dawson
Saint Mary's University

**Key Words**: Boxplot; Outlier; Significance; Spreadsheet; Simulation.

Abstract

It is common to consider Tukey's schematic ("full") boxplot as an informal test for the existence of outliers. While the procedure is useful, it should be used with caution, as at least 30% of samples from a normally-distributed population of any size will be flagged as containing an outlier, while for small samples (N<10) even extreme outliers indicate little. This fact is most easily seen using a simulation, which ideally students should perform for themselves. The majority of students who learn about boxplots are not familiar with the tools (such as R) that upper-level students might use for such a simulation. This article shows how, with appropriate guidance, a class can use a spreadsheet such as Excel to explore this issue.

1. Introduction

 Exploratory data analysis suddenly got a lot cooler on February 4, 2009, when a boxplot was featured in Randall Munroe's popular Webcomic **xkcd**.

Figure 1
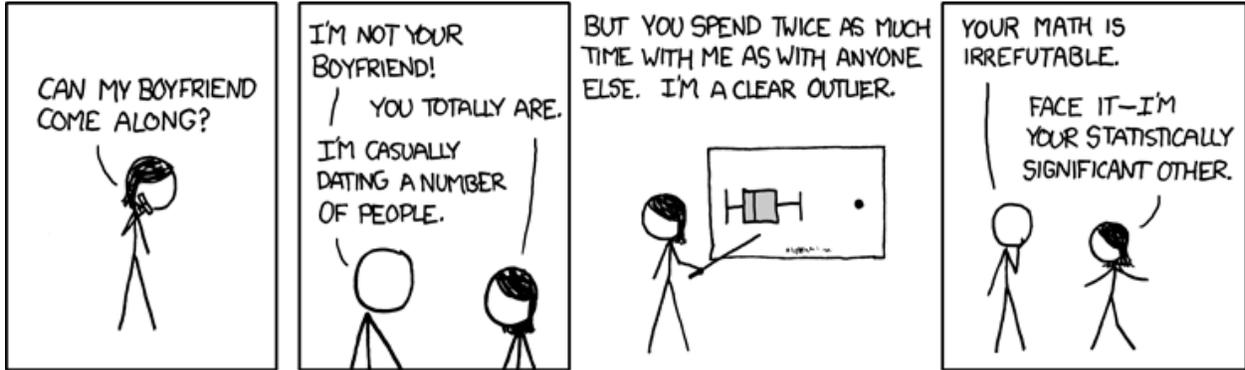
"**Boyfriend**" (Feb. 4 2009, http://xkcd.com/539/)   Used by permission of Randall Munroe

But is Megan justified in claiming "statistical significance"? We shall explore this intriguing question using Microsoft Excel.

The *boxplot*, introduced by Tukey (1977) should need no introduction among this readership. Tukey originally introduced two variants, the *skeletal* boxplot which contains exactly the same information as the "five number summary" and the *schematic* boxplot that may also flag some data as *outliers* based on a simple calculation. Other variants such as the "notched boxplot" (McGill, Tukey, and Larson 1978) were introduced later; and generalizations such as the "violin plot" (Hintze and Nelson 1998) and "bagplot" (Rousseeuw, Ruts, and Tukey 1999) have also been suggested.

The skeletal boxplot consists of a *box* extending from the first quartile (Q1) to the third quartile (Q3); a mark at the median; and *whiskers* extending from the first quartile to the minimum, and from the third quartile to the maximum. Typically, the whiskers will be significantly longer than the box; short whiskers may indicate a distribution like the uniform distribution, with sharp cutoffs. The skeletal boxplot is widely used in data analysis sections of grade school math courses, and both the NCTM Standards (National Council of Teachers of Mathematics 2000) and the more recent Common Core State Standards Initiative (Common Core State Standards Initiative 2010, standard 6.SP.4) mandates its introduction in the sixth grade. While Bakker, Biehler, and Konold (2004) have argued that boxplots may be too conceptually advanced for younger students, they are a widely accepted tool at the high school and undergraduate level.

The schematic boxplot (more often called a "full boxplot" today) divides the data based on four invisible boundaries, namely, two *inner fences* and two *outer fences*. As usual, we define the *interquartile range* (IQR) to be Q3-Q1. The inner fences are $Q1 - 1.5$ IQR and $Q3 + 1.5$ IQR, while the outer fences are $Q1 - 3$ IQR and $Q3 + 3$ IQR. The whiskers extend to the most extreme data within the inner fences. Data outside the inner fences but inside the outer fences (like Megan) are considered mild outliers and each is marked with a symbol; they are also sometimes labeled. Data outside the outer fences are considered to be extreme outliers and are marked with a different symbol. There is no general agreement on the choice of symbols.

For a large sample from a normally-distributed population the quartiles should be close to $\mu \pm 0.67\sigma$. Thus, the inner fences should be close to $\mu \pm 2.67\sigma$ and the outer fences close to $\mu \pm 4.67\sigma$, respectively. We conclude that, if the population is normally distributed, only about 0.8% of the data will be found beyond the inner fences, and about three in a million beyond the outer fences. Outliers may be evidence of a contaminated data set; they may be evidence that a population has a non-normal distribution; or, they may appear in a sample from a normally-distributed population.

It is tempting to interpret a boxplot outlier as evidence that the population is non-normal or that the sample is contaminated, and students often do so. Effectively, they treat the boxplot as an informal hypothesis test for normality, or (as Megan appears to be doing) population homogeneity. However, this can be very misleading! The probability of 0.8% applies to an individual datum; the probability that *some* datum in the sample is flagged as an outlier may be very much higher. Indeed, for N=1000, we would expect to see around eight flagged data.

For small sample sizes, we might expect fewer extreme data. However, quartiles, like most other statistics, vary more for small sample sizes. If we take only a small sample, we are more likely to get an IQR that is unrepresentatively small, resulting in unusually narrow fences. When this happens, it is more probable that data will be flagged as outliers. Except for small N and rather trivial distributions, computing exact probabilities is impossible.

2. Studying Significance with a Simulation

Over time, a working statistician who uses boxplots will develop an understanding of how they should be interpreted in the context of the type of data with which [s]he customarily works. Our concern here is developing the beginnings of such an understanding in undergraduate students taking a first, perhaps terminal, course in statistics. Students should have the opportunity to learn both the power and the limitations of the boxplot.

Of course, this can be illustrated with a few well-chosen slides in a conventional lecture. However, my classroom experience has been that such a lesson is not well retained, and that students continue to take a very literal view of boxplot outliers as evidence either that the distribution is non-normal or that the flagged datum is somehow "wrong". This paper suggests an alternative approach based on experiential learning, via a simulation. There is a large literature relating to the use of computer simulations in statistics teaching. An extensive bibliography and review up to about the year 2000 appears in (Mills 2002).

One limitation of simulations, noted at least as far back as the work of delMas, Garfield, and Chance (1999), is that merely watching or playing with a preprogrammed simulation may lead to passive learning and poor understanding. They describe a classroom study that shows that requiring the students to predict the outcome of the simulation and then test the predictions significantly increases the effectiveness of the exercise. Lane and Peres (2006) discuss the effectiveness of the simulations in the Rice Virtual Lab in Statistics (RVLS). Some of these simulations, designed to train the student to a more sophisticated approach to interpreting graphical displays of data, are particularly relevant here. Examples include the simulations on

"Histograms, Bin Width, and Cross-Validation" and "Regression by Eye." Lane and Peres reach conclusions similar to those of delMas, Garfield, and Chance both on the risk of passive learning and on the value of having the students form an initial hypothesis – right or wrong – that will be supported or modified on the basis of the simulation.

Another way to engage students is to get them to take an active part in constructing the simulation. In many cases, this participation must be carefully directed in order to be effective; as Mayer (2004) points out, neither the increased behavioral activity nor the increased student autonomy of a hands-on exercise is a sufficient condition for effective learning. Fay and Mayer (1994) give a particularly relevant example of students learning LOGO programming, where a guided discovery approach proved more effective than a pure discovery approach.

The professional statistician who wants to create a simulation will probably use a statistical programming language such as R; and statistics majors should probably be encouraged to do the same. However, lower-level courses for non-majors are unlikely to use R (except perhaps with a "wrapper" such as pmg - see Verzani 2008). There are several reasons for this. The learning curve for R is steep for a one-semester terminal course at the first year level; and there may well be pressure from client departments to teach a package such as SPSS "because that's what they will be using later."

Dedicated "statsheet" packages such as SPSS or MINITAB enforce a "subjects × observations" structure. This discourages beginners from doing meaningless calculations, and clarifies syntax, but it makes simulations unwieldy. In a simulation, an instance of a statistical process is a "single replication" and should, to be consistent with the "statsheet" philosophy, be carried out in a single row; but this is problematic except for the very simplest simulations that can be carried out using commands such as MINITAB's RMEAN.

What options do we have for software that is readily available, easily usable for simulations, and not too intimidating for the intro stats course? There are packages such as *Fathom* that support simulations - see, for instance, (Tabor 2010) for an example of *Fathom* in use - but their cost may be prohibitive for an individual course at a school, or at a university that has already chosen another standard statistics package. A natural answer is Microsoft Excel. Most students already have some familiarity with Excel, and it is already loaded on most university network servers and most students' laptops. And, as will be shown in the next section, Excel does have the power to simulate a multistage statistical process such as the construction of a boxplot.

The author is aware that Excel is widely disparaged - see, *e.g.,* (Nash and Quon 1996) or (Cryer 2001) - as a tool for data analysis, and indeed supports this view. Over several versions, Excel has used incorrect algorithms for many statistical calculations, and it encourages impressionable neophytes to produce graphics with a very high level of what Tufte (1983) has labeled "chart junk." As a final irony, it does not even support the boxplot as a built-in plot type! However, it is much better suited for many simulation purposes. It is more flexible than "statsheet" software such as MINITAB, and easier to use than most programming languages.

The Appendix gives a simulation exercise using Microsoft Excel 2007 that might be handed out to a class. The student starts out with a blank spreadsheet, and creates the first simulation a step

at a time. Some of the formulae are given; others must be created by the student by modifying the given ones.

Of course, the completed spreadsheet (available as BoxplotSim.xls from the JSE website) could be put online, so that students could perform the simulation simply by pressing F9 to recalculate the entire spreadsheet (generating in the process fresh random numbers.) However, this "black box" approach would not leave the students with much understanding of what the simulation did. Instead, the student is required to focus on one component of the simulation at a time, learning the necessary Excel functions and how to use them. After this, he or she is expected to modify the spreadsheet, constructing new spreadsheet simulations to answer related questions.

3. Results and Interpretation

In this section, typical results of some simulations of this type are given. Simulations were performed for sample sizes of 5, 9, 13, 17, and 21, for both Uniform [0,1] and normal distributions, with 10,000 repetitions. Sizes were chosen to be of the form 4k+1 (where k is a positive integer) in order to have simply-defined quartiles. For the uniform distribution on the unit interval, the theoretical quartiles are at 0.25 and 0.75 and the corresponding inner fences are at -0.5 and 1.5, outside the range of the distribution. Thus any datum flagged as an outlier is so flagged because of inaccurate quartile estimates, not on its own merits. The results are presented in the following table.

Table 1: Samples (out of 10,000) that have flagged mild outliers

Size	5	9	13	17	21
Uniform[0,1]	2609	1510	855	526	282
Normal	3408	3081	3004	3061	3165

For the uniform distribution, very small samples yield outliers rather often; but as the sample size increases the incidence drops off rapidly. However, for the normal distribution, the incidence of samples with flagged outliers never drops below about 30%, and above sample size 13 starts to rise. It is interesting that in the sample size range examined the incidence varies rather little. (The counts above may be treated as binomial random variables, with standard deviation bounded above by $\sqrt{N}/2$; so both the drop and subsequent rise are statistically significant.) A second simulation was done with 1000 repetitions and a sample size of 105. For the uniform distribution the fences approximated the theoretical out-of-bounds values; no outliers whatsoever were observed. In contrast, for the normal distribution, a whopping 557 data sets – more than half - were flagged as having at least one outlier.

If we consider the boxplot as an informal test of the null hypothesis that the sample is from a normally-distributed [and uncontaminated] population, a flagged outlier in a properly-distributed sample corresponds to a "Type I error". But, as we have just seen, whatever the sample size, the α of such a test is never less than 30%. Thus, the existence of a datum flagged as a mild outlier can never be taken, on its own, as significant evidence against the purity of the sample or the normality of the population.

The first simulation was repeated for extreme outliers (replace "2.5" by "4" and "-1.5" by "-3" in the formula). Recall that for a large sample from a normally-distributed population, only about three data in a million will be beyond the outer fences; the small samples examined here tell a very different story!

Table 2: Samples (out of 10,000) that have flagged extreme outliers

Size	5	9	13	17	21
Uniform[0,1]	1055	279	86	26	7
Normal	1295	727	440	269	227

With a sample size less than 10, an extreme outlier does not allow us to reject the hypothesis of normality at a significance level of 5%. For medium-sized samples, however, the p-value becomes smaller. A final sequence of simulations (with 1000 repetitions) extends this to larger sample sizes.

Table 3: Samples (out of 1000) that have flagged extreme outliers

Size	21	41	61	81	101
Normal	28	10	3	1	2

The number of extreme outliers continues to drop with sample size throughout the range of sizes that we can easily simulate. However, as the individual sample gets large, the outer fences will approach their theoretical values of about $\mu \pm 4.67\,\sigma$. The proportion of a normally-distributed population that lies outside this range is very small but nonzero (about three in a million), so for large values of N we expect the number of extreme outliers to be roughly Poisson ($\lambda=3\times10^{-6}N$). For values of N below about 10^4, the probability that this variable will take any value but 0 is very low; but with $N=10^6$, for instance, the probability of at least one extreme outlier would be about $1-e^{-3}$, or about 95%.

4. Conclusions

A standard spreadsheet, such as Microsoft Excel, can be used in simulations to study the interpretation of boxplots. Such a simulation is simple enough that it can be carried out (with guidance) by students in lower-level undergraduate classes who might have difficulty using a language such as R for the same purpose.

These simulations demonstrate that, regardless of size, at least 30% of samples drawn from a normally-distributed population will have one or more data flagged as outliers. It is generally incorrect to assume that a single flagged datum implies an exotic population distribution or contaminated data, although such a conclusion can be drawn from extreme outliers if the sample size is not too large or too small.

For mild outliers, statistical significance is only achieved for short-tailed distributions such as the uniform distribution; and even then the sample size cannot be too small. We deduce that if Megan is correct, her boyfriend must have quite a social life!

Appendix
An Excel Simulation Exercise to Study Outliers

JSE Editor's Note: An alert reader found a couple errors in the Appendix. The Appendix was updated October 31, 2012. This is the updated version.

(Hints for the teacher are given in italics. A version of the exercise without these hints is given on the JSE web page at www.amstat.org/publications/jse/v19n2/Appendix.docx. Questions in **boldface** *are to be answered by the student. You may modify the exercise freely for the needs of your class. You are encouraged to open Excel and follow along while reading. The exercise starts with a blank spreadsheet, as displayed when Excel opens or when the "New" item is selected on the menu.*

The artist, Randall Munroe, has given blanket permission for nonprofit use of the cartoon that introduces the main article, provided that the source is given. You may wish to include it in a class handout. However, teachers of younger students should be aware that the **xkcd** *website contains some other material that parents, principals, or school boards might consider unsuitable; please use your judgment.)*

You are familiar with boxplots, and should be developing the ability to look at a boxplot and get a feel for the data set involved. In particular, you should (like Megan, in the cartoon) be aware that data flagged as outliers may differ in some important way from the rest of the population. But is every datum that is flagged as a boxplot outlier really an "outsider", or do well-behaved samples have outliers too? Before we start, please answer the following question in the space provided. Try to give the best answer you can, but remember that we haven't done the simulation yet! You'll be allowed to change your mind later if you want.

If a boxplot shows an outlier, can we be pretty sure that that datum does not belong, or that the distribution is not normal? Explain your answer.

(This exercise uses a predict-test-check model; this has been found (see, e.g., *[Fay and Mayer (1994)](#) or [Lane and Peres (2006))](#) to be more effective than less-structured exploration-based learning. Please ensure that the students do answer this question before proceeding, rather than answering it out of order once they know the answer.)*

Now we will use Microsoft Excel to do a simulation and check your guess. We will do 1000 replications of the outlier labeling process for N=9, with normally distributed simulated data. We start with a single example.

> (1) Put a function in cell A2 that will generate the appropriate random variable. To start with, we will use =NORMSINV(RAND()) The built-in function RAND() generates a uniformly generated random number from 0 to 1; NORMSINV converts this to a **S**tandard **NORM**ally distributed random variate. The "=" at the beginning tells Excel that the cell contains a formula that it should evaluate.

(2) Click on square A2; it should now have a heavy black border with a little black square (the "Fill Handle") in the bottom right corner.

Figure 1: Using the Fill Handle to create a 9-element sample
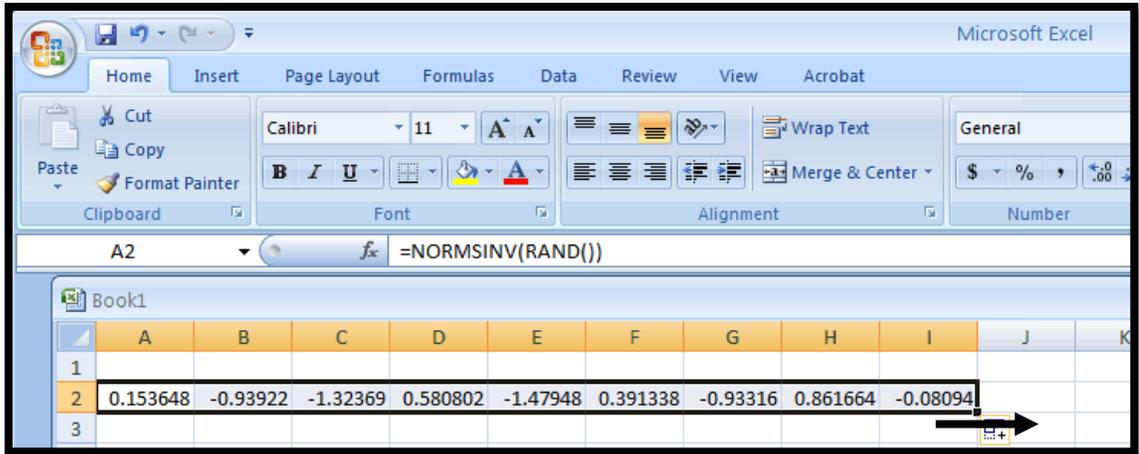
Drag the Fill Handle across to square I2. The 1×9 rectangle should fill with normally-distributed random numbers, with mean 0 and standard deviation 1. Click on each cell to see the formula that it contains; they should be all the same, but each one generates a different random number.

(There are other ways of filling a region of the spreadsheet with identical or analogous formulae, but this is probably the easiest . Students can use whatever method they prefer; the results should be the same.)

(3) In cell K2 put the formula   =MIN(A2:I2). You may want to use the Formula Bar at the top to enter or edit it. Beware of typos; like most programs, Excel is fiercely literal-minded. If you have got it right, the minimum of the nine simulated data in the first row should appear in cell K2. Use cell K1 to label this column "Minimum".

(A pair of cell coordinates separated by a colon represents a "range", a linear or rectangular block of cells. Excel has many functions, like MIN()*, that act on ranges. We will see some others below.)*

(4) In cell L2 put   =QUARTILE(A2:I2,1), which will evaluate the first quartile (with nine data, this is the third smallest) of the first row of simulated data. Label the column (if you use the abbreviation Q1 be sure not to put an equals sign before it!)

(5) Create a formula for the third quartile and put it into cell M2. Test it.

My third quartile formula:

(6) Create a formula for the maximum and put it into cell N2. Test it.

My maximum formula:

(7) Skip a column and put the formula =M2-L2 into cell P2. This should compute the interquartile range; check that it does.

(8) Skip a column and put the formula =L2-1.5*P2 into cell R2. This should compute the lower fence. Create a formula that computes the upper fence and put it into cell S2.

My upper fence formula:

(9) Skip a column and put the formula =K2 < R2 into cell U2. This is a *Boolean* expression, which takes values "TRUE" and "FALSE". It should be TRUE if and only if there is a lower outlier. **Why does this formula work?**

(10) Put a formula into cell V2 that determines whether there is an upper outlier. Test and make sure your formula works. (Pressing F9 will recalculate all formulae, generating fresh random numbers: you can use this to test the formula both in cases where there is a lower outlier and where there is not.)

My upper outlier detection formula:

(11) Skip a column and put the formula =IF(OR(U2,V2),1,0) into cell W2. OR(U2,V2) is Boolean and has the value TRUE if either of U2 and V2 is TRUE, and FALSE otherwise. IF takes three inputs, the first of which must be Boolean. If the Boolean is TRUE it returns the second value, otherwise the third. Thus W2 displays _____if the simulated data have at least one outlier, and _____ otherwise (**fill in the blanks**!)

(12) You could use the F9 key to repeat your simulation several times, but there is an easier way, which lets you tally the results automatically. Drag the mouse along the second row from A2 to W2; you should see a heavy outline around the selected range, with a black Fill Handle at the bottom right corner. Drag this downward to the corner of cell W1001. You should get a thousand replications of the simulation.

Figure 2: Using the Fill Handle to replicate the simulation

(13) By clicking on different cells, you can see the formulae that they contain. Note that K2 contains the minimum of the range (A2:I2), K3 contains the minimum of the range (A3:I3), and so on.

(This is standard spreadsheet user interface behavior: by default, addresses are relative to the cell being copied. You might say that if cell B1 contains the formula =A1+A2, Excel thinks of this as "Add the cell above me to the one on its right." If you copy it to E10, the formula becomes =D10+D11, and so on – still adding the cell above to the one on its right. To avoid this (say you wanted all your formulae to reference one constant in cell A1) use a dollar sign before one or both coordinates in the formula. This will get values in just the same way, but it will copy differently. If you copy =$A$1+A2 from cell B2 to cell E10, it will become =$A$1 + D11, "add A1 to the cell above and to my right." Copying = A$1+A2 from cell B2 to cell E10, it will become =D$1 + D11. Your students will not need to know this for this lab, but it may be useful if they do other simulations, including the challenge exercise at the end of this appendix.)

(14) Now put the formula =SUM(W2:W1001) into cell X2. Use X1 to label it "Number." It should display the total number, out of 1000 simulations, in which at least one outlier was found.

(If the student has done everything correctly, the number should be around 300 out of 1000. The completed spreadsheet, [BoxplotSim.xls](), appears on the JSE website.)

For about what proportion of samples of size 9, drawn from a normally distributed population, did the boxplot process indicate an outlier? Does this change the opinion you held before the experiment? Why?

Press F9 to recalculate everything. Is your estimate of the proportion the same this time? If not, how different is it?

*(If the class has already studied binomial confidence intervals, the following question may be substituted: **The outcome of each replication is an independent random variable taking the values 1 (if there is an outlier) or 0 (if there is not). Putting 1000 of these together we have a binomial experiment with N=1000 and P = (proportion of samples with an indicated outlier.) Construct a 95% confidence interval for P.**)*

"Statistical significance" is traditionally associated with observations that would have a probability of 5% or less under the null hypothesis. Comment on Megan's conclusion, if the null hypothesis is that, with regard to time spent on dates, she is part of a normally distributed population.

If you look carefully at Megan's chart, it is possible that the dot representing herself is an extreme (>3IQR) outlier. Redo your simulation to study the frequency of these; describe your findings. What formulae do you need to change?

Repeat the simulation for other data set sizes (good ones are 5, 13, 17, 21, because the quartiles are simply defined). Describe what you observe. Under what conditions, if any, would Megan be correct?

(Stronger students could be asked, as a challenge, to create a simulation to determine how frequent multiple outliers are. To do this, they would, after finding the fences, create a second 9x1000 workspace, with each cell containing 1 if the corresponding random number was an outlier, and 0 otherwise. Partial absolute addressing (see note after (13)) may be useful here, to allow entries in all nine columns to reference the same fence values. For instance, the formula in AA2 might be =IF(OR((A2<$R2),(A2>$S2)),1,0). Copying this into cell AB2 gives the formula =IF(OR((B2<$R2),(B2>$S2)),1,0), while copying into cell AA3 gives the formula =IF(OR((A3<$R3),(A3>$S3)),1,0). The random numbers change within the row, the fence values only change from row to row.)

Acknowledgement

References

Bakker, A., Biehler, R. & Konold, C. (2004), "Should young students learn about box plots?" *Curricular Development in Statistics Education, Sweden* 263-173.

Cryer, J. D. (2001), "Problems Using Excel for Statistics" [online], available at www.cs.uiowa.edu/~jcryer/JSMTalk2001.pdf

delMas, R. C., Garfield, J., and Chance, B. L. (1999), "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning", *Journal of Statistics Education,* 7(3) (1999)   [online], available at http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

Fay, A. L. and Mayer, R. E. (1994), "Benefits of Teaching Design Skills Before Teaching LOGO Computer Programming: Evidence for Syntax-Independent Learning", *Journal of Educational Computing Research* 11 287-210.

Hintze, J.L. and Nelson, R. D. (1998), "Violin Plots: A Box Plot-Density Trace Synergism", *The American Statistician* 52(2) 281-84.

Lane, D. M. and Peres, S. C. (2006), "Interactive Simulations in the Teaching of Statistics: Promise and Pitfalls", in ICOTS-7, available online at http://www.ruf.rice.edu/~lane/papers/interactive-simulations.pdf

Mayer, R. E. (2004), "Should There Be a Three-Strikes Rule Against Pure Discovery Learning?", *American Psychologist* 59(1) 14-19 .

McGill, R., Tukey, J. W., and Larsen, W. A. (1978), "Variations of Box Plots". The American Statistician, 32(1) 22-16.

Mills, J. D (2002), "Using Computer Simulation Methods to Teach Statistics: A Review of the Literature", *Journal of Statistics Education* 10(1) [online], available at http://www.amstat.org/publications/jse/v10n1/mills.html

Nash, J. C and Quon, T. K. (1996), *"Issues in Teaching Statistical Thinking with Spreadsheets", Journal of Statistics Education* 4(1) [online], available at http://www.amstat.org/publications/jse/v4n1/nash.html

National Council of Teachers of Mathematics (2000), *Principles and Standards for School Mathematics,* Reston, VA.

National Governors Association Center for Best Practices and Council of Chief State School Officers (2010), *Common Core State Standards for Mathematics*, available at http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot", *The American Statistician*. 53 (4), 382–387.

Tabor, J. (2010), "Investigating the Investigative Task: Tests for Skewness", *Journal of Statistics Education* 18(2) [online], available at http://www.amstat.org/publications/jse/v18n2/tabor.pdf

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley.

Verzani, J. (2008), Using R in Introductory Statistics Courses with the pmg Graphical User Interface, *Journal of Statistics Education* 16(1) [online], available at http://www.amstat.org/publications/jse/v16n1/verzani.pdf

Robert Dawson
Saint Mary's University
933 Robie Street
Halifax, Nova Scotia,
Canada   B3L 3C3
rdawson@cs.stmarys.ca
(902) 420-5796