Analyzing class performance on tests: gray swans

David M. Harrison*
Department of Physics, University of Toronto
Toronto, ON M5S 1A7, Canada

I. INTRODUCTION

When we give a test or other assessment to a class, the distribution of scores can have a variety of shapes: flat, double-peaked, ramped, and more. But because of regression to the mean, the most common distribution is bell-shaped or approximately Gaussian. This type of distribution is so common that the statisticians call it *normal*. Normal distributions have been extensively analyzed.

A favorite example of the ubiquitous nature of the normal distribution is a *quincunx*, also known as a *Galton board* or a *bean board*. Figure 1 shows a quincunx.[1] Devised by Galton in about 1860, it consists of a number of balls dropped one at a time onto a peg located so that each ball bounces to the left or to the right with equal probability. Below that peg are two pegs, each positioned so that any ball that collides with it also has equal probability of bouncing to the left or to the right. This continues for a number of layers, and then the balls are collected in bins at the bottom. As can be seen, as the number of balls becomes large, the distribution of balls in the bins goes to a Gaussian.

Figure 1. A Quincunx

* email: david.harrison@utoronto.ca

When we wish to characterize the overall performance of the class on the test, we commonly calculate the mean and the standard deviation of the scores. However, these measures assume that the underlying probability distribution function (pdf) is a true Gaussian:

$$\mathrm{pdf(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \tag{1}$$

The maximum amplitude of the distribution is such that the area under the curve from $x = -\infty$ to $x = +\infty$ is exactly 1. But on a test where the score is out of 100% and the distribution is bell-shaped, the actual pdf is closer to:

$$\mathrm{pdf(x)} = \begin{pmatrix} A e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} & 0 \le x \le 100 \\ 0 & \text{otherwise} \end{pmatrix} \tag{2}$$

where the amplitude so the total area under the curve is 1 is:

$$A = \frac{1}{\sqrt{\frac{\pi}{2}}\sigma\left[\mathrm{erf}\left(\frac{100-\bar{x}}{2\sqrt{\sigma}}\right) - \mathrm{erf}\left(\frac{\bar{x}}{2\sqrt{\sigma}}\right)\right]} \tag{3}$$

and erf is the error function:

$$\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{4}$$

For gaining a rough idea of how a class performed on a test, the difference between Eqns. 1 and 2 can usually be ignored. But to compare the size of the performance difference of two or more groups of students, such as by gender or some other factor, we will show that the difference between Eqns. 1 and 2 can be significant. Such comparisons are central to much science education research, but are also useful for classroom teachers; a recent example is a comparison of evaluation results over a 13 year span for a teacher who changed her pedagogy during that period.[2]

If we take Eqn. 2 to be a model of student performance, than we wish to determine the value of $\bar{x}$, σ, and the uncertainty in $\bar{x}$.

Here we explore the differences between these 2 pdf's. Nassim Nicholas Taleb has written passionately about how assuming a normal distribution often leads to catastrophic mistakes because that distribution under-estimates the number and impact of outliers.[3] He calls these outliers *black swans*. The effects we shall discuss are much less dramatic, and we call their causes *gray swans*. Although the discussion doesn't introduce anything

that is very new, it is apparent from reading journals devoted to science education research that the issues that are discussed are not as well known as they should be.

II. NON-INFINITE CLASS SIZES

For data whose pdf is a true Gaussian, Eqn. 1, any finite number of data points N is just a sample and the mean and standard deviation can only be estimated.

$$\bar{x}_{\text{est}} = \frac{\sum_{i=1}^{N} x_i}{N}$$
$$\sigma_{\text{est}} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \bar{x}_{est}\right)^2}{N-1}} \qquad (5)$$

The statistical uncertainty in the value of $\bar{x}_{est}$ is estimated by the standard "error" of the mean $\Delta\bar{x}_{\text{est}} = \sigma_m = \sigma_{\text{est}} / \sqrt{N}$. The uncertainty in the value of estimated standard deviation $\Delta\sigma_{\text{est}} = \sigma_{\text{est}} / \sqrt{2N-2}$.

For a test whose pdf is given by Eqn. 2 with $\bar{x} = 75$ and $\sigma = 15$, we imagine a multiple-choice format consisting of 20 questions, each worth 5 points with no partial grades given. Then a Monte Carlo (also known as a *random variate*) procedure for a class of 100 students gave the grade distribution shown in Figure 2.

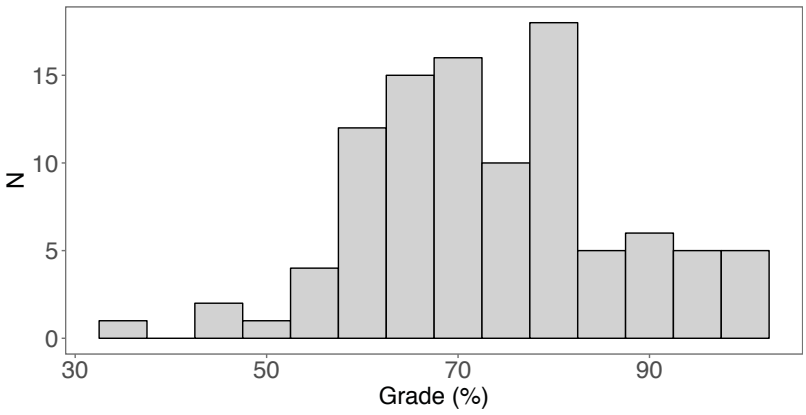


Figure 2. A Monte Carlo generated grade distribution.

Blindly applying Eqn. 5 to this distribution gives $\bar{x}_{\text{est}} = 73.4$ and $\sigma_{\text{est}} = 13.3$, so $\bar{x}_{\text{est}} = 73.4 \pm 1.3$. The median of the distribution is m = 70.0. The uncertainty in the median is discussed in Section IV below.

Mathematically, the expected value of the mean, which we call $\bar{x}_{\text{expected}}$, is given by:

$$\bar{x}_{\text{expected}} = \int_0^{100} x A e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx \qquad (6)$$

The general solution to Eqn. 6 is mathematically very complicated, but for $\bar{x} = 75$ and $\sigma = 15$, $\bar{x}_{\text{expected}} = 73.43$.

We can also perform a least-squares fit of a Gaussian to the distribution of Fig. 2 using a Levenberg-Marquardt algorithm. Taking the uncertainty of the number of students in each bin to be the square root of that number,[4] the result of the fit is $\bar{x}_{\text{fit}} = 73.3 \pm 1.5$, $\sigma_{\text{fit}} = 12.2 \pm 1.3$, with $\chi^2 = 16.2$ for 11 degrees of freedom. The maximum amplitude of the Gaussian is 13.8 ± 2.0. The stated uncertainties are from the diagonal elements of the covariance matrix of the fit. Figure 3 shows the result.

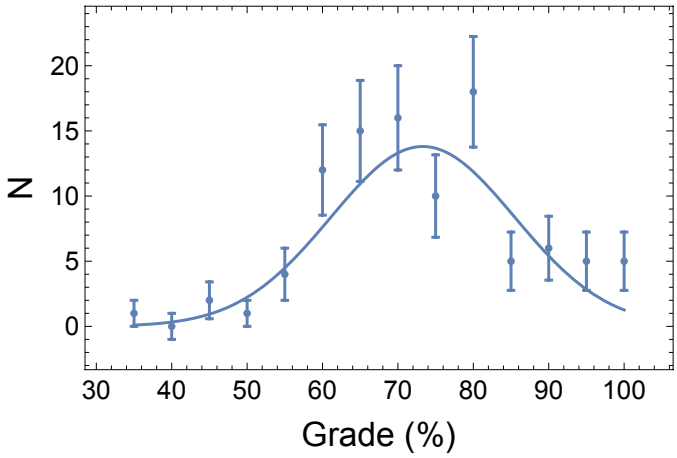


Figure 3. Fitting the distribution of Fig. 1 to a Gaussian.

So all three methods of trying to determine $\bar{x}$, using Eqn. 5 on the data, integrating Eqn. 2, and fitting the data to a Gaussian, gave values lower than the actual value of 75.

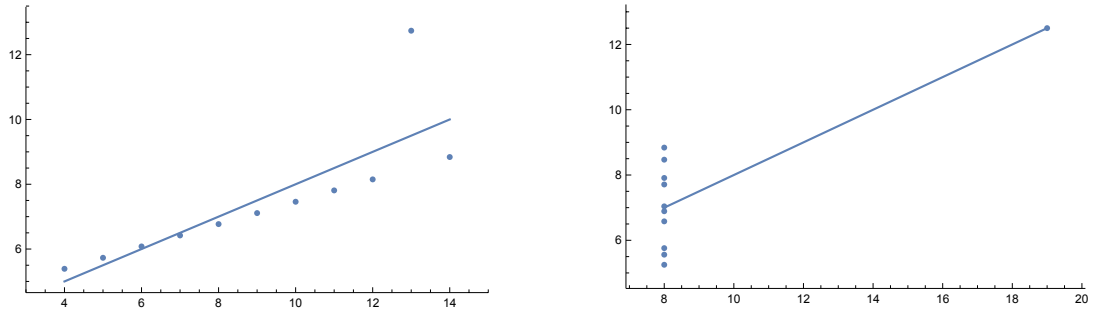
One well-known problem with the mean is that a single outlier datapoint can seize control. For example, if we have a class of 5 students who take the 20-question test we have been describing, the grades could be (75, 80, 75, 30, 70). For this distribution, the mean is 66. The grade of 30 is a black swan. The median is "robust" for the presence of a few black swans; for this data the median is 75.0.

Standard least-square regression algorithms for fitting data to a model have the same problem with black swans and for the same reason. As Emerson and Hoaglin point out:

> Various methods have been developed for fitting a straight line of the form
>
> $$y = ax + b$$
>
> to the data (x_i, y_i), $i = 1,\ldots,n$. The best-known and most widely used method is least-squares regression, which involves algebraically simple calculations, fits

> neatly into the framework of inference built on the Gaussian distribution, and requires only a straightforward derivation. Unfortunately, the least-squares regression line offers no resistance. A wild data point can easily seize control of the fitted line and cause it to give a totally misleading summary of the relationship between y and x.[5]

Figure 4 shows the result of fitting two synthetic datasets to a straight line; the datasets were devised by Anscombe.[6] In both cases a single wild datapoint has seized control of the fit.



(a) (b)

Figure 4. Fitting two datasets, each with an outlier, to a straight line.

The fits shown in Fig. 4 illustrate the importance of visual representations of the data. For both fits the intercept is 3.0 ± 1.1 and the slope is 0.50 ± 0.12. The sum of the squares of the residuals are 13.7562 and 13.7425 respectively and both fits have 9 degrees of freedom. So just looking at the numbers, one could conclude that the datasets are very similar. However, the plots make it obvious that the datasets are quite different.

If Jeff Bezos walks into a bar, the mean wealth of the bar's patrons immediately goes up by several billion dollars, although no non-Bezos drinker is any richer: in terms of the mean wealth Bezos is a black swan.[7] Even without black swans, for datasets with long tails the data in those tails are what we are calling gray swans, which can skew the value of the mean. That is why, for example, in describing the typical income of a sample population the median is preferred. Figure 5 illustrates for weekly income in the United Kingdom for 2009-2010.[8]

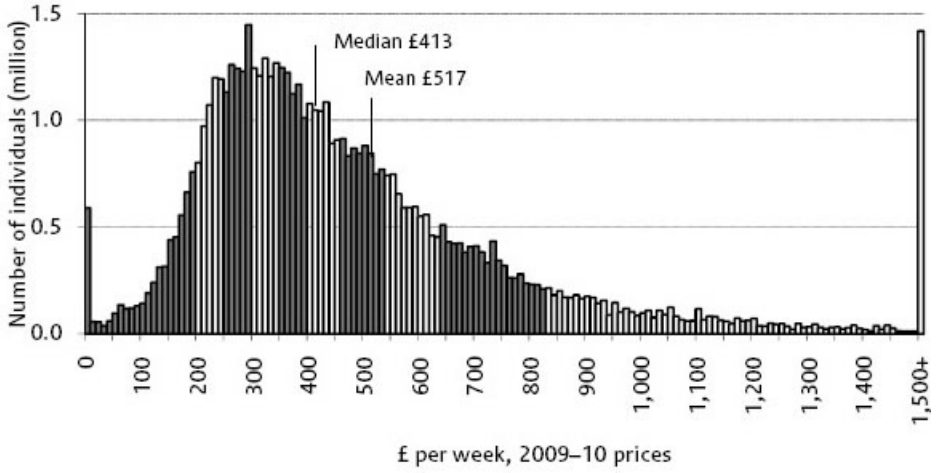


Figure 5. Income per week in the UK in 2009 – 2010.

For both black and gray swans, a common technique is to use a *trimmed mean*, which removes the largest and/or smallest values before calculating the mean. For the test results shown in Fig. 2, a reasonable way for forming such a dataset is to remove the tail for low grades by defining:

$$\delta \equiv 100 - \overline{x}_{\text{est}} \tag{7}$$

and then removing all grades from the dataset less than:

$$\overline{x}_{\text{est}} - \delta = 2\overline{x}_{\text{est}} - 100 \tag{8}$$

For the data shown in Fig. 2, this eliminated the 3 lowest of 100 grades, which raised the calculated mean from 73.4 to $\overline{x}_{\text{trimmed}} = 74.4$. It also raised the median from 70.0 to 75.0, but lowered the calculated standard deviation from 13.3 to $\sigma_{\text{trimmed}} = 12.2$.

III. A COURSE WITH 1000 IDENTICAL SECTIONS

We used a Monte Carlo method to generate 1000 instances of grades on the 20-question multiple choice test already discussed, each with 100 students. This could correspond to a course with 1000 sections of the course with the 100 students in each section being statistically the same, and with the same instructor using the same pedagogy for each section. For each section we calculated $\overline{x}_{\text{est}}$, $\overline{x}_{\text{trimmed}}$, the median *m*, and σ_{est}. Figure 6 shows the histograms of these values. The vertical red lines in Figs 6(a), 6(b), and 6(c) are the value of $\overline{x} = 75$, and the vertical red line in Fig 6(d) is the value of $\sigma = 15$.

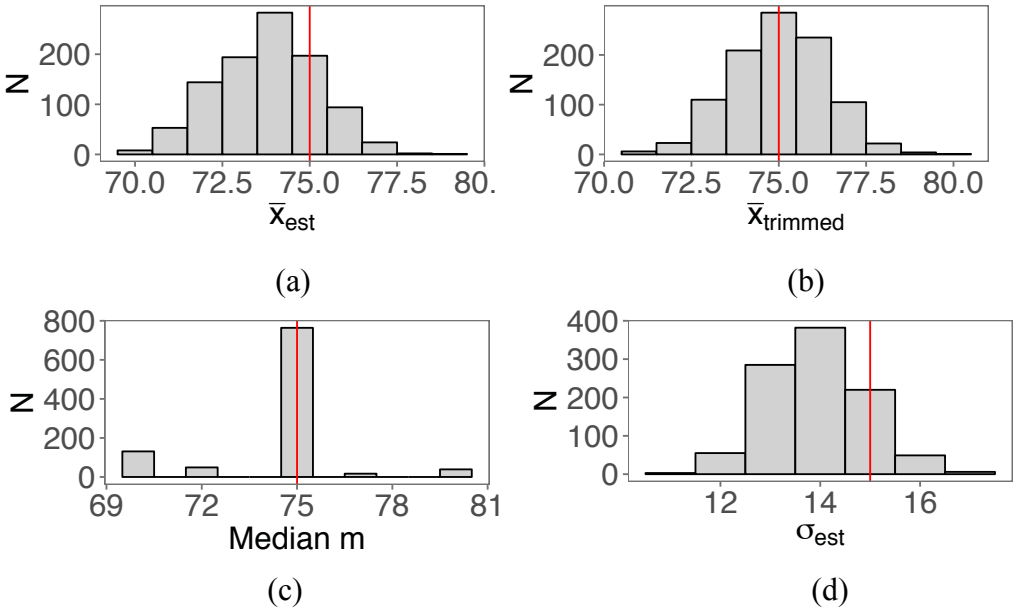


Figure 6. 1000 sections each with 100 students for the 20-question multiple-choice test.
(a) $\overline{x}_{\text{est}}$ (b) $\overline{x}_{\text{trimmed}}$ (c) m (d) σ_{est}.

Note that the median $m = 70 < 75$ for the data of Fig. 2 is due to sampling errors and the gray swans in the tail for lower grades. In general the median is resistant to such swans: almost 80% of the values in Fig. 6(c) are exactly 75.0

The spread of values in Fig. 6 is inversely related to the number of students in each section. Figure 7 shows the result of the same calculation for 1000 sections, but with each having 1000 students. The histogram of the median is not shown since all 1000 values were exactly 75.0.

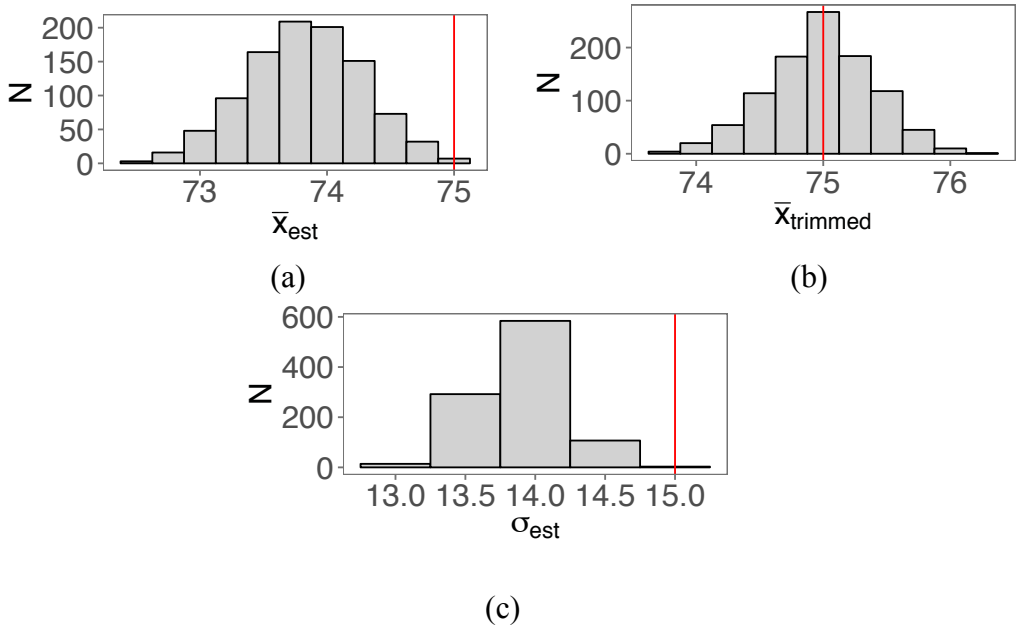


Figure 7 1000 sections each with 1000 students for the 20-question multiple-choice test.
(a) $\overline{x}_{\text{est}}$ (b) $\overline{x}_{\text{trimmed}}$ (c) σ_{est}.

IV. MORE ABOUT THE MEDIAN

We can define the *quartiles* for a distribution so that 25 % of the distribution is less than q_1, the value of the median $m = q_2$, and 75% of the distribution is less than q_3. The interquartile range (*IQR*) is then defined as $q_3 - q_1$.

We are used to characterizing the width of the distribution of grades on a test by specifying the grades that include the middle 68.27% of the class or the middle 95.45% of the class. However, these percentages are only natural for a true Gaussian distribution and correspond to ranges of $\bar{x} \pm \sigma$ and $\bar{x} \pm 2\sigma$ respectively. For any distribution it is at least as natural to characterize the width of the distribution of the grades for the middle 50% of the class, which is the *IQR*.

For a normal distribution with $m = q_2 = \bar{x} = 0$, $q_1 = -.6745\sigma$ and $q_3 = +.6745\sigma$, so $IQR = 1.3490\sigma$. This value for the *IQR* is true for any normal distribution.

Similarly, it can be shown that the area under a Gaussian between $q_1 - 1.5 \times IQR$ and $q_3 + 1.5 \times IQR$ is 99.30%, which corresponds to slightly less than $\pm 3\sigma$.[9] These values are taken to define *cutoffs* for any distribution and values outside of this range are taken to be outliers.[10]

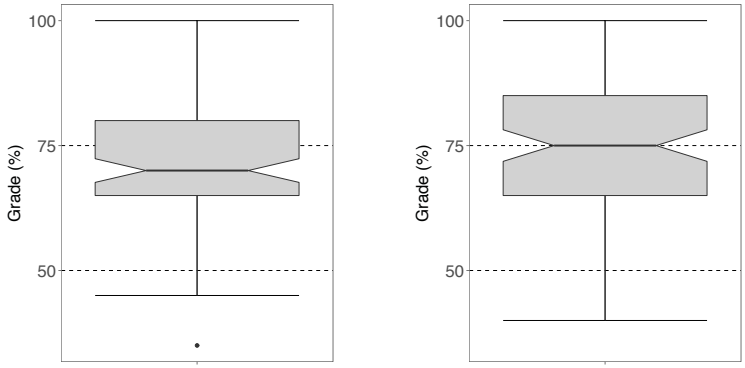
If one wishes to assign a 95% confidence level statistical uncertainty to the value of the median, similar to $2 \times \sigma_m = 2 \times \sigma / \sqrt{N}$ for a normal distribution, a heuristic expression is $1.58 \times IQR / \sqrt{N}$.[11] Thus the value of the median is $m \pm 1.58 \times IQR / \sqrt{N}$.

The value of 1.58 in the above expression is a compromise. Uncertainties are commonly used to decide if the medians for two or more groups are the same or are different. As discussed in Section 7 of Ref. 11, if the widths of the distribution of values in the groups are vastly different, a value of 1.81 would be appropriate. If the widths are roughly equal, 1.81 results in a test that is far too stringent; in this case a value of 1.28 is more appropriate. The value 1.58 is an empirically selected compromise between 1.28 and 1.81, and has been widely adopted. However in specific circumstances some of other value could be more appropriate.

An alternative method of estimating uncertainties is based on Monte Carlo methods, similar to those discussed in Section III above.[12] Since many science education researchers do not know these methods, it is questionable whether they are worth the effort to learn and use. However, they are increasingly used in the experimental physical sciences.

For visualizing the grade distribution using quartiles and medians, the boxplot is very useful. It was invented by John Tukey.[13] Figure 8(a) is the boxplot for the grades shown in Fig. 2. The "waist" on the boxplot is the median, the "shoulder" is the upper quartile, and the "hip" is the lower quartile. The vertical lines extend to the largest/smallest datapoint value less/greater than the cutoffs defined above. The dot is a datapoint outside

the cutoffs and is therefore considered to be an outlier. The "notch" around the median value represents the statistical uncertainty in the value of the median, which as discussed above, is $\pm 1.58 \times \mathrm{IQR} / \sqrt{N}$. For this data, the uncertainty is equal to ± 2.4 . The value the median, then, is 70.0 ± 2.4 which is not within uncertainties of $\bar{x} = 75.0$.



(a) (b)

Figure 8. (a) Boxplot of the test grades shown in Fig. 2. (b) A more typical distribution

Although the median is resistant to swans, it is not immune. As discussed, in the simulated 1000-section course with each section having 100 students, just over 20% of the sections did not have a median grade exactly equal to 75.0. Figure 8(b) shows a more typical grade distribution for another section of this simulated course. The median in this case is 75.0 ± 3.3 .

V. A Real-World Example

So far, we have only used simulated test grades. Figure 9 shows the actual grade distribution for the first term test of an 800-student introductory physics course for life science students for three categories of students. What those categories are is not important for our purposes, but they are for three different measured personality types.[14]

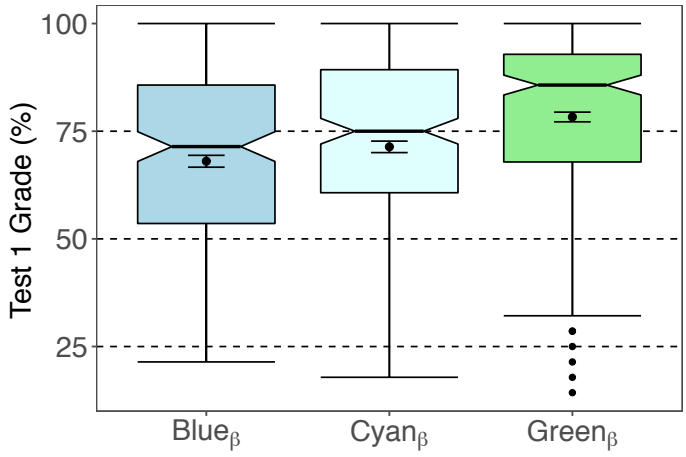


Figure 9. Boxplots of test grades for three categories of students.

Also shown as dots within each "box" are the values of the mean of the grades, and the "error" bars are $\sigma_m = \sigma_{\text{est}} / \sqrt{N}$. Table I shows the values of the median, the mean, and the trimmed mean. When comparing the given uncertainties in the table for the median and the mean, recall that the uncertainty in the median corresponds to twice the uncertainty in the mean.

Table I. Median, mean, and trimmed mean of test grades for three categories of students.

Category	Median	Mean	Trimmed Mean
Blue_{β}	71.4 ± 3.5	68.0 ± 1.4	75.0 ± 1.1
Cyan_{β}	75.0 ± 3.0	71.4 ± 1.3	77.4 ± 1.0
Green_{β}	85.7 ± 2.3	78.3 ± 1.1	82.1 ± 0.9

The boxplots for all three categories are asymmetric about the median. The asymmetry is largest for the distribution on the right, which also has a number of outliers. The effect of these asymmetries and outliers is to lower the mean compared to the median, and this lowering is largest for the distribution on the right.

For comparing groups of students, such as the study that generated Fig. 9, we wish to compare performance of the groups. Measuring performance using the mean instead of the median tends to reduce the differences, which we believe is because the mean is not as appropriate as the median. In Ref. 14 we only report the medians and their uncertainties.

Using the trimmed mean reduces the difference between categories even more; we only included this column in the table for completeness. In this this case we don't know how to interpret the values. In general procedures that throw out data are dangerous, and should be avoided whenever possible

VI. DISCUSSION

For a grade distribution on a test which is roughly bell-shaped and whose histogram shows a maximum at around 50% and with negligible students with grades close to 0% or 100%, whether one characterizes the results by the mean or the median doesn't make any appreciable difference. However, we have shown that if the distribution has a maximum value that is appreciably greater than 50%, the standard way of characterizing the distribution using $\bar{x}_{est}$ and σ_{est} gives values that are too low. If the maximum value is appreciably less than 50%, the cutoff for grades < 0 means that the value of $\bar{x}_{est}$ will be too high, although σ_{est} will also be too low. If we compare two groups, one with a maximum much greater than 50% and the other with a maximum much less than 50%, we have the worst of both worlds and the difference in performance as measured by means can be significantly less than is actually the case. Although these biases are due to gray swans and not catastrophic, they nonetheless can obscure the reporting of results in the context of education research.

We have introduced a method for trimming the data that largely eliminates the biases in calculating a value for the mean. However, the method still underestimates the value of the standard deviation, which in turn underestimates the value of the uncertainty in the mean. It also reduces the measured difference between different categories of students. As mentioned, we think this procedure should be avoided.

For distributions that are not bell-shaped, such as typical results on the conceptual assessments like the Force Concept Inventory[15] or many regular class tests, characterizing the distribution with robust measures such as the median and the interquartile range is necessary. We have shown that even for bell-shaped distributions, these measures are usually preferable to ones that assume a normal distribution because of the cutoffs for grades less than 0% or greater than 100%.

In some sense all tests are a ranking exercise: we are attempting to sort students by their knowledge as demonstrated on the test. Providing feedback to the students on how they are doing compared to their classmates is an important way for students to assess how they are doing, and students take these numbers very seriously. But we have shown that the calculated mean and standard deviation of the test are somewhat misleading. In addition, beginning students struggle to understand what the standard deviation means. Giving the median and the quartiles gives students a more accurate and easier to interpret picture of how they rank.

For deeper analysis, one often calculates the Pearson correlation coefficient for results of individual questions compared to the overall performance on the test. However, this coefficient assumes a normal distribution. The Spearman correlation coefficient does not make this assumption, and is therefore preferred.[16]

For comparing two grade distributions, which is common in education research, one often uses the well-known Student's T-Test.[17] However, this too assumes that the distributions are normal. Two alternatives are the Mann-Whitney U-Test[18] and an extension, the Kruskal-Wallis one-way analysis of variance.[19] Both of these are based on the median, not the mean. Both typically return p-values, which are interpreted identically to the p-value of Student's T-Test. But, they both assume that the distributions have the same shape and differ only in the value of the medians.[20]

However, there is a growing realization that blindly relying on p-values, typically choosing a value of 0.05 to decide whether or not two distributions are the same or are different, can lead to erroneous conclusions.[21] Thus the phrase "p hacking" as a pejorative is in growing use, and a growing number of researchers in many fields, both in the physical and the social sciences, are refusing to calculate or report p-values.

Instead of or in addition to p-values, many are now calculating *effect sizes*.[22] However, one of the most common effect sizes seen in the literature seems to be Cohen's *d*, which assumes a normal distribution of the two samples.[23] Cliff's δ makes no such assumption.[24]

VII. CONCLUSION

Methods of analyzing tests or other assessments assuming a normal distribution are well known. However, this appears to us to be their only virtue. We have shown that methods based on robust measures such as the median are resistant to both black and gray swans, and avoid the biases inherent in assuming normal distributions when the grades are constrained to be between 0 and 100%. These methods are readily available for most common software. We can think of no circumstance where assuming a normal distribution and using the methods based on this assumption is preferable to using robust measures.

ACKNOWLEDGEMENT

We have benefited from discussions with David C. Bailey, Dept. of Physics, Univ. of Toronto

REFERENCES

[1] From https://www.physlab.org/class-demo/galton-board/ and used by permission.
[2] J.-A. Brown, "Using Psychology in the Physics Classroom, Five Steps to Improving Classroom Effectiveness," The Phys. Teach. **56**(1), 32 (2018). https://doi.org/10.1119/1.5018687
[3] N.N. Taleb, *The Black Swan: The Impact of the Highly Improbable*, 2nd ed. (Random House, New York, 2010).
[4] For the bin with 0 students, we take the uncertainty to be ± 1.
[5] J.D. Emerson and D.C. Hoaglin, "Resistant Lines for y versus x," in D.C. Hoaglin, F. Mosteller, and J.W. Tukey eds., *Understanding Robust and Exploratory Data Analysis* (John Wiley, Toronto, 1983), p. 129.
[6] F.J. Anscombe, "Graphs in Statistical Analysis," American Statistician **27**(1), 17 (1973). Our examples are datasets 3 and 4 respectively.
[7] Paraphrased from P. Krugman, "From Whom the Economy Grows," The New York Times (August 30, 2018).
[8] From W. Jin, R. Joyce, D. Phillips, and L. Sibieta, "Poverty and Inequality in the UK: 2011," Report C118, Institute for Fiscal Studies (2011). doi: http://dx.medra.org/10.1920/co.ifs.2011.0118 (Retrieved October 27, 2017).
[9] $\bar{x} \pm 3\sigma$ has an area under the curve of 99.73%.
[10] J.D. Emerson and J. Strenio, "Boxplots and Batch Comparison," p. 58 in Ref. 5.
[11] R. McGill, J.W. Tukey, and W.A. Larsen, "Variations of box plots," American Statistician **32**, 12 (1978). http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.1978.10479236 (Retrieved October 28, 2017). Note that in this article the multiplier is 1.57, not 1.58: since the uncertainty itself is largely heuristic, the difference in these values is trivial. We believe the value of 1.57 is probably a typo.

[12] A nice introduction is C. E. Papadopoulos and H. Yeung, "Uncertainty estimation and Monte Carlo simulation method," Flow Measurement and Instrumentation **12**(4), (2001) 291. https://doi.org/10.1016/S0955-5986(01)00015-2.
[13] J. Tukey, *Exploratory Data Analysis* (Addison-Wesley, Reading MA, 1977), p. 39.
[14] J.J.B. Harlow, D.M. Harrison, M. Justason, A. Meyertholen, C. Sealfon, and B. Wilson, "Personality types and student performance in an introductory physics course, part 2," submitted to Phys. Rev. PER.
[15] See, for example, Figure 1 of J.J.B. Harlow, D.M. Harrison, and A. Meyertholen, "Correlating student interest and high school preparation with learning and performance in an introductory physics course," Phys. Rev. ST Phys. Educ. Res. **10**, 010112 (April 7, 2014).
[16] For some examples, see http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/ (Retrieved October 29, 2017).
[17] See, for example, E.M. Pugh and G.H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, Don Mills Ontario, 1966), pg. 172 ff.
[18] H.B. Mann and D.R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," Annals of Mathematical Statistics **18**, 50 (1947).
[19] W.H. Kruskal and W.A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," Jour. of the American Statistical Association **47**, 583 (1952).
[20] A discussion of the Mann-Whitney test compared to other methods is A.J. Vickers, "Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data," BMC Medical Research Methodology **5**, 35 (2005).
[21] A favorite example is C. Aschwanden, "Science Isn't Broken," (August 19, 2015), https://fivethirtyeight.com/features/science-isnt-broken/ (Retrieved October 29, 2017).
[22] See for example, G.M. Sullivan and R. Feinn, "Using Effect Size – or Why the P-Value is not Enough," Jour. of Graduate Medical Education **4**(3), 279 (Sept. 2012).
[23] J. Cohen, "A power primer," Psychol. Bull. **112**(1), 155 (1992).
[24] N. Cliff, "Dominance statistics: Ordinal analysis to answer ordinal questions," Psychol. Bull. **114**(3), 494 (1993).

Last revision: October 27, 2018