The GRIMMER test: A method for testing the validity of reported measures of variability

Jordan Anaya[1]

[1] omnesres.com, email: omnesresnetwork@gmail.com, twitter: @omnesresnetwork

Corresponding Author:

Jordan Anaya[1]

Charlottesville, VA, US

Email address: omnesresnetwork@gmail.com

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

The GRIMMER test: A method for testing the validity of reported measures of variability

Jordan Anaya[1]

[1]**Omnes Res**

ABSTRACT

GRIMMER (Granularity-Related Inconsistency of Means Mapped to Error Repeats) builds upon the GRIM test and allows for testing whether reported measures of variability are mathematically possible. GRIMMER relies upon the statistical phenomenon that variances display a simple repetitive pattern when the data is discrete, i.e. granular. This observation allows for the generation of an algorithm that can quickly identify whether a reported statistic of any size or precision is consistent with the stated sample size and granularity. My implementation of the test is available at PrePubMed and currently allows for testing variances, standard deviations, and standard errors for integer data. It is possible to extend the test to other measures of variability such as deviation from the mean, or apply the test to non-integer data such as data reported to halves or tenths. The ability of the test to identify inconsistent statistics relies upon four factors: (1) the sample size; (2) the granularity of the data; (3) the precision (number of decimals) of the reported statistic; and (4) the size of the standard deviation or standard error (but not the variance). The test is most powerful when the sample size is small, the granularity is large, the statistic is reported to a large number of decimal places, and the standard deviation or standard error is small (variance is immune to size considerations). This test has important implications for any field that routinely reports statistics for granular data to at least two decimal places because it can help identify errors in publications, and should be used by journals during their initial screen of new submissions. The errors detected can be the result of anything from something as innocent as a typo or rounding error to large statistical mistakes or unfortunately even fraud. In this report I describe the mathematical foundations of the GRIMMER test and the algorithm I use to implement it.

Keywords: Standard deviations, standard errors, variances, statistics, reproducibility, replicability

THEORETICAL FOUNDATIONS

In 1983 Magic Johnson received 304.5 votes for Most Valuable Player (link). Votes for Most Valuable Player can only be whole numbers, and the redditors at r/nba recognized the impossibility of this sum and proposed some interesting reasons for the apparent half vote.

If something like this was discovered in the scientific literature it would probably also cause a bit of confusion. Imagine a lab reported a statistic that could only be a whole number, such as the number of mice used in an experiment. If they claimed their experiment involved 10.5 mice it would be unclear how many mice they actually used, and similarly to the half MVP vote, if the number was taken seriously it may engender some interesting speculation as to how this half mouse came to be, or perhaps there were two quarter mice, or even four eighth mice.

Just as a simple statistic such as a sum can be nonsensical and potentially humorous when dealing with discrete data, it is possible for other statistics to be just as nonsensical, albeit less conspicuously so and likely not as humorous. It is because detecting incorrect values for anything but the simplest of statistics requires more effort than just checking if the reported statistic has the same precision as the data that these errors have thus far gone undetected by the scientific community. Only recently has there been progress on evaluating the statistics of granular data, and the work revealed a striking number of nonsensical values (Brown and Heathers, 2016). One can only imagine how many errors will be revealed with further advances in this field and widespread adoption of the techniques.

Review of the GRIM test

This work and future work on statistics with granular data are possible because of the solid foundations laid by the GRIM test (Brown and Heathers, 2016). The authors of the GRIM test made the simple observation that when the values of data sets are granular, the means are also granular, and this makes certain means mathematically impossible, i.e. "inconsistent" with the data. The authors referred to their test checking for inconsistent means as the Granularity-Related Inconsistency of Means (GRIM) test.

The GRIM test is elegant in its simplicity. Given a data set with granularity G and sample size N, the granularity of the mean is $\frac{G}{N}$. For example, with a data set of 10 values where the values are reported as integers (a granularity of 1), the means of all possible sets can be enumerated:

$$0, 0.1, 0.2, 0.3, 0.4, \ldots, 1.0, 1.1, 1.2, 1.3, 1.4, \ldots$$

The granularity of the means in this case was $1/10 = 0.1$. Any means reported that are not a multiple of 0.1 would be incorrect and would fail the GRIM test.

The GRIM test has no upper limit on the size of means it can test. For example, even means around a million will still have to be a multiple of 0.1. Where the limitations of the test arise are in the sample size and the number of reported decimals of the mean. Going back to our example, if the researcher rounded his or her means to the nearest integer the possible reported means would now be:

$$0, 1, 2, 3, 4, \ldots$$

In this case the GRIM test is useless because due to rounding the granularity of the means is smaller than the granularity of the reported means. Stated generally, if the mean is reported to decimal places D, then the GRIM test can detect inconsistent means given

$$\frac{G}{N} > 10^{-D}$$

For example, if G is 1, i.e. integer values, and the means are reported to two decimals (D is 2), the test is applicable for sample sizes up to 99, after which the granularity of the means are equal to or less than the granularity of the reported means. This is because at two decimals the granularity of the reported means is .01, and at a sample size of 100 the granularity of the means is also .01.

A naive algorithm for the GRIM test would be to enumerate all possible means, round them to the same decimals as the reported mean, and then check if the mean is one of the possibilities. Although this would work, the authors of the GRIM test created a clever algorithm for testing the consistency of means:

1. Multiply the reported mean by the sample size
2. Round the result to the nearest integer
3. Divide by the sample size
4. Round the result to D decimal places and compare to reported mean

This algorithm assumes a granularity of 1 for the values. If the data is present at a different granularity you must round to the nearest granular value instead of rounding to the nearest integer at step 2.

Beyond the GRIM test

A natural extension of the GRIM test would be to apply it to other commonly reported statistics, for example standard deviations. While it might be obvious that the standard deviations of granular data should also be granular, it is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

unclear *a priori* what that granularity might be. Going back to our example, a sample size of 10 with a granularity of 1 produces means with a granularity of 0.1, but what is the granularity of the standard deviations it produces? The formula for population standard deviation, henceforth referred to simply as standard deviation or σ, is

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

It is unclear to me how to determine the granularity of standard deviations from this formula.

Another problem with standard deviations is the presence of the square root. The ultimate goal of a granularity test is to determine if the fractional component of the reported statistic is consistent regardless of the value of the integer component. We saw that when working with means 10.1 was a consistent value, as was 100.1, as was 1,000.1, etc. It is difficult to imagine that standard deviations would display any sort of consistent fractional values. The curve of $\sqrt{x}$ is just that, a curve, and values become compressed as $x \to \infty$, with the effect that the integer portions of the values are not independent of the fractional components.

To get around this we can use another commonly reported statistic, the variance, which is σ^2. With the elimination of the square root there may exist a simple granularity G that is independent of the size of the variance. But again, it is unclear how to determine what this granularity might be. A naive approach would be to employ a brute force algorithm that records all possible variances and then checks a reported statistic against this table of values. However, there is no upper limit to variances, and computing all possible variances for a given data set can quickly become computationally intensive.

Nota bene: From now on the data will be assumed to be integers and thus have a G of 1.

Using the itertools library for Python it is possible to generate all possible combinations of a data set of size N and possible unique values $0, 1, 2, 3, \ldots, X$ with the function `combinations_with_replacement`. The number of possible unique combinations that the function will generate is

$$\frac{(X+N)!}{X!N!}$$

What this equation shows is that when N or X gets large the number of possible combinations that need to be tested quickly become unmanageable. Despite this I decided to start calculating as many variances as I could and see what I observed.

Enumerating the unique possible variances for a sample size of 5 and sorting them results in this sequence:

$$0.0, 0.16, 0.24, 0.4, 0.56, 0.64, 0.8, 0.96, 1.04, 1.2, 1.36, 1.44, 1.6, 1.76, 1.84,$$
$$2.0, 2.16, 2.24, 2.4, 2.56, 2.64, 2.8, 2.96, 3.04, 3.2, 3.36, 3.44, 3.6, 3.76, 3.84, \ldots$$

At first glance it may appear that the granularity of the variances, G_{σ^2}, is 0.16, however there are times when the step is .08 instead of 0.16, such as the step between the values 0.56 and 0.64.

A close inspection of the sequence reveals that the fractional values of variances that have an even integer are the same regardless of the size of the even integer. Similarly, the fractional values of variances that have an odd integer are the same regardless of the size of the odd integer. I will refer to the fractional values for even integers as the even pattern (EP), and the fractional values for odd integers as the odd pattern (OP). I will also take the liberty to refer to variances with an even integer component as even variances (EV) and variances with an odd integer component as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

odd variances (OV).

These repeating fractional values are not limited to a sample size of 5. Enumerating the unique possible variances for a sample size of 6 and sorting them results in this sequence:

$$0.0, 0.13\overline{8}, 0.2\overline{2}, 0.25, 0.3\overline{3}, 0.47\overline{2}, 0.5\overline{5}, 0.58\overline{3}, 0.6\overline{6}, 0.80\overline{5}, 0.8\overline{8}, 0.91\overline{6},$$
$$1.0, 1.13\overline{8}, 1.2\overline{2}, 1.25, 0.3\overline{3}, 1.47\overline{2}, 1.5\overline{5}, 1.58\overline{3}, 1.6\overline{6}, 1.80\overline{5}, 1.8\overline{8}, 1.91\overline{6}, \ldots$$

In this case the odd pattern is the same as the even pattern. This leads to the following theorem:

Theorem 1

If N is even: $EP = OP$
If N is odd: $EP \neq OP$

I am completely unaware of any previous reports of repeating patterns of variances of discrete numbers and this may be the first time this is reported in the literature.

It appears we now have everything we need to apply granularity testing to variances, and consequently standard deviations and standard errors since they can be derived from variances. However, the power of this granularity test to detect inconsistent values appears weaker than the GRIM test. Recall that the GRIM test can detect inconsistent values for means rounded to one decimal for sample sizes up to 9. However, at only a sample size of 5 we see that a test for variances is already beginning to break down if the reported statistic is rounded to 1 decimal. For example, the variances 0.16 and 0.24 would both get rounded to 0.2 and be indistinguishable.

But does the test have to be less powerful than the GRIM test? When researchers report a variance or standard deviation they also often report a mean. Can we somehow use that mean to narrow down which variances in EP and OP we should test? Intuition says yes. For example, imagine the reported variance is 0.0. The only way to have a variance of 0 is for every value in the data set to be the same. If every value in the data set is the same then the mean would have to have a fractional value of 0 since the mean would just be the integer value that is repeated in the data set.

To investigate this I began recording which means are consistent with which variances. Below are the means coupled to the variances for a sample size of 5:

Table 1. Means mapped to variances ($N = 5$)

(a) Even pattern

σ^2	Means
0.0	$0.0, 1.0, 2.0, \ldots$
0.16	$0.2, 0.8, 1.2, \ldots$
0.24	$0.4, 0.6, 1.4, \ldots$
0.4	$1.0, 2.0, 3.0, \ldots$
0.56	$0.8, 1.2, 1.8, \ldots$
0.64	$0.4, 0.6, 1.4, \ldots$
0.8	$1.0, 2.0, 3.0, \ldots$
0.96	$0.8, 1.2, 1.8, \ldots$

(b) Odd pattern

σ^2	Means
1.04	$1.4, 1.6, 2.4, \ldots$
1.2	$1.0, 2.0, 3.0, \ldots$
1.36	$0.8, 1.2, 1.8, \ldots$
1.44	$0.6, 1.4, 1.6, \ldots$
1.6	$1.0, 2.0, 3.0, \ldots$
1.76	$1.8, 2.2, 2.8, \ldots$
1.84	$1.4, 1.6, 2.4, \ldots$
3.04	$1.4, 1.6, 2.4, \ldots$

What I hope this table shows is that only certain means are consistent with certain variances. In fact, for a given variance the means that are consistent for that variance always either have fractional component F, or fractional

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

component $1-F$. For instance, an even variance ending in 0.16 can have a mean with a fractional component 0.2, or a fraction component $1-0.2=0.8$. This rule does not hold for all sample sizes (see below for more details).

This observation now provides us with a second check. Actually, it provides a third check if we first apply the GRIM test on the reported mean. As a result, when a researcher reports a mean and a variance, for the values to be consistent the mean must first pass the GRIM test, the variance must then match either pattern *EP* or *OP*, and then the mean must be consistent with the *EP* or *OP* value. Failing any of these three checks indicates the values were reported incorrectly or possibly fabricated.

Incorporating the mean into the test increases the power of the test. Now if the variances for a sample size of 5 are rounded to a single decimal, and the mean is provided, it is possible to determine if a reported value of 0.2 originated from a variance of 0.16 or 0.24. More importantly, if a researcher reports a variance of 0.2 but reports a mean of 1.0, they would not pass the test despite the fact that two possible variances round to 0.2. And of course if they report any means with an odd number as a fractional value such as 1.1 they would also fail the GRIM test.

Given that the authors of the GRIM test likely chose that name because it can give a grim view of a person's work, and because a variance can be thought of as an "error" of a measurement and my test may provide an even grimmer view of an article, I refer to my test as the Granularity-Related Inconsistency of Means Mapped to Error Repeats (GRIMMER) test.

More on those patterns

The reader may be wondering why I started the discussion of patterns with the pattern for a sample size of 5 instead of a smaller sample size and presumably simpler pattern. The reason for this is for a sample size of 2, 3, or 4, there is not a simple repetitive pattern. However, these seem to be the only sample sizes for which this is the case. I have recorded the variances up to a sample size of 99 and have been able to observe a pattern in each case.

The variances for a sample size of 2 do show a pattern, however the pattern is not simple:

Table 2. Variances and steps for $N=2$

σ^2	$\sigma^2_{i+1}-\sigma^2_i$
0.0	0.25
0.25	0.75
1.0	1.25
2.25	1.75
4.0	2.25
6.25	2.75
9.0	3.25
12.25	3.75

In this case the variances do not follow a repetitive pattern, but the differences do. There are multiple ways to reproduce this variance sequence, one such way is to let X be $0,1,2,\ldots$ and plug X into these equations:

$$X^2, X^2+X+0.25$$

The variances for sample sizes 3 and 4 are more complicated than this and I haven't worked on trying to define equations that can reproduce them. It is unclear how often a test for sample sizes 2, 3, and 4 would be used, but adding this functionality to the GRIMMER test will be worked on at a future date.

Another interesting observation that I made involves the length of the repetitive sequences.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Table 3. Length of variance patterns

N	**len** EP + **len** OP
5	15
6	24
7	28
8	32
9	36
10	60
11	66
12	72

This leads us to the next theorem:

Theorem 2

The length of the full pattern (EP and OP) is a multiple of N

Despite knowing about this rule it is still unclear how to predict how long the pattern will be for a given sample size. For example, at a sample size of 97 the full pattern is 49 times the sample size, but at a sample size of 99 the full pattern is only 24 times the sample size.

The patterns are also not as simple as the patterns for sample sizes 5 and 6 would suggest. Starting at a sample size of 9 the variances with zero as an integer component do not match the even pattern. The values are contained within the even pattern, but there are missing values. As a result, I had to define another pattern which I refer to as the zero pattern (ZP).

As mentioned before, the means for a given fractional value do not always follow the pattern F, $1-F$. For example, at a sample size of 9 variances ending in $.\overline{4}$ always map to means that end in either $.\overline{3}$, $.\overline{6}$, or .0. And at a certain sample size some means begin to display a F_1, $1-F_1$, F_2, $1-F_2$ pattern, or even a F_1, $1-F_1$, F_2, $1-F_2$, F_3, $1-F_3$ pattern. It is unclear how to predict when a given fractional value of a variance will map to 2, 3, 4, 6, or possibly more unique fractional components of means.

IMPLEMENTATION

Implementing a test for variances requires two things: (1) identifying the patterns for each sample size and the matching means; (2) creating an algorithm that uses the knowledge of the patterns to check if a mean-variance pair is consistent. Neither of these two things should be difficult, however floating point issues can cause problems and if not handled correctly affect the accuracy of the test. Because dealing with floating points is nontrivial I will describe how I went about implementing the GRIMMER test and some tricks I used.

Finding the patterns

To identify the patterns for a given sample size we have to enumerate the variances for every possible unique combination of a data set that contains the values $0, 1, 2, 3, \ldots, X$. The question then is how many variances do we need to calculate before we can see the pattern? It might be tempting to think that if we know what the maximum variance for a set is for a given X we would have saturated all variances up to that value. For example, the maximum variance for $N = 6$ occurs with this set: $[X, X, X, 0, 0, 0]$. If $X = 4$ then the maximum variance happens to also be 4, and we might think that we have identified every variance that exists from 0 to 4. However, there will be missing variances.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Given a sample size N, and maximum value X, it is unclear how to determine at what variance value below which the variances will be saturated. One thing I tried was to calculate the variances for a given X, then calculate the variances for $X+1$, sort the values and see at which value below which the variances were identical. Although this works fairly well, I was able to identify cases where the variances were not saturated and $X+2$ added new variances.

As a result, I am unsure how to choose the minimum value for X that will reveal the patterns for a given N. However, once a value for X is chosen (6 seems to suffice), it is not difficult to be fairly confident that the variance pattern has been saturated. From Theorem 2 we don't know the exact length of the pattern, but we know that it must be a multiple of the sample size. When inspecting the variances, if the fractional values for an integer of 1 equal the fractional values for an integer of 3, that is assumed to be pattern OP. If the fractional values for an integer of 2 equal the fractional values for an integer of 4, that is assumed to be pattern EP. If EP and OP are consistent with Theorem 2 the patterns are considered saturated and correct.

Even using only an X of 6 computing times can approach an entire day for higher sample sizes. One trick I employed to reduce the number of variances that had to be calculated was to only calculate variances for sets which contain X. It can be shown that the variances calculated for all possible sets with the highest value will contain the variances for all the possible sets which do not contain the highest value. This is shown by using the fact that adding any array, $c[i,\ldots,i]$, with constant values to another array will not affect the variance. As a result, any array which does not contain the maximum value can have the array, $[1,\ldots,1]$, added to it until it contains the maximum value. It will then have the same variance as the array with the maximum value, and therefore should be ignored.

One consequence of only computing sets with the maximum value however is that the means are not saturated for low variances. Just as I am unsure how to prove variances are saturated, I am unsure how to prove the means mapping to a given variance are saturated. When identifying the means I therefore took the means from the variances that had the most means and assumed they were saturated. All files and code used are present at the GitHub repository.

Implementing an algorithm

Checking if a reported variance matches either EP or OP is fairly straightforward, but checking a reported standard deviation requires a few more steps. One method might be to enumerate all possible variances starting with 0, convert them to standard deviations, and stop once a standard deviation matches or a standard deviation is greater than the reported standard deviation. However, this involves unnecessary computing and would take some time for very large standard deviations.

It is not difficult to narrow down the possible variances that need to be tested. Imagine someone is interested in checking if a reported standard deviation of 1.00 is consistent. Depending on how they do their rounding the original value could have been anything from .995 to 1.005. As a result, this represents the range of standard deviations that need to be checked. The algorithm for standard deviations is thus:

1. Determine how many decimals (D) the SD is reported to
2. Identify lower and upper bounds by $(SD \pm {}^{0.5}/_{10^D})^2$
3. Floor and ceil these values just to be safe
4. Enumerate the possible variances between these values
5. Convert to SD and round to D decimals
6. Check if any match the reported SD
7. If mean was provided perform GRIM test and check if mean matches the variance

A similar algorithm would be used for standard errors.

The issue with standard deviations (and standard errors) is the presence of the square root. The square root compresses values that would normally be distinguishable. For example, at a sample size of 5 the variances 100.56 and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

100.64 would clearly be distinguishable at a precision of two decimals. However, converting these values to standard deviations and rounding to two decimals results in the same number, 10.03. Because of this, as standard deviations and standard errors get larger it becomes more difficult to detect inconsistent values and the test is better suited for variances or small standard deviations and standard errors.

When using this algorithm for variances and means stored to a large number of places or applying the test to large numbers, floating point errors can creep in. One strategy I employed with Python is not relying upon modulo to get fractional values as it is error prone. Instead I opted to convert the value to a string and split at the decimal, and then add the decimal back while it is still a string. The Python `str()` method only converts numbers to a certain number of places so I opted for the `repr()` method instead. A more advanced technique is needed when floating point errors can't be avoided. If a list of floats gets altered at the 12^{th} decimal place through a floating point error the order of the list will still be the same. As a result, when trying to map a variance to its consistent means the variances do not have to match exactly, only the index position in the list must be known.

Future additions

It is possible to trick the GRIMMER test. For example, if your data is a Likert-type scale and the smallest possible value is 1, then the mean could not possibly be less than this value. However, the GRIMMER test does not know about this, and will happily treat a mean of 0.0 as consistent. In fact, the GRIMMER test accepts negative means and is applicable for data sets that contain negative numbers. Adding this check to the test would not be difficult, but this type of error is so obvious that a test shouldn't be needed for it.

However, there is a less obvious addition that could increase the power of the GRIMMER test. If a researcher has a data set with positive integers and reports the sample size as 5 with a mean of 2.0 and a variance of 100.8, that seems impossible. And yes, it is impossible. However, the GRIMMER test assumes that values can be negative, and in that circumstance it is entirely possible to have a variance that large with a mean that small. But if the values are positive integers, for a given variance there is a minimum consistent mean. I am unsure how to determine what this minimum mean is when provided details about the data set. If a mean passes the GRIM test the GRIMMER test assumes that the mean is correct and proceeds to check if the mean-variance pair is correct without trying to determine if the mean is too small for the variance. If you believe the reported mean is clearly too small for the variance it would not be difficult to check what the maximum possible variance for that mean is. If the reported variance is less than this maximum variance then clearly there is a problem. If not, then it is unclear if the mean is too small or not and a brute force approach will be required.

It is easy to extend the GRIMMER test to granularities other than 1 (integer steps). To do so simply construct new ZPs, EPs, and OPs by multiplying the variances for a granularity of 1 by G^2_{new}. The means mapped to each variance must then be multiplied by G_{new}. Although this is trivial, it is unclear how useful this would be to the scientific community. In order to detect inconsistent variances for a granularity less than 1 the sample sizes would either have to be very small or the variances would have to be reported to more than 2 decimals.

Other measures of variability such as deviation from the mean also seem to show simple repetitive patterns. It is unclear how often these measures are reported in the literature and if the scientific community would be interested in a test for these statistics.

DISCUSSION

There are a lot of questions to ask about the GRIMMER test. How powerful is the test? Who should be using it? What should people do if they find inconsistent values? Can errors or fraud in the literature explain the current reproducibility crisis? I will suggest answers to these questions below, but ultimately scientists, publishers, and individual fields will have to decide on how they want to use the test and the implications of its results.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Power considerations

The first question a researcher might ask about the GRIMMER test is just how good is it? Well, that depends. It depends on the sample size, granularity of the data, number of decimals reported, and for standard deviations and standard errors, size of the statistic. It also depends on whether the mean is available. The test can be performed without a mean, but it is more powerful when combined with a mean. And when a mean is provided, remember that any errors detected by the GRIMMER test will be in addition to errors detected by the GRIM test. In fact, when the GRIM test fails it is impossible to perform the full version of the GRIMMER test since the full version checks if the the mean is consistent with the variance, and if the mean itself is inconsistent then by default the mean-variance pair (MVP) is inconsistent. If the mean fails the GRIM test and the user still wants to check the statistic of variability, he or she can simply leave out the mean in the web application.

What then is the chance of passing the GRIMMER test assuming the researcher picked some values out of a hat? The first thing that would have to occur is passing the GRIM test, which is powerful in its own right. For a granularity G, sample size N, and decimals D, the probability ($P_{failure}$) of failing the GRIM test is

$$P_{failure} = \begin{cases} 1 - \dfrac{N}{10^D G} & \text{if } \frac{G}{N} > 10^{-D} \\ 0 & \text{otherwise} \end{cases}$$

To see how this equation works imagine a researcher reported a mean to 2 decimals for integer data. If the sample size was 1, there would be a 99% chance of failing the GRIM test. For a sample size of 1 the only way to pass the test is to have a fractional value of 0.00. And at 2 decimals there are 100 possible fractional values. At a sample size of 10 the researcher would then have a 90% chance of failing the test, and at a sample size of 50 the researcher would have a 50% chance of failing the test. As a result, if means are reported to a high precision the GRIM test on its own provides an excellent check for data and if a researcher is just randomly making up values the GRIM test will likely catch them.

So what is the $P_{failure}$ of the GRIMMER test given that the researcher passed the GRIM test? It is not possible for me to write a general equation for the GRIMMER test since I don't have a general equation that describes the pattern for each sample size. In addition, for standard deviations and standard errors $P_{failure}$ is complicated by the fact that it is a function of the size of the statistic. However, it is possible to simulate the probabilities.

Figure 1 shows the results of one such simulation. The simulation was performed with the SD and mean at 2 decimals as I believe 2 decimals is what is most likely to be reported in the literature. In addition, the test was performed assuming a sample standard deviation s. Every possible s to 2 decimals was enumerated, i.e. $.00, .01, .02, \ldots$, and for each s the GRIMMER test was performed with every possible GRIM consistent mean. As a result, any failures of the test are in addition to those that would have been detected by the GRIM test. The figure essentially shows what percent of mean-standard deviation pairs are inconsistent.

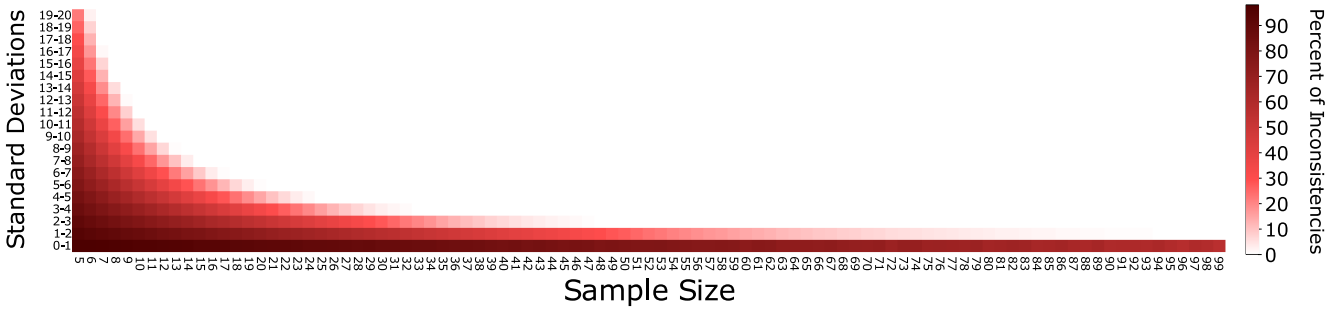

Figure 1. Sample SD $P_{failure}$ for $D = 2$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

As expected, the figure shows that $P_{failure}$ is higher for smaller sample sizes. The figure also shows the striking effect that the size of the standard deviation has on $P_{failure}$. Starting at a sample size of 51 the $P_{failure}$ is 0 for sample standard deviations above 2. Interestingly, if s is below 1 the test is still highly effective for large sample sizes, having a $P_{failure}$ of 56% at a sample size of 99. Keep in mind that the test will be much less powerful for a D of 1, and much more powerful for a D above 2.

Where the test really shines is in testing variances. Because the values are not compressed by a root operation, the $P_{failure}$ is independent of the size of the variances. Figure 2 shows the results of a simulation with sample variances and means at $D = 2$. Whereas the GRIM test has a $P_{failure}$ of 1% at a sample size of 99, the $P_{failure}$ is 51% for the GRIMMER test.

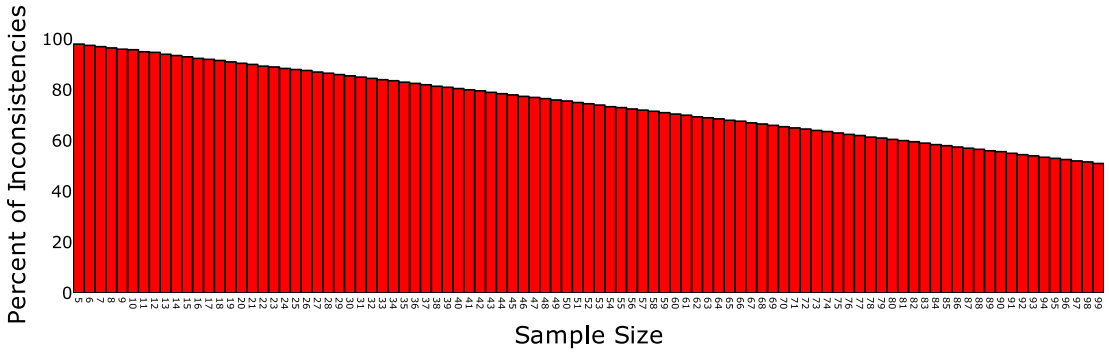

Figure 2. Sample Variance $P_{failure}$ for $D = 2$

The graph also shows that $P_{failure}$ might follow a line. In fact, at each sample size the $P_{failure}$ drops by about 0.5% and $P_{failure}$ for s^2 at $D = 2$ can be approximated by this equation

$$P_{failure} \approx \max \left\{ \begin{array}{l} -0.5\%(N-5)+98\% \\ 0 \end{array} \right\}$$

Setting $P_{failure}$ equal to 0, this equation indicates that the test would be able to detect inconsistent s^2-mean pairs up to a sample size of 200. This also suggests that the test could detect inconsistent s-mean pairs at a sample size of 200 provided $s < 1$.

How should the test be used?

With great power comes great responsibility, and although researchers are free to use the test however they want, I don't intend for the test to be used to conduct witch hunts. As mentioned, if all the statistics are fabricated, the GRIM test alone will likely be sufficient to detect the fraud. I envision the GRIMMER test to primarily be a check for honest errors that could happen to anyone and if a paper contains a certain number of GRIMMER (or GRIM) errors it may indicate a general sloppiness of the work and may make you think twice about trusting the results of the paper.

Failing the GRIMMER test does not imply the researcher fabricated the statistic. Although the number is undeniably wrong, there are a lot of possible explanations for why the researcher reported an incorrect statistic. For one, the GRIMMER test assumes that the only rounding that occurred was in the final reporting of the statistic. If for some reason the researcher calculated the statistic by hand and rounded at individual steps, he or she may fail the test. Even if the only rounding occurred at the final step, it is possible to fail the GRIMMER test due to different rounding conventions. By default the GRIMMER test always rounds up when the last digit is a 5, but some software may round down, or perform bankers' rounding. This affects sample sizes which are a multiple of 8, and to account for this the web application allows users to either round up or round down.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

There are a variety of other reasons a researcher may fail the test. They could have made a simple typo when reporting the mean, variability, or the sample size. Or perhaps the data has an unusual granularity, for example $0, 0.5, 1, 2, 3, \ldots$, and the reader assumed the granularity was 1. Or maybe the statistic was reported as a sample standard deviation but the researcher accidentally calculated a population standard deviation. The possibilities are endless.

Regardless of the reason for the inconsistency, all a researcher has to do to resolve the inconsistency is provide the data that was used. If a researcher provides the data and an error is confirmed it is unclear if this mistake is large enough to warrant publishing a correction. And if the researcher refuses to provide the data used it will be up to the reader to decide what to do next.

Where I think the test is best used is in providing confidence in a reported study. If I were to plan a project where the premise relied upon a small number of publications being correct, I would want to know if the publications contained any errors. If I checked the statistics in the studies and key values failed the GRIMMER test I might think twice about dedicating a lot of time and resources to a project that builds upon the results of the studies, and might hold off until I contacted the researchers to try to understand why their publications contain errors.

Implications for reproducibility

It is said that we are in the midst of a reproducibility crisis Ioannidis (2005), although it should be noted that some believe these concerns are overblown (Patil et al., 2016; Jager and Leek, 2014). Regardless of the exact state of scientific reproducibility, I'm sure most would agree that we could be doing better. One of the assumptions of a replication study is that the study was performed and reported correctly. But if the results of a study appear too good to be true, maybe they are.

One can't help but wonder if at least some of the current reproducibility problems aren't due to clerical errors, incorrect analyses, or fudging of statistics. There are many reasons why science may be irreproducible Begley and Ioannidis (2015), but reporting errors do not have to be one of them. If journals performed a check on all granular statistics at submission these errors could be corrected before the work is published. A cynic may claim this will give researchers committing fraud a chance to hide the evidence. To counteract this if a submitted paper has a large number of GRIM or GRIMMER errors the journal could request to see the raw data to investigate if fraud indeed occurred.

It should be noted that if the data itself is fabricated and the summary statistics are calculated correctly the GRIMMER test will not detect the fraud. However, fabricating the raw data requires more work than making up the summary statistics and there are techniques that exist to identify fabricated data (Hartgerink et al., 2016).

As journals are the gatekeepers of the scientific literature and supposedly add value to publications, they must take at least some blame for allowing blatant statistical errors to be published. Ideally they would require all researchers to share their raw data, thus allowing for a simple reanalysis of the data and eliminating the need for the GRIMMER test. But because of meddling "research parasites" who might use the data to make important scientific advances, this is a controversial policy (Longo and Drazen, 2016; Devereaux et al., 2016). If journals are not going to request the raw data the least they could do is request that summary statistics be reported to a large number of decimals to increase the power of granularity testing. I understand it is unseemly for a high precision to be present in publications, but the unrounded numbers could be present as supplemental information.

CONCLUSIONS

I set out to create a granularity test for measures of variability. In doing so I discovered a statistical phenomenon where the fractional values of variances of discrete data repeat indefinitely, and only certain fractional components of means are possible for certain variances. I named the resulting consistency test the GRIMMER test, and developed a web application for easy application of the test.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016

Consistency testing for granular data is a novel concept introduced by the GRIM test, and the present work represents an important advance in this fledgling field. It will be interesting to see if researchers are able to extend granularity testing to other commonly reported statistics such as the t-statistic, and if more general properties of granular data will be discovered, perhaps culminating in an aptly named GRIMMEST test.

It will be up to the scientific community to determine if they want to open Pandora's box and identify all the errors in the published literature. While researchers have this right, and at the very least should use the test to check publications which affect their work, I believe the test is best used to prevent future errors from being published. If researchers applied the test to their own work prior to submission, or journals made the test part of their standard screening procedure, the accuracy of scientific research can only improve.

METHODS

All the code for enumerating the variances, extracting the patterns, and implementing the GRIMMER test, along with producing the figures in this paper, is present at the GitHub repository. In addition, the web application for the test is open source. The code was run with Python 2.7 on an i7-6700K processor running Windows 7 64-bit.

ACKNOWLEDGMENTS

I would like to acknowledge the creators of the GRIM test, Nicholas Brown and James Heathers, not only for being the founding members of this field, but also for preprinting their work. If not for this, granularity testing would still be unknown to the world and this publication would not have been possible. I must also acknowledge that through conversation with Nicholas Brown I learned that he was interested in extending the GRIM test to other statistics and I only started this work after my discussions with him.

REFERENCES

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science improving the standard for basic and preclinical research. *Circulation research*, 116(1):116–126.

Brown, N. J. L. and Heathers, J. A. J. (2016). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *PeerJ Preprints*, 4:e2064.

Devereaux, P. J., Guyatt, G., Gerstein, H., Connolly, S., and Yusuf, S. (2016). Toward fairness in data sharing. *N Engl J Med*, 375(5):405–7.

Hartgerink, C., Wicherts, J., and van Assen, M. (2016). The value of statistical tools to detect data fabrication. *Research Ideas and Outcomes*, 2:e8860.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8):e124.

Jager, L. R. and Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1):1–12.

Longo, D. L. and Drazen, J. M. (2016). Data sharing. *N Engl J Med*, 374(3):276–7.

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspect Psychol Sci*, 11(4):539–44.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2400v1 | CC BY 4.0 Open Access | rec: 29 Aug 2016, publ: 29 Aug 2016