Article

Recurring Errors in Studies of Gender Differences in Variability

Theodore P. Hill [1,*,†] and **Rosalind Arden** [2,†]

1   School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
2   Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science, London WC2A 2AE, UK
*   Correspondence: hill@math.gatech.edu
†   These authors contributed equally to this work.

Abstract: The past quarter century has seen a resurgence of research on the controversial topic of gender differences in variability, in part because of its potential implications for the issue of under- and over-representation of various subpopulations of our society, with respect to different traits. Unfortunately, several basic statistical, inferential, and logical errors are being propagated in studies on this highly publicized topic. These errors include conflicting interpretations of the numerical significance of actual variance ratio values; a mistaken claim about variance ratios in mixtures of distributions; incorrect inferences from variance ratio values regarding the relative roles of sociocultural and biological factors; and faulty experimental designs. Most importantly, without knowledge of the underlying distributions, the standard variance ratio test statistic is shown to have no implications for tail ratios. The main aim of this note is to correct the scientific record and to illuminate several of these key errors in order to reduce their further propagation. For concreteness, the arguments will focus on one highly influential paper.

Keywords: intrasex variability; variability ratio; statistical errors; gender differences; sex differences; greater male variability hypothesis; tail ratios

1. Variability Ratios and Greater Male Variability

As noted by psychologists, "Few topics in psychology can rival sex differences in their power to stir controversy and captivate both scientists and the public" ([1] p. 1) and its history "has been long and contentious" [2]. The so-called greater male variability hypothesis, which dates back to Darwin's observations, says that for many traits in many species, including humans, males vary more than females. Unfortunately, recent research on this important subject continues to suffer from several basic statistical, inferential, and logical errors. The purpose of the present note is to identify and explain some of these errors, and, for specificity, it pays special attention to one paper [3] in the flagship publication of the American Mathematical Society, which has been widely cited (nearly 200 Google Scholar citations, and quoted in leading scientific publications, including *Proceedings of the National Academy of Sciences*, *Scientific American*, and *Science*). The goal here is to identify and explain some of these basic errors.

2. Variance Ratios and the Variance Ratio Test

The most widely used method for quantifying gender differences in variability is extremely simple in principle, and interprets "variability" as "variance" in the formal statistical sense. (By variability, Darwin almost certainly did not mean statistical variance since that term was first introduced by R.A. Fisher in 1918.) In this method, the so-called variance ratio (VR) is defined, by convention, as the variance of the male data divided by the variance of the female data. Thus, with this interpretation of variability, there is greater male variance (GMV) in a particular trait of a given sexually-dimorphic species if

the variance ratio for that trait is greater than 1, and there is greater female variance (GFV) if the variance ratio is less than 1.

It should be noted that there are also other standard tests for homogeneity of variance among populations; one such test is the classical test of Levene [4]. A few studies on gender differences in variability, such as [5,6], employed Levene's standard test, but it has been argued that "the results of Levene's test alone should not be interpreted as evidence for or against GMV but rather, the latter should rely on the broader pattern of results, including the variance ratio and the ratio of males and females in different regions of the distribution" ([7] p. 4). Most current research studies on gender differences in variability, however, use the variance ratio method, and that is the focus of this article.

In research on sex differences, in general, the subject of quantification is crucial in interpreting the results of the studies. Del Giudice's treatise [1] provides a concise and systematic review and comparison of the main methods used to measure sex differences and similarities, including standard metrics such as Cohen's *d* for comparing differences in means, and the variance ratio method discussed here, which is the standard for comparing differences in variation.

It has been noted that "the substantive interpretation of [Cohen's] *d* values is a persistent source of confusion" ([1] p. 5); the same is true for variance ratio values. There is no accepted rule for determining exactly how much larger than 1 the variance ratio must be to consider it significant, and of course, this depends on the goal of the study under question. As noted in [1], what counts as "small" or "large" depends entirely on the area of research, the variables under consideration, and the research question. Hyde and Mertz [8], for example, label variance ratios between 1.05 and 1.20 as not radically different from 1; similarly, Kane and Mertz's finding of VR = 1.08 for the mathematics performance led them to conclude that this is "pretty clear data *debunking* the greater male variability hypothesis" ([9] emphasis added). In contrast to the interpretation of variance ratio values, on the other hand, a subsequent study of global gender differences in variability in mathematics and science achievement, based on variance ratios between 1.09 and 1.14, concluded that "the variability of boys' performance in science was larger than that of girls" ([10] p. 651 and Table 5).

A crucial factor missing in almost all greater male variance studies is the importance of the actual distribution of the data. If the data are known to be normally (Gaussian) distributed, the tail behaviors [11] are always extreme. As shown in [12], given any number of different normally-distributed distributions, one of those will always completely overwhelm all of the others in the right tail (but this conclusion is not true for other bell-shaped distributions, such as the common Cauchy distribution.) As a corollary in the variance ratio setting, the tail behaviors are always extreme unless both variances are identical.

Proposition 1. *Suppose that the male and female values for a given trait are normally distributed. Then if VR > 1, the male population will completely dominate both the upper and lower tails of the combined population, and if VR < 1, the female population will completely dominate both tails.*

Proof. If VR $\neq$ 1, then the distributions of male and female values are not identical, and the conclusion follows directly from Theorem 1(iii) in [12]. □

The points at which domination begins for high and low values depend on the specific values of the parameters for the normal distributions. The next example illustrates this dependence of the variance ratio method on the underlying distribution by presenting three different hypothetical male/female pairs of distributions with identical variance ratios but with radically different proportions of males and females at the high- and low-end values.

Example 1. *Let F and M denote the distributions for females and males, respectively, of a given trait in a given species. It is easy to see that the following three cases have identical variance ratios of 1.21:*

(i)   F is uniformly distributed between 150 and 160, and M is uniform between 139 and 150.
(ii)   F is uniformly distributed between 130 and 140, and M is uniform between 140 and 151.

(iii)   *F is normally distributed with a mean of 150 and a standard deviation of 10, and M is normally distributed with a mean of 140 and a standard deviation of 11.*

In these three cases, the high- and low-end statistics are radically different. In case (i), even though VR > 1, all F values are larger than all M values; in case (ii), all M values are larger than all F values; and in case (iii), even though the M average is smaller, the M values strongly dominate the high-end values, which follows from Proposition 1.

Similarly, there are *F* and *M* distributions with large variance ratios (e.g., VR = 1.21 as in the above example), but the tails of both *F* and *M* are exactly the same, with the differences in variance occurring solely in the center of the distributions. The construction of such pairs of distributions is left to the interested reader.

Thus, if the underlying distribution of the data is not known or not specified, then a given value for the variance ratio does not imply anything about gender differences in high-end values, low-end values, or any other range. If the data are known to be normal, on the other hand, then the means and variances completely determine the relative proportions of each gender in any given range, including the tail ratios of M/F proportions above any given cutoff. In short, this establishes the following general fact.

F   The numerical value of a variance ratio alone, without additional assumptions about the underlying distributions, implies nothing about the tail ratio comparisons.

If one is interested in gender differences in mathematical ability, as in the Kane and Mertz study, then it may be useful to look at raw data without calculating variance ratios. For example, the official historical summary of 20 years of results of SAT-M, the mathematics portion of the Scholastic Aptitude Test, shows greater male standard deviations in every single year, with an average variance ratio of approximately 1.14 ([13] p. 25). If the tail ratio behavior is more important than differences in variance, however, note that the male standard deviations there are not much larger than those of females, but the mean values of the males are noticeably higher. This combination can have significant effects on the right tails since, as Feingold pointed out, "what might appear to be trivial group differences in both variability and central tendency can accumulate to yield very appreciable differences between the groups in numbers of extreme scorers" ([14] p. 11); see also Example 3 in [12].

3. Variability Ratios in Mixtures of Distributions

One argument that is being used for deciding whether or not there are generic differences in variability between men and women, for example in mathematics performances, as in the Kane and Mertz study, is by computing the variance ratios across countries and comparing those values with the overall worldwide variance ratio. Kane and Mertz argue that if there were greater male variance, the variance ratios for *all* countries should be greater than unity and similar in value ([3] p. 13). As *Science* magazine reported, "If the greater male variability hypothesis, which posits that men have a greater range of intelligence than women, is true, then that variability would persist, consistently, across all 86 countries" [9]. The same logic is also repeated elsewhere in the scientific literature, e.g., "the male greater variability hypothesis does not accommodate the staggering cross-country differences found here" ([15] p. 438).

That is, the underlying argument here is based on the following claim (C):

(C)   If VR > 1 worldwide for a particular trait in humans, then the corresponding variance ratios for each country should all be greater than 1 and similar in value.

These studies do not give justification for claim (C); although (C) may appear intuitive and plausible, as stated, it violates a standard statistical fact, namely, the formula for the variance of a weighted finite mixture of distributions. This will be seen in the next example, which is hypothetical and illustrates how a union of countries (in this case, only two) could exhibit greater male variability as a whole, even though not all of the individual countries do. That is, VR > 1 for the union, but not for each country.

Example 2. *A population consists of two countries C_1 and C_2, with equal numbers of people in each, divided equally in each among men and women. Measurements of height administered to everyone in the overall population result in means m_1 and m_2 and standard deviations σ_1 and σ_2 for the men in countries C_1 and C_2, respectively, and means f_1 and f_2 and standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$ for women. Applying the standard formula (e.g., equation (1.21) in [16]) for the moments of finite-weighted mixtures of distributions, the variance of the men's scores in the overall population is given by $(2(\sigma_1^2 + \sigma_2^2) + (m_1 - m_2)^2)/4$ and that of women is $(2(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) + (f_1 - f_2)^2)/4$. Letting VR denote the variance ratio in the overall population and VR_i the variance ratios in C_i, $i = 1, 2$, it follows immediately that*

$$VR_1 = \frac{\sigma_1^2}{\hat{\sigma}_1^2}, VR_2 = \frac{\sigma_2^2}{\hat{\sigma}_2^2}, VR = \frac{2(\sigma_1^2 + \sigma_2^2) + (m_1 - m_2)^2}{2(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) + (f_1 - f_2)^2}. \tag{1}$$

For example, if $m_1 = m_2 = 173, \sigma_1^2 = 5, \sigma_2^2 = 1, f_1 = f_2 = 172, \hat{\sigma}_1^2 = 1, \hat{\sigma}_2^2 = 2$, then Equation (1) implies that $VR_1 = 5, VR_2 = 0.5$, and $VR = 2$, so there is greater male variability in the overall population, but greater female variability in C_2. This contradicts claim (C).

Fortunately, we can turn to real data to pursue this example. Men have a higher variance in height than women (as a species) yet women have more variance in some countries. As with all conclusions based on real data, of course, the possibility of significant sampling errors must always be taken into account. (Human height is one of science's most studied and documented measurements and has been recorded and analyzed in great detail, over time and geographic location, in part because height is easy to measure and is an indicator of important factors such as nutrition and genetics.)

Example 3. *According to Roser et al. ([17]), the mean height of men worldwide is 178.4 cm with a standard deviation of 7.59 cm, while the mean height of women is 164.7 cm with a standard deviation of 7.07 cm. The variance ratio for adult human height worldwide is, therefore, $VR > 1.07$, implying greater male variance for human height. The variance ratios by country and birth year, on the other hand, range from less than 0.5 to greater than 2.5. Thus, there is greater variance worldwide in the heights of men than the heights of women, even though the variance ratios for height vary significantly among countries, some of which exhibit greater female variance. This contradicts claim (C).*

Note that the raw values of a statistic or collection of statistics indicate nothing about their causes. For example, the collection of estimated values of human heights in various countries says nothing about the relative importance of genetic, cultural, or nutritional factors. The heights are simply a record of observations, and further implications require further assumptions.

4. Correlation and Causation

As Holland observed in his classic text *Statistics and Causal Inference* [18], "Problems involving causal inference have dogged at the heels of statistics since its earliest days", and recent research involving gender differences in variability is certainly no exception. These studies are generally not balanced, i.e., they either conclude that their data support greater male variance for a given species and trait, or they conclude that the data do not support greater male variance for that trait and species. Although greater female variance has been found in some cases (see [19] Appendix A), the vast majority of conclusions invariably state that greater male variance holds or does not hold.

Whether one gender is more or less variable than the other (in the strict variance ratio sense or some other measure) is often, in itself, less interesting than the conclusions usually associated with it. For example, Kane and Mertz, in the first paragraph of their study, state that if a greater male variance were true, then "it could account for the fact that all Fields medalists have been male" ([3] p. 10). Thus, tail ratios are primarily their main object of interest.

Unless the distributions are assumed to be normal, however, this violates the basic fact *F* since this is a conclusion about tail ratios. On the other hand, if the distributions are assumed to be normal, then by Proposition 1, the finding of VR = 1.08 implies that right tails, such as Fields medalists, should indeed be strongly male-dominated.

Kane and Mertz then conclude that the non-uniformity in variance ratios that they found is largely an artifact of "a complex variety of sociocultural factors rather than intrinsic differences", i.e., cultural factors as opposed to "innate, biologically determined differences between the sexes" ([3] pp. 10–11). This same secondary causal inference conclusion is repeated in the AAAS *Science* article, namely that "cross-cultural analysis seems to rule out several causal candidates, including coeducational schools, low standards of living, and innate variability among boys" [9]; the *Scientific American* review of this study quotes Mertz as arguing that "The vast majority of the differences between male and female performance must reflect social and cultural factors".

As illustrated in Example 3 above, however, the raw values of a statistic or collection of statistics indicate nothing about their causes. In the Kane and Mertz study, however, they did suggest that other factors played a role, namely "The finding that males' variance exceeds females' in some countries but is less than females' in others and that both range all over the place suggests it can't be biologically innate, unless you want to say that human genetics is different in different countries" ([20] p. 4).

However, evidence shows that human genetics *do* differ within and among countries, even within the continent of Europe [21]. In ([3] Table 2), Taiwan and Tunisia have extreme variance ratios in the 2007 TIMSS scores for eighth graders, namely, 1.31 for Taiwan and 0.91 for Tunisia. Whether the significantly different variance ratio values for these two countries are primarily artifacts of sociocultural factors, rather than, say, a more balanced combination of sociocultural and innate biological factors, is a matter for further study; it does not follow from computing variance ratios. Yet based on the Kane and Mertz study, *Scientific American* concludes "Now that the greater male variability hypothesis has fallen short, nature is not looking as important as scientists once thought" ([20] p. 5).

Fortunately, this line of reasoning in greater male variance studies is not ubiquitous. For example, Taylor and Barbot report "In light of evidence that gender differences in creative variability are inconsistent across domains and tasks, broad claims about the causes and consequences of gender differences in creativity based on GMV should be avoided" ([7] p. 8).

In short, variance ratio values alone imply nothing about over- or under-representation of males or females for any given trait, and nothing useful about the relative importance of biological or cultural factors. Variance ratios simply compare variances.

5. Design of Experiments

A recurring methodological error in research on gender differences in variability is to base conclusions about adult members of a species, especially humans, on observations of younger members of the same species. As pointed out in [22], the extent of sex differences may depend on normal maturation and socializing influences as well as on genetics, so considerable time may be required before significant differences emerge.

In their paper, Kane and Mertz [3] tested the greater male variability hypothesis with respect to mathematics performances in humans, specifically referring to both Fields medalists and women in technical, management, and government positions. That is, their experiments were designed to draw conclusions about adult humans, not infants or children. In the description of their method, however, the authors clearly state that most of their measures of mathematics performances are based on mathematics assessments of fourth-graders and eighth-graders from numerous countries ([3] p. 11). Thus their conclusions about gender differences in adult humans are based on tests of pre- and early-adolescent children, not adults. Similarly, as Block observed, 75% of the studies reported in the highly influential text *The Psychology of Sex Differences* [23] are based on research

participants who were 12 years old or younger, and almost 40% employed preschool children ([22] pp. 289–290).

However, it has been established that, at those ages, boys and girls follow different developmental trajectories in many basic traits, both physical (e.g., height [17]) and cognitive (e.g., school performance [24]). For example, Arden and Plomin [5] addressed questions about the over- and under-representation of boys and girls at the low and high extremes of measures of cognitive abilities by studying sex differences in the variance of test scores across childhood. Among other conclusions, they found that "From age 2 to age 4, girls in our study were highly significantly over-represented in the top tail" ([5] p. 44). Thus, employing the same logic and methodology as in [3] to extrapolate data from tests conducted on children from age 2 to 4 to conclusions about adults, this finding by Arden and Plomin would imply that among adults, women are highly significantly over-represented in the top tail of intelligence. Similarly, extrapolating predicted heights of adults from data on fourth- and eighth-graders would conclude that adult women are generally taller than men.

To draw reasonable inferences about gender differences in variability (or any other traits) among human adults from tests on children, therefore, requires serious formal justification.

6. Discussion

If the goal of the research is to conclude something about the tail ratios, such as different proportions of males and females among top-level mathematicians, as in [3], then without knowledge of the underlying distributions, the variance ratio value alone says nothing. Similarly, a comparison of variance ratio values alone, e.g., for different countries, does not imply any answer to the standard nature vs. nurture debate. It is not the goal of the present article to argue the validity or invalidity of greater male variance or greater female variance in general or with respect to cognitive or mathematical abilities in humans, but simply to correct the scientific record concerning a series of faulty logical arguments being propagated in the scientific literature.

Perhaps Darwin was correct, and that there appears to be generally more variation among males than among females with regard to many different traits. For example, using three different measures for differences in variability (Levene's test, the variance ratio test, and the distances between cumulative distribution functions), Lehre et al concluded the following:

> "The data presented here show that human greater male intrasex variability is not limited to intelligence test scores, and suggest that generally greater intrasex variability among males is a fundamental aspect of the differences between sexes" ([6] pp. 220–221).

However, perhaps Darwin was wrong. For a survey of studies published since 2000, both supporting and not supporting greater male variance, the reader is encouraged to look at ([19] Appendix A).

As emphasized in a similar report that was critical of research in this field: "where findings appear especially inconsistent [this] should motivate and direct investigations toward fruitful new studies" ([22] p. 307). Similarly, we agree with the conclusion of [2] that "we do not mean to suggest that the quoted studies are completely uninformative. The data collected by these authors make a useful contribution to the literature, and could perhaps be reanalyzed in ways that avoid some of the problems with the original analysis". We hope that this article will help shed light on the interesting differences between men and women.

References

1. Del Giudice, M. Measuring Sex Differences and Similarities. In *Gender and Sexuality Development, Focus on Sexuality Research*; VanderLaan, D.P., Wong, W.I., Eds.; Springer: New York, NY, USA, 2022; pp. 1–38. [CrossRef]
2. Del Giudice, M.; Gangestad, S.W. No Evidence Against the Greater Male Variability Hypothesis: A Commentary on Harrison et al.'s (2022) Meta-Analysis of Animal Personality. *Evol. Psychol. Sci.* **2023**. [CrossRef]
3. Kane, J.M.; Mertz, J.E. Debunking Myths about Gender and Mathematics Performance. *Not. AMS* **2012**, *59*, 10–21. [CrossRef]
4. Levene, H. Robust Tests for quality of Variances. In *Contributions to Probability and Statistics*; Essays in Honor of Harold Hotelling; Olkin, I., Ed.; Stanford University Press: Stanford, CA, USA, 1960; pp. 278–292.
5. Arden, R.; Plomin, R. Sex differences in variance of intelligence across childhood. *Personal. Individ. Differ.* **2006**, *41*, 39–48. [CrossRef]
6. Lehre, A.C.; Lehre, K.; Laake, P.; Danbolt, N. Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Dev. Psychobiol.* **2009**, *51*, 198–206. [CrossRef] [PubMed]
7. Taylor, C.L.; Barbot, B. Gender differences in creativity: Examining the greater male variability hypothesis in different domains and tasks. *Personal. Individ. Differ.* **2021**, *174*, 110661. [CrossRef]
8. Hyde, J.S.; Mertz, J.E. Gender, culture, and mathematics performance. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 8801–8807. [CrossRef] [PubMed]
9. Kaufman, R. It Doesn't Add Up. *Science* **2011**. [CrossRef]
10. Reilly, D.; Neumann, D.L.; Andrews, G. Sex differences in mathematics and science achievement: A meta-analysis of National assessment of Educational Progress assessments. *J. Educ. Psychol.* **2015**, *107*, 645–662. [CrossRef]
11. Voracek, M.; Mohr, E.; Hagmann, M. On the Importance of Tail Ratios for Psychological Science. *Psychol. Rep.* **2013**, *112*, 872–886. [CrossRef] [PubMed]
12. Hill, T.P.; Fox, R.F. Extreme Tail Ratios and Overrepresentation among Subpopulations with Normal Distributions. *Stats* **2022**, *5*, 977–984. [CrossRef]
13. Kobrin, J.L.; Sathy, V.; Shaw, E.J. *A Historical View of Subgroup Performance Differences on the SAT Reasoning Test*; Technical Report No. 2006-5; College Board: New York, NY, USA, 2006. Available online: https://files.eric.ed.gov/fulltext/ED562569.pdf (accessed on 17 April 2023).
14. Feingold, A. The additive effects of differences in central tendency and variability are important in comparisons between groups. *Am. Psychol.* **1995**, *50*, 5–13. [CrossRef]
15. Dilmaghani, M. The gender gap in competitive chess across countries: Commanding queens in command economies. *J. Comp. Econ.* **2021**, *49*, 425–441. [CrossRef]
16. Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer: Berlin/Heidelberg, Germany, 2006; Chaper 1.2.4; pp. 10–11.
17. Roser, M.; Appel, C.; Ritchie, H. Human Height. Our World in Data. 2019. Available online: https://ourworldindata.org/human-height (accessed on 17 April 2023).
18. Holland, P.W. Statistics and Causal Inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945–960. [CrossRef]
19. Hill, T.P. An Evolutionary Theory for the Variability Hypothesis. *arXiv* **2021**, arXiv:1703.04184v14.
20. Begley, S. Anything Boys Can Do... *Sci. Am.* **2012**, *306*, 25. Available online: https://www.scientificamerican.com/article/anything-boys-can-do/ (accessed on 17 April 2023).
21. Moskvina, V.; Smith, M.; Ivanov, D.; Blackwood, D.; Clair, D.S.; Hultman, C.; Toncheva, D.; Gill, M.; Corvin, A.; O'Dushlaine, C.; et al. Genetic Differences between Five European Populations. *Hum. Hered.* **2010**, *70*, 141–149. [CrossRef] [PubMed]
22. Block, J.H. Issues, Problems, and Pitfalls in Assessing Sex Differences: A Critical Review of "The Psychology of Sex Differences". *Merrill-Palmer Q. Behav. Dev.* **1976**, *22*, 283–308.
23. Maccoby, E.E.; Jacklin, C.N. *The Psychology of Sex Differences*; Stanford University Press: Stanford, CA, USA, 1974.
24. Buczylowska, D.; Ronniger, P.; Melzer, J.; Petermann, F. Sex Similarities and Differences in Intelligence in Children Aged Two to Eight: Analysis of SON-R 2Ð8 Scores. *J. Intell.* **2019**, *7*, 11. [CrossRef] [PubMed]