UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Naïve Statistics: Intuitive Analysis of Variance

Permalink

https://escholarship.org/uc/item/4sv5n2v4

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 30(30)

ISSN

1069-7977

Authors

Trumpower, David L.
Fellus, Olga

Publication Date

2008

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Naïve Statistics: Intuitive Analysis of Variance

David L. Trumpower (david.trumpower@uottawa.ca)
Olga Fellus (ofell070@uottawa.ca)
University of Ottawa, Faculty of Education, 145 Jean-Jacques-Lussier Street
Ottawa, ON K1N 6N5 Canada

Abstract

In the present study, we explored the ability of statistics-naïve students to perform an intuitive analysis of variance and compare their performance with that of more experienced statistics students. Participants were shown several sets of data that varied with respect to within group variability and/or between group variability. They were asked to rate the strength of evidence provided by each dataset in support of a hypothetical theory. Results indicate that statistics-naïve students are able to perform a rudimentary form of analysis of variance with some accuracy, demonstrating at least a partial understanding of the importance of both within and between group variability. In one instance, statistics-naïve students actually performed in a more expert-like manner than did statistics-experienced students. However, statistics-naïve students also displayed a tendency to overweigh the relative importance of evidence provided by between group variability. This tendency persisted in statistics-experienced students.

Keywords: naïve statistics; statistical understanding; intuitive knowledge; conceptual knowledge; expert-novice differences

Introduction

Statistics is a notoriously difficult course for many students. It has been estimated that as many as 67-80% of graduate students experience uncomfortable levels of statistics anxiety (Onwuegbuzie & Wilson, 2003). One reason may be that they lack an intuitive conceptual understanding of statistics, or worse have *mis*conceptions. Alternatively, it is possible that students do possess an intuitive understanding of statistical principles, but have difficulty articulating and applying it in class.

It has long been recognized that humans must perform some sorts of intuitive statistical analyses (e.g., mean and variance estimation, correlation detection) in order to function in the uncertain environment of our natural world. On a daily basis we test hypotheses about group membership and, more generally, about relationships – e.g., “Is this person trustworthy or not?”, “Does the amount of sleep that I get affect my performance at work?”, etc. Peterson and Beach (1967), in a review of the literature on intuitive estimation of descriptive statistics, were impressed by the accuracy of such estimates. More recently, Kareev (2000) has even suggested that the limits of human working memory capacity are optimized for allowing detection of correlations.

Despite the ability of individuals to estimate statistics, these estimates are sometimes subject to certain biases. Peterson and Beach (1967) noted that estimates are often conservative, not making use of all available data. For example, subjects’ estimates of the mean of skewed distributions are typically biased toward the median. Similarly, Kareev, Arnon, and Horwiz-Zeliger (2002) have shown that estimates of variance tend to be low. It is well known that we are often biased toward making use of the most salient data when making judgements about averages (i.e., the *availability heuristic*; Kahneman & Tversky, 1973) and correlations (Chapman & Chapman, 1969). It is also well known that we tend to make use of data which confirms our preconceived hypotheses about such things as correlations between variables (i.e., confirmation bias, Wason & Johnson-Laird, 1972; for a review, see Klayman & Ha, 1987).

If students do hold certain intuitive biases or misconceptions which affect statistical estimation and inference, then instruction in statistics may be impeded. In physics, another difficult subject for many, students have been shown to possess faulty intuition. Furthermore, these naïve misconceptions are highly resistant to change. McCloskey, Caramazza, and Green (1980) have demonstrated that a large proportion of students that had just completed a university course in physics maintained their false belief that a bullet fired from a gun with a curved barrel would travel in a curved path upon exiting the gun. This is just one of many intuitive misconceptions that students hold concerning the physical world and that persist despite instruction to the contrary (see e.g., Halloun & Hestenes, 1985).

Although much has been written about informal, intuitive detection of statistical properties in the natural world, very few have discussed transfer of intuitive statistical reasoning to the more formal statistical computation required in the classroom. Our hope is that identification of intuitive successes in statistics will allow us to alleviate statistics anxiety, while identification of intuitive failures will allow for more focused instruction to better overcome such biases or misconceptions.

In the present study, we explore the ability of statistics-naïve students to perform an intuitive analysis of variance and compare their performance with that of more experienced statistics students. Intuitive successes and failures are identified and discussed.

Method

Participants

Twenty students enrolled in courses within the Faculty of Education at the University of Ottawa volunteered to serve as participants. Thirteen were graduate students who had just completed an introductory-level course in statistics (henceforth referred to as *Experienced*), whereas the other seven were statistics-naïve undergraduate students who had just completed a teacher education course in curriculum design (henceforth referred to as *Naïve*).

Materials & Procedure

Both the Experienced and Naïve participants were given a worksheet with the following cover story and instructions on the first page:

> *Suppose two scientists/entrepreneurs are considering whether or not to develop a golf ball freezer that can be attached to a regular golf bag. They have a theory that frozen golf balls travel farther than normal (i.e., unfrozen) golf balls. To test their theory, the scientists/entrepreneurs devise an experiment in which a robotic arm will be used to hit normal and frozen golf balls, all with the exact same force, after which the distance that each ball travels will be measured. In order to remain completely unbiased, the scientists/entrepreneurs will allow independent researchers (who are completely unaware of their theory) to conduct the experiment.*
>
> *Listed on the following page are hypothetical results from several such experiments. For each experiment, rate the amount of support (1=weak, 10=strong) that you think the test would provide for the claim that frozen golf balls go farther than normal golf balls, and briefly explain why.*

On the second page of the worksheet, participants were shown four hypothetical datasets, each listing the distances traveled by 3 frozen and 3 unfrozen golf balls. Beside each dataset was a 10-point rating scale as well as blank space in which participants could write explanations for their ratings.

Hypothetical datasets varied with respect to between and/or within group variability, as summarized in Table 1. Because each dataset was comprised of the same number of scores, the F-statistic comparing the distances traveled by the frozen and unfrozen golf balls can be considered a measure of the relative amount of evidence that each dataset provides against the null hypothesis that there is no difference between the distance traveled by frozen and unfrozen golf balls. It should be noted that participants were shown only the raw scores in the hypothetical datasets. Means, standard deviations, and F-statistics are provided here for illustrative purposes only.

Order of presentation of the four datasets was counterbalanced across participants, separately for both Naïve and Experienced. Participants were allowed as much time as necessary to make and explain their ratings, but most completed the task within approximately 15 minutes.

Table 1: Means (and standard deviations) of conditions, and resulting F-statistics, of hypothetical datasets.

Dataset	Unfrozen	Frozen	F
1	300 (50)	304 (50)	.01
2	300 (50)	400 (50)	6
3	300 (1)	304 (1)	24
4	300 (1)	400 (1)	15000

Results

Participants' ratings were analyzed using a 2 Expertise (Naïve, Experienced) x 4 Dataset split-plot ANOVA with repeated measures on the second factor. Both main effects of Expertise, F(1,18)=7.95, p=.011, and Dataset, F(3,54)=44.44, p<.001, were significant, but were qualified by a significant Expertise by Dataset interaction, F(3,54)=2.77, p=.05 (see Figure 1).

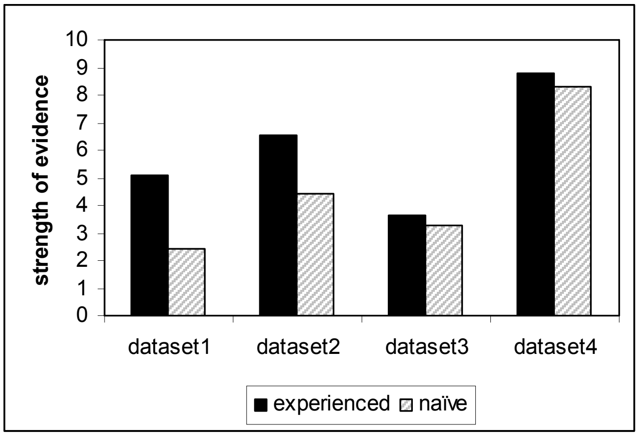

Figure 1. Mean ratings by Expertise and Dataset.

Next, all possible 2 Expertise x 2 Dataset interaction contrasts (and if non-significant, simple pairwise contrasts collapsed across Naïve and Experienced participants) were conducted. The Scheffe procedure was used to determine the critical values used for testing all contrasts (Maxwell & Delaney, 2003).

It can be seen in Table 1 that Datasets 1 and 2 differ only with respect to the magnitude of the difference between the means of the distances traveled by the unfrozen and frozen golf balls. This is also true for Datasets 3 and 4. Thus, comparisons of the mean ratings for these datasets indicate participants' ability to detect and appreciate the importance of differences in means. Neither the interaction contrast comparing Experienced and Naïve participants' ratings of Datasets 1 and 2, F(1,18)<1, nor Datasets 3 and 4, F(1,18)<1, was significant. Both Experienced and Naïve participants accurately rated Dataset 2 as providing stronger evidence than Dataset 1, F(1,18)=14.60, p=.001, and Dataset 4 as providing stronger evidence than Dataset 3 F(1,18)=84.70, p<.001.

Datasets 2 and 4 differ only with respect to the magnitude of within-group variability. The standard deviations of the distances traveled by the unfrozen and frozen golf balls was smaller in Dataset 4 than in Dataset 2. Likewise, Datasets 1 and 3 differ only with respect to within-group variability – standard deviations are smaller in Dataset 3. Thus, comparisons of the mean ratings for these datasets indicate participants' ability to detect and appreciate the importance of within-group variability. The interaction contrast comparing Experienced and Naïve participants' ratings of Dataset 2 and Dataset 4 was not significant, F(1,18)=3.32, p=.085. Both Experienced and Naïve participants accurately rated Dataset 4 as providing stronger evidence than Dataset 2 (F(1,18)=46.58, p<.001). The interaction contrast comparing Experienced and Naïve participants' ratings of Dataset 1 and Dataset 3 was, however, significant, F(1,18)=8.79, p=.008. Naïve participants accurately rated Dataset 3 as providing stronger evidence than Dataset 1, but Experienced participants rated Dataset 3 as providing *weaker* evidence than Dataset 1. Although both Naïve and Experienced participants show some ability to detect and appreciate the importance of within group variability, the Naïve participants appeared more consistent in their application of this ability.

Datasets 1 and 4 differed with respect to both within and between group variability. The distributions displayed in Dataset 4 contained both smaller standard deviations and a larger difference between group means than those in Dataset 1. Therefore, Dataset 4 provided the most evidence in support of the scientist/entrepreneurs' theory whereas Dataset 1 provided the least. The interaction contrast comparing Experienced and Naïve participants' ratings of Datasets 1 and 4 was not significant, F(1,18)=3.26, p=.088. Not surprisingly, both Experienced and Naïve participants accurately rated Dataset 4 as providing stronger evidence than Dataset 1, F(1,18)=63.49, p<.001.

The final contrast involved Datasets 2 and 3. Whereas Dataset 2 displayed a larger difference between group means, it also displayed larger standard deviations than Dataset 3. Thus, in order to accurately determine the relative amount of evidence provided by these two datasets, one must combine the information provided by the within and between group variability in each dataset. A rough estimate of the strength of evidence can be determined via the ratio of between:within group variability (similar to that provided by the F-statistic). The difference in group means in Dataset 2 is 100 yards and the standard deviations are 50. This results in a 2:1 ratio of between:within group variability (more formally, an F-statistic equal to 6) for Dataset2. In Dataset 3, the difference in group means is 4 yards and the standard deviations are 1, resulting in a 4:1 ratio (or an F-statistic equal to 24). Hence, Dataset 3 provides stronger evidence than Dataset 2. While the interaction contrast comparing Experienced and Naïve participants' ratings of Dataset 2 and Dataset 3 was not significant, F(1,18)=3.25, p=.088, both Experienced and Naïve participants *inaccurately* rated Dataset 2 as providing stronger evidence than Dataset 3, F(1,18)=16.97, p=.001.

To summarize the results:

1. Both Naïve and Experienced participants were able to detect and appreciate the importance of between group variability. Both judged the datasets depicting larger between-group differences in means as providing stronger evidence.
2. Participants were also able to detect and appreciate the importance of within-group variability to some extent. Surprisingly, however, Naïve participants appeared more consistent in their application of this ability. Both Naïve and Experienced participants judged Dataset 4 as providing stronger evidence than Dataset 2, but only Naïve participants rated Dataset 3 as providing stronger evidence than Dataset 1.
3. Both Naïve and Experienced participants demonstrated some difficulty in combining between- and within-group variability to make strength of evidence ratings as indicated by lower ratings for Dataset 3 than Dataset 2.

Participants' written justifications for their ratings may help to explain the difficulties of both Naïve and Experienced participants exemplified in result #3 and the counterintuitive superiority of Naïve participants exemplified in result #2 above. These written justifications will be explored in the Discussion section which follows.

Discussion

It appears that statistics-naïve students do possess an intuitive understanding of some basic statistical concepts. They are able to perform a rudimentary analysis of variance on simple datasets. However, this study also shows that they hold at least one bias which runs counter to expert statistical reasoning. Furthermore, initial training in statistics may not eliminate this bias, and may actually create the potential for other misconceptions. Each of these major findings will be discussed separately, followed by a brief general discussion.

Between-group variability

Both statistics-naïve and statistics-experienced students were able to detect and appreciate the importance of between-group variability. Where two datasets differed only with respect to the size of the between-group difference in means, students rated the dataset with the larger difference in means as providing stronger evidence in support of an effect of an independent variable.

This ability to appreciate differences in means might seem trivial when such differences are obvious. In Datasets 3 and 4, all within-group standard deviations were 1 so that the distributions of frozen and unfrozen golf balls did not overlap. Thus, the 100 yard difference between the means of the frozen and unfrozen golf balls in Dataset 4 and the 4 yard difference between the means of frozen and unfrozen golf balls in Dataset 3 are both quite salient.

However, this ability of participants to appreciate differences in means was displayed even when the between-group differences might have been partially obscured by large within-group variance. For example, the difference in means in Dataset 1 was 4 yards in favor of the frozen golf balls, whereas the difference in Dataset 2 was 100 yards in favor of the frozen golf balls. But, in both datasets the within-group standard deviations were 50, resulting in some overlap of the distributions of frozen and unfrozen golf balls. Nonetheless, participants accurately rated Dataset 2 as providing stronger evidence than Dataset 1.

Within-group variability

Both statistics-naïve and statistics-experienced students were also able to detect and appreciate the importance of within-group variability, under certain circumstances. The difference between the mean distances traveled by the hypothetical frozen and unfrozen golf balls was 100 yards in both Datasets 2 and 4. But, the within-group standard deviations in Dataset 2 were 50 whereas they were just 1 in Dataset 4. Consequently, participants rated the dataset with less within-group variability (Dataset 4) as providing stronger evidence.

When the difference between group means was smaller, though, as was the case in Datasets 1 and 3, only the statistics-naïve participants rated the dataset with less within-group variability (Dataset 3) as providing stronger evidence.

Poorer performance of statistics-experienced students in this situation may be attributed to their failure to conceptualize the data as coming from an independent-groups research design. Inspection of participants' written justifications of their ratings reveals that this, indeed, was the case. Of the 13 Experienced participants, the written justifications of 10 indicated that they treated the data as resulting from a correlated groups design. These participants had subtracted the distances of individual frozen golf balls from the distances of the unfrozen golf balls (i.e., they computed difference scores), as would be done if one were to hand-compute a correlated groups t-test or sign test. They then justified their ratings by citing the magnitude of the difference scores (e.g., "…2+ differences of 50 yards…") and the number of differences in which the frozen ball traveled further (e.g., "…in 2/3 trials frozen > normal…". Because the distances in Dataset 3 contained small between-group differences, as well as small within-group variability, computation of difference scores resulted in three small differences all in favor of the frozen balls. As was indicated by several participants' justifications, though, the reliability of small but consistent differences were questioned (e.g., "…unclear if 3 positive differences are significant enough…". Computation of difference scores in Dataset 1, in which the distances contained much greater within group variability, resulted in two large differences in favor of the frozen balls and one large difference in favor of an unfrozen ball. For Experienced participants, two large differences in favor of frozen balls despite one even larger difference in favor of an unfrozen ball were perceived as providing stronger evidence than three small differences in favor of the frozen balls. Apparently, large but inconsistent differences trumped small but consistent differences.

Computing difference scores within the other datasets results in three large differences in favor of frozen balls for Dataset 4, and two large differences in favor of frozen balls with one "tie" for Dataset 2. If one only considers large positive (i.e., in favor of frozen balls traveling farther) differences as acceptable evidence, while largely ignoring small or negative differences, then the strength of evidence provided by Datasets 1-4 is consistent with the Experienced participants' ratings. That participants would ignore large negative difference scores that seemingly disconfirm the hypothetical scientist/entrepreneurs' theory is not entirely surprising considering the well documented confirmation bias (Klayman & Ha, 1987).

Potentially the presentation format of the datasets might have contributed to the misconception that data derived from a correlated groups design. Data were presented in two side-by-side columns. It should be noted, however, that only one of the seven Naïve participants displayed any indication of computing difference scores in their written justifications. Thus, it was not the format per se, but the format coupled with the Experienced participants recent exposure to similar looking datasets that were analyzed with correlated groups techniques (i.e., correlated groups t-test and sign test) in their statistics course, that may have caused the confusion. Future studies are planned in which presentation format is altered in an effort to eliminate this misconception.

Ratio of between:within group variability

Although statistics-naïve students, and to a lesser extent statistics-experienced students, appeared able to detect and appreciate the importance of both between- and within-group variability, they were not entirely successful in *combining* such information. Datasets 2 and 3 differed in two ways. The difference in the mean distances of the frozen and unfrozen balls was larger but the standard deviations of the distances were also larger in Dataset 2 than in Dataset 3. So, while the larger between-group variability in Dataset 2 might suggest that it provides stronger evidence, the larger within-group variability might suggest that it provides *weaker* evidence, than Dataset 3. In order to accurately discriminate between the strength of evidence provided by these two datasets one must consider the magnitude of the between-group variability *relative to* the within-group variability (i.e., "error"). Although the ratio of between:within group variability is larger in Dataset 3, both statistics-naïve and statistics-experienced students believed that Dataset 2 provides stronger evidence. Students did not combine their intuitive knowledge about between- and within-group variability in an appropriate manner.

Perhaps the most plausible explanation for this discrepancy between the actual and perceived strengths of evidence provided by these datasets is that students have a tendency to place more weight on between-than within-group variability. As was discussed earlier, this is almost certainly true for statistics-experienced students. Experienced participants gave higher ratings to datasets with large but variable difference scores than datasets with small but consistent difference scores. Naïve participants, however, did not compute difference scores. Rather, their written justifications indicate that they estimated group means and then compared the difference of means (e.g., "…average basically the same: 301 vs. 304…). Regardless of whether one estimates between-group variability by computing difference scores or by estimating and then comparing group means, Datasets 2 and 4 have greater between-group variability than Datasets 1 and 3, which is consistent with students' ratings.

This is not to say that students completely ignore within-group variability. Where two datasets had the same between-group variability, Naïve participants rated the one with less within group variability as providing stronger evidence.

General discussion

The present study is our initial attempt to document and explain students' naïve conceptual understanding of analysis of variance. We have shown that statistics-naïve students understand, at some level, the importance of both between- and within-group variability. However, we have also shown that statistics-naïve students place a greater importance on between-group variability. Finally, we have shown that this strong focus on between-group variability persists even after students are exposed to an introductory course in statistics.

The strong focus on between-group variability may be understandable considering that we are typically exposed to the concept of an *average* well before entering a statistics course. Also, a large difference between the distances of frozen and unfrozen golf balls, whether computed as mean difference scores or as a difference between group means, is likely to be viewed as confirming evidence for theories about independent variable effects, whereas large within-group variability is more likely to be viewed as potentially disconfirming evidence. If operating with a confirmation bias, then again the strong focus on between-group variability is understandable. This assumption is consistent with Friederich's (1993) Primary Error Detection and Minimization (PEDMIN) interpretation of confirmation bias.

Our hope is that by identifying students' naïve abilities and biases we can improve instruction. Making students aware of their accurate naïve knowledge may help to alleviate statistics anxiety. And, making instructors aware of pre-existing biases may help target instruction toward these more difficult areas. If so, then perhaps statistics will no longer be sadistic.

References

Chapman, L.J., & Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74 (3),* 271-280.

Friedrich, J. (1993). Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena. *Psychological Review 100(2),* 298-319.

Halloun, I.A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics, 53,* 1056-1065.

Kahneman, D., & Taversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237-251.

Kareev, Y. (2000). Seven (indeed plus minus two) and the detection of correlations. *Psychological Review*, *107,* 397-402.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General, 131(2),* 287-297.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211-228.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science, 210,* 1139-1141.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

Onwuegbuzie, A.J. & Wilson, V.A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments—a comprehensive review of the literature. *Teaching in Higher Education, 8(2),* 195–209.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68 (1),* 29-46.

Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning: Structure and content*. London: Batsford.