Combining Estimates of Epistemic and Aleatory Uncertainty to Reduce Overprecision

Daniel G. Goldstein, Microsoft Research, USA
David Rothschild, Microsoft Research, USA

Expectations of prices and quality are critical in consumers’ decisions, and should be shaped by the distributional properties of the outcome and the uncertainty in a person’s estimate of the average. We show that people do not naturally focus on all these sources, leading to excessively narrow confidence intervals.

[to cite]:

[url]:
http://www.acrwebsite.org/volumes/1024553/volumes/v45/NA-45

[copyright notice]:
This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
Advances in Consumer Research
Volume 45, ©2017

Challenging Intuitions on Intuitive Statistics

Chairs: Quentin André, INSEAD Europe Campus, France
Nicholas Reinholtz, University of Colorado Boulder, USA
Bart de Langhe, University of Colorado Boulder, USA

Paper #1: High Chances and Close Margins: How Different Forecast Formats Shape Beliefs
Oleg Urminsky, University of Chicago, USA
Sondre Skarsten, University of Chicago, USA
Luxi Shen, The Chinese University of Hong-Kong, China

In this research, we contrast two common ways of communicating an uncertain forecast, either as a chance (e.g., the probability of winning a game) or as an expected margin (e.g., the point spread or predicted amount by which a game would be won). We find a robust chance-margin discrepancy, in which people tend to treat a chance forecast as conveying greater certainty about the directional outcome than the statistically equivalent margin forecast. This discrepancy is not explained by other accounts, such as anchoring or neglect of information, and has consequences for subsequent judgments.

Across five experiments, we tested this chance-margin discrepancy in voting, sports and statistical reasoning and with various judgment and decision variables. In Study 1 (N=126), we use forecasts from fivethirtyeight.com about the 2016 Presidential election, which were updated regularly based on aggregating public election polls, and which were communicated both in terms of the chances of each major candidate winning the election, as well as the expected vote margin. Participants were either presented with the site’s forecast for chances and estimated the margin forecasts, or saw the margin forecasts and estimated the chance forecasts. When participants were shown the chance forecast (86.4% chance of Clinton winning) they over-estimated the forecast margin (64.7% vs. 54.2%, t=3.87, p<.001). Conversely, participants who were shown the margin forecast (54.2%) underestimated the chance forecast (62.8% vs. 86.4%, t=11.46, p<.001). This pattern was replicated in each of five other such studies conducted during the pre-election period, as the outcome forecasts varied (all ps <.001).

In Study 2 (N=143), we ruled out pre-existing beliefs and attention to the forecasts as alternative explanations. Participants
saw either chance or share forecasts for the election in four unspecified states. We replicated over-estimation of share forecasts when shown chance forecasts and under-estimation of chance forecasts when shown share forecasts in each state. Furthermore, participant estimates were sensitive to the state-specific forecast information. In a follow-up study, we further address an anchoring account, by demonstrating that estimates based on forecasts were significantly different from (and less accurate than) estimates based on equivalent non-forecast anchors.

In Study 3 (N=220), we tested the discrepancy between beliefs based on chance forecasts and margin forecasts in a new domain. First, we used a similar design with forecasts of football game outcomes, both for forecasted chance of winning and margin of winning (e.g., point spread). Participants who were shown the point-spread significantly under-estimated the forecasted chance of winning (62% vs. 71%, t=8.4, p<.001). Participants who were shown the chances of winning only directionally over-estimated the point-spread (6.7 vs. 6.5, t=1.31, p=.19).

Study 4 (N=398) tested the effects of forecast framing using objective statistical information, and identified an important boundary condition. We presented participants with a statistical scenario involving more or less informative sample of colored marbles drawn from one of two jars, for which we could calculate the objectively correct answers. When the sample was larger (and therefore more informative) participants showed the margin in the sample underestimated the chance of the marbles coming from the more likely jar (60.7% vs. 78%, t=15.1, p<.001). Conversely, participants told the chances of the sample coming from the more likely jar overestimated the corresponding proportion in the sample (66.3% vs. 60%, t=6.57, p<.001). However, when the sample was small (and therefore less informative), both share and chance information yielded underestimation.

In the last study, we tested the consequences of forecast framing. In Study 5 (N=198), participants were shown the change over the prior ten days in fivethirtyeight.com’s forecasts, either in terms of chances (from 85% Clinton/15% Trump to 69% Clinton/31% Trump) or in terms of margin (from 53% Clinton/47% Trump to 52% Clinton/48% Trump). Participants then answered several questions about their perceptions of the election, support for the candidates and behavioral intentions.

Participants who saw the change in chance forecasts (vs. change in margin forecasts) gave more extreme assessments of the change in the election (t=3.25, p=.001). Participants were more likely to judge the change in chance forecasts as “very good news” or “very bad news” (45%) than the equivalent change in margin forecasts (25.5%). The degree to which participants were worried about the election was unaffected. Participants’ own intention to vote was high (77% “very likely”) and unaffected by the forecast format. However, participants who were shown the change in chance forecasts were significantly more likely to say they would remind a friend or neighbor to vote (50% vs. 31.6% “very likely”, M=3.8 vs. 3.2, t=2.71, p=.007).

Overall, our results demonstrate that forecasts are seen as more extreme when framed as predicting chances rather than as predicting margins, even when the two formats are used to communicate equivalent information. Given that forecasts are often communicated in only one format, these findings suggest that the supposedly irrelevant choice of format for a forecast can have a remarkably meaningful impact on the interpretation of the forecast and even change one’s attitude and behavioral intention.

---

**EXTENDED ABSTRACT**

It’s Absolutely Relative: On the Processing of Absolute and Relative Differences

Consider a consumer deciding between a 40 inch TV and a 50 inch TV. When evaluating the difference in screen size the consumer can consider the absolute (i.e., 10 inches) or the relative (i.e., 25% larger) difference. Predominant decision theories assume that the marginal utility/value for additional inches of TV is a function of the absolute differences in size (Kahneman and Tversky 1979; von Neumann and Morgenstern 1947).

Although absolute and relative dimensions both reflect differences, they are not perfectly correlated, and consumers can as such use either or both dimensions when making decisions. For instance, prior research on proportion dominance has demonstrated sensitivity to relative differences. In the iconic jacket-calculator problem, 68% of participants were willing to drive 20 minutes to save $5 on a $15 calculator, but only 29% were willing to drive to save $5 on a $125 jacket (Tversky and Kahneman 1981).

The observation that consumers are sensitive to both absolute and relative differences is well documented (Krishna et al. 2002; Choi and Coulter 2009); however, disagreement exists regarding their respective influence. Within marketing and related literature, researchers have argued that processing relative differences is easier and more natural than processing absolute differences (Saini and Thota 2010; Azar 2007; Dehaene 2011; Choi and Coulter 2009). This line of reasoning suggests that when making decisions, consumer perceptions are informed by their Approximate Number System (ANS), the set of cognitive features that allow consumers to discriminate visual quantities (e.g., smaller vs. larger sets of dots). This system supposes that the mind only possesses a natural ability to process relative differences, and that representations of absolute differences must be constructed on the spot using this ANS. This system is what gives psychophysical functions their particular concave shapes (for a review see Dehaene 2011).

For consumer decisions that involve number-based judgments (e.g., was $100 now $90), researchers have made the opposite prediction, and suggested that processing relative differences is more cognitively taxing than processing absolute differences (Kruger and Vargas 2008; Chen and Rao 2007; Bettman, Johnson and Payne 1990). They argue that to understand that a sale price of $90 is 10% off of an original $100 price, the consumer has to first process the absolute difference ($10) then divide by the base price ($100). As such, the absolute difference is a necessary precondition for processing relative differences.

The current research aims to test the predictions of these two accounts and identify the cognitive antecedents of these two modes of processing. Whereas past research has studied these differences at the group level, the current research introduces a within-participant paradigm to assess the use of relative and absolute differences in different settings. In study 1, 301 participants were presented price promotions with relative (5%, 10%, 15%, 20%, 25% off) and absolute ($5, $10, $15, $20, $25 off) discounts manipulated orthogonally.

For each level of prices promotion, participants indicated the furthest they would be willing to drive (WTD), in miles, to receive that particular promotion. To assess the influence of relative and absolute differences, data were analyzed using a Bayesian hierarchical linear model (HLM), allowing for the estimation of sensitivities to relative and absolute differences for each participant.

In study 1, participants were randomly assigned to either the control (“was SA now SB”), % off (“was SA now SB, X% off”), or S-off conditions (“was SA now SB, $Y off”). Our results indicate that
participants in the control condition were more sensitive to absolute differences than relative differences ($b_{absolute} = 0.042$ vs. $b_{relative} = 0.009$, $p < .0001$). In the $5\%$-off condition, the use of absolute (relative) differences increased (decreased) ($b_{absolute} = 0.050$ and $b_{relative} = 0.006$) compared to the control condition ($p < .10$). This change was markedly larger in the $%\%$-off condition ($b_{absolute} = 0.030$ and $b_{relative} = 0.020$) compared to the control condition ($p < .0001$), suggesting that participants are more prone to utilizing absolute differences, even when the relative-difference information is explicitly provided.

Results of study 1 also indicated strong negative correlations in the fixed (rs -.60 to -.34) and random (rs -.47 to -.32) effect parameters. Importantly these effects were asymmetric: while a greater reliance on absolute differences limits the use of relative differences, greater use of relative differences has a much larger attenuation effect on absolute-difference usage. This suggests that relative differences require fewer resources, and can thus be used while processing absolute differences. On the other hand, absolute differences are more cognitively taxing and prevent the use of relative differences. Taken together, those results can explain the conflicting evidence regarding absolute and relative differences: the within-participant-level analysis indicates that while relative-difference processing is indeed more automatic than absolute-difference processing, consumers utilize absolute differences despite its more laborious processing.

To ensure generalizability, studies 2a ($N = 300$) and 2b ($N = 299$) replicated the findings of study 1 using product attributes instead of prices (study 2a) and by only manipulating relative or absolute differences (study 2b). Similar results were found. In study 3, we investigated whether our findings would hold true when assessing absolute differences. On the other hand, absolute differences are more easily processed by relative differences compared to numeric quantities. Furthermore, consistent with study 1, processing of relative differences inhibited processing of absolute differences, but not the other way around.

A fourth study on 202 participants replicated the conditions from study 1, with the critical difference that the final price was omitted in the $5\%$-off and the $5\%$-off conditions, so that the participants would have to compute it themselves. Opposite of study 1, results indicated a greater use of relative differences than absolute differences in both the $%\%$-off ($b_{absolute} = 0.015$ and $b_{relative} = 0.029$) and $%\%$-off ($b_{absolute} = 0.020$ and $b_{relative} = 0.031$) conditions, $p < .01$. Additionally, contrary to the previous studies, processing absolute differences inhibited relative-information processing and not the other way around. Taken together, results of studies 1-3 suggest that consistent with the numerical cognition argument, relative-information processing is more efficient, but consumers appear to utilize a different mode of processing quantitative differences when they must calculate the differences themselves (study 4).

**Variance Spillover in Intuitive Statistical Judgments**

EXTENDED ABSTRACT

For more than 50 years, researchers have been studying the intuitive statistical capabilities of humans (Peterson and Beach 1967). While some consensus has emerged, many unresolved questions remain. In this project, we address one of these questions, which has important marketing implications: Can consumers accurately encode and represent the distributional properties of multiple numerical distributions? Consumers frequently encounter numerical information (prices, qualities, quantities, etc.) and—to make optimal decisions—must remember and use this information. For example, when considering whether to buy a new computer, a consumer must assess the relative likelihood of finding a comparable computer at a better price with continued search. The solution to this problem relies on the variance of computer prices on the market (and the consumer’s search cost). Importantly, consumers make many decisions of this nature each year and thus—to make optimal decisions—must maintain mental representations of different price (and other numerical) distributions simultaneously.

When presented with a distribution of numerical values, previous investigations suggest that people are surprisingly accurate when making judgments about central tendency (e.g., the mean or median; Beach and Swenson 1966, Peterson and Miller 1964). However, there is less consensus regarding people’s ability to judge dispersion (e.g., variance). People seem biased towards using absolute, rather than squared deviation, when judging variance (Beach and Scopp 1968, Goldstein and Taleb 2007). Further, people seem to confound differences in coefficient of variation (variance divided by mean) with differences in variance (Hofstatter 1939, Lathrop 1967). Some recent research, however, is more optimistic. Goldstein and Rothschild (2014) found that people could reproduce distributions accurately from memory using a “distribution-builder”, in which they can “draw” the distribution. This research suggests that people’s mental representations of variance may be accurate and that previously identified biases may stem from the elicitation methods used.

The current investigation extends previous work to contexts in which people must learn about multiple numerical distributions at the same time: in the present research, distribution of prices for two different products. We believe this extension is particularly important for understanding consumer decision making. Consumers tend to learn about products over time through infrequent exposures to information and these exposures are often interspersed with exposures to information about other, different products. For example, you probably don’t know the exact price of a banana at your local supermarket, but you probably have a pretty good idea, and could probably identify whether the current price is a good deal or a bad deal. However, you have probably never directed any real attention to learning about the price of bananas; instead, you’ve been exposed to banana prices irregularly and infrequently. And in the interim, you have seen prices for many other products (apples, cereal boxes...). In our research, we ask how accurately can people learn about numerical distributions in this type of context, and contribute to the literature on internal reference price formation (Mazumdar, Raj, and Sinha 2005).

In three studies, we presented people with a sample of prices for two different products (e.g., red wine and white wine, within the products there was no differentiation other than price). Prices were displayed one-at-a-time, for several seconds each, over-imposed on a picture of the target product. Prices for the two products were intermixed. After seeing all of the prices, participants were asked to reproduce the price distributions for each product individually (i.e., a separate distribution for each product) using a distribution builder tool (André 2016; Goldstein and Rothschild 2014). We then assessed the accuracy of these expressed distributions with respect to the actual distributions the participants had seen.

In study 1, we examined participants ability to simultaneously learn the central tendency of two different price distributions: prices of red wine and prices of white wine. Within each product category, the variance of prices was held constant (SD = $6$), but the mean price of each category varied between participants. The mean price for one product category was $25$ in all conditions (the focal product), but we varied whether the mean price of the other category was either $10$ higher, $10$ lower, or the same (the distractor product). This design allowed us to determine whether prices from the distractor distribution influence the estimate of central tendency from the
focal distribution. Our analysis suggests they do not: The means of the (reproduced) focal distribution were not different between conditions and, further, were not different than the true mean ($25).

In studies 2 and 3, we instead examined participants ability to simultaneously learn the variance of two different price distributions. In study 2, we used a 2 x 2 between-participant design. In all conditions, the mean price of one product (e.g., red wine) was $23 and the mean price of the other product (e.g., white wine) was $28 (products were counterbalanced). We cross-manipulated the variances of the price distributions, such that each product was either high (SD = $9) or low (SD = $3). This design allowed us to test for “variance spillover”: whether participants’ impression of variance in prices for one category was influenced by the variance in prices for the other category. Our analysis suggests this is indeed the case: variances estimates were influenced by the true variance of the focal distribution, but also of the non-focal distribution. For example, when the true variance of the red wine price distribution was high, participants underestimated this variance when the variance of the white wine price distribution was low (vs. high).

In study 3, we examined a potential moderator of the variance spillover effect. This study largely reproduced study 2, but included an additional factor: whether the product categories were similar (red wine vs. white wine) or dissimilar (red wine vs. smartphone cases). We found that the variance spillover effect was attenuated, but not eliminated when the products were dissimilar. This suggests that categorization processes may underlie the observed result. In fact, the contamination of variance estimates may be adaptive from a Bayesian perspective—you can learn about the likely dispersion of one category by observing another similar category.

Combining Estimates of Epistemic and Aleatory Uncertainty to Reduce Overprecision

EXTENDED ABSTRACT

Overprecision, or excessive confidence that one knows the truth, is considered the most robust form of overconfidence (Moore, Tenney, Haran, 2016), and has been called “the most significant of the cognitive biases” (Kahneman, 2011). It is typically studied with the confidence-interval paradigm, in which people are asked to state two numbers that should bracket the true realization of an outcome with a given probability. Many prior demonstrations of overprecision show that people are miscalibrated: for example, stated 80% confidence intervals tend to contain the correct answers much less than 80% of the time.

Let us consider a person asked to construct a confidence interval for the price of gas at a given station in the USA. This person will face two sources of uncertainty. The first source is its epistemic uncertainty, which is tied to its ability to generate an accurate estimate about the average price of gas in the USA. Second, even if this person knew the average price of gas in the USA with absolute confidence, there would still remain some uncertainty about the price at this given station. This second type of uncertainty, called aleatory uncertainty, is tied to the standard deviation of the outcome: here, the typical amount by which a station deviates from the true average.

In the present research, we investigate the link between consumers’ perception of aleatory uncertainty, epistemic uncertainty, and overall uncertainty. In particular, we seek to uncover the sources of overprecision: is it a consequence of consumers’ inability to generate accurate estimates for aleatory and/or epistemic uncertainty? Or does it reflect a failure to incorporate both types of uncertainty when generating confidence intervals?

We develop a paradigm in which we ask people for three different kinds of confidence intervals. First, as done in past research on overprecision, we ask the standard confidence interval of what a particular outcome will be (such as the price of gas at a particular gas station, or the temperature on some future date in a particular city). Second, we ask participants to generate a measure of aleatory uncertainty, by asking them how much that value varies around the average (for example how much the price of gas around the country varies around the mean or how much the temperature in a particular city on a given day varies from year to year). Finally, we ask respondents to generate a measure of epistemic uncertainty, by asking them how far their estimate of the average is likely to be from the true average. Since the standard confidence interval should reflect both epistemic and aleatory, there is a direct theoretical relationship between first value (the standard confidence interval for the value), and the second and third values (variation around the true average and uncertainty in the location of the true average). As such, this paradigm allows us to investigate consumers’ accuracy in estimating aleatory and epistemic uncertainty, and their ability to combine both estimates into a single measure of confidence.

We observe, in line with previous research (e.g., Moore, Tenney, Haran, 2016), that people consistently create too-narrow confidence intervals: elicited 80% intervals contained the correct answer less than 80% of the time. We also document several important findings. First of all, we find that people have reasonable estimates of both aleatory uncertainty in the world and epistemic uncertainty in their minds. Second, the stated confidence intervals exhibit a strong correlation with the participants’ estimate of epistemic uncertainty, but a weak or null correlation with aleatory uncertainty. Those two effects are robust to the order of questions: if participants are asked to generate their measures of aleatory and epistemic uncertainty first, this does not make them incorporate both types of uncertainty into their predictions. In other words, consumers do not learn from the order of questions. However, we show that a procedure that combines peoples’ estimates of aleatory and epistemic uncertainty into a single estimate of confidence yield better-calibrated confidence intervals. The findings above are replicated in several consumer domains (such as estimates of prices and quality) as well as in general knowledge domains.

Our findings suggest that although people possess the ability to generate reasonably accurate estimates for both kinds of uncertainty, they do not incorporate aleatory uncertainty when asked to generate a confidence interval, they do possess the ability to do so with reasonable accuracy. This “aleatory uncertainty neglect” has significant implication for theory, as it suggests a strong causal mechanism for consumers’ overprecision. Furthermore, well-calibrated intervals can be obtained by mechanically combining separate estimates of two kinds of uncertainty. This result has implications the design of decision aids for consumers, and survey tools for pollsters and marketing researchers who wish to elicit well-calibrated confidence intervals.

REFERENCES
André, Quentin (2016), DistributionBuilder (Version 1.0.0) [Computer Software].


Hofstätter, Peter R. (1939), Über die Schätzung von Gruppeneigenschaften, Barth.


Kahneman, Daniel (2011), Thinking, fast and slow, Macmillan.


Urmins, O. (2016) Probability distributions of common repeated events are misestimated, Working paper