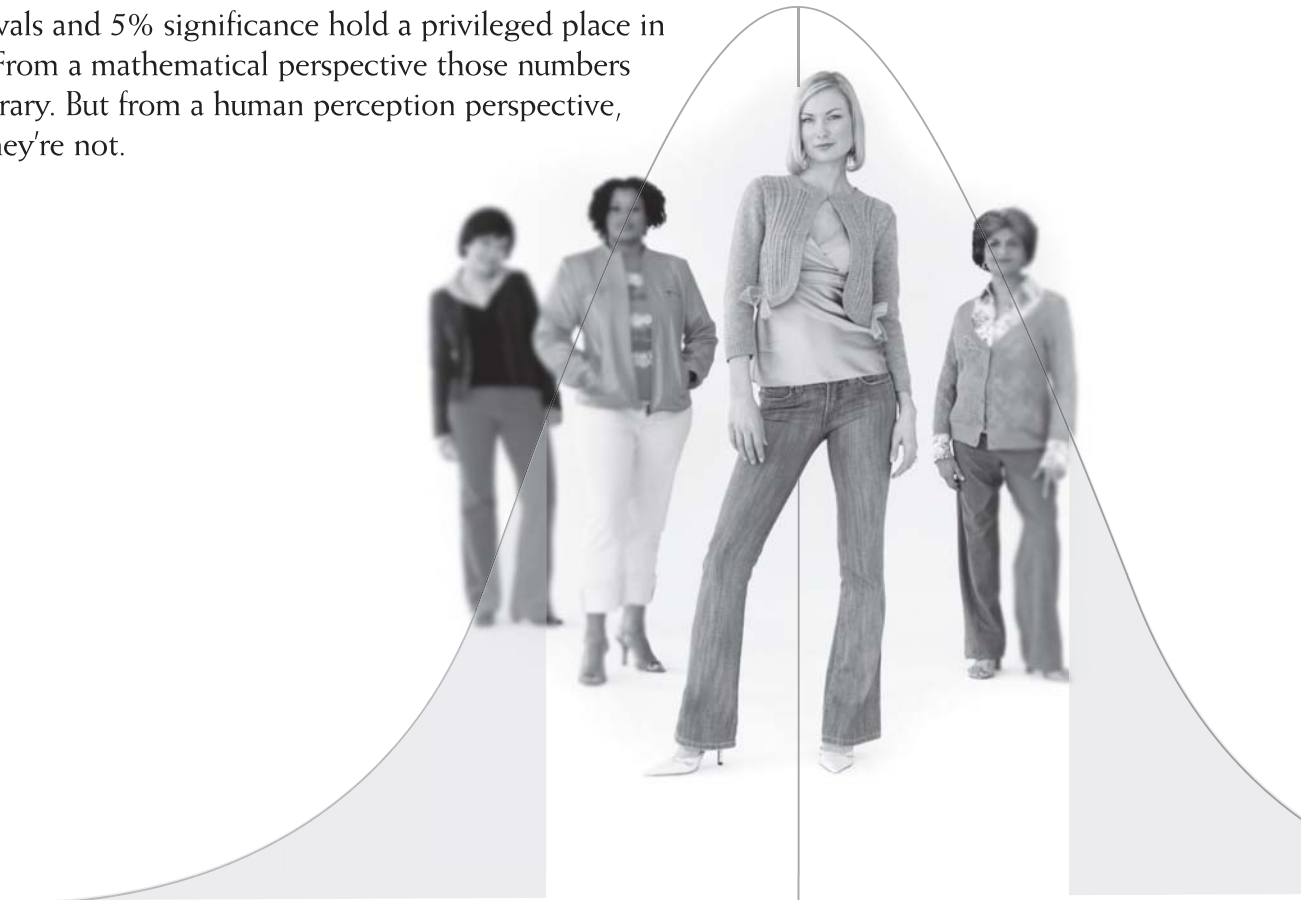


Seeing Significance: Is the 95% Probability Range Easier To Perceive?

Ian Ayres, Antonia R. Ayres-Brown, and Henry J. Ayres-Brown

95% intervals and 5% significance hold a privileged place in statistics. From a mathematical perspective those numbers seem arbitrary. But from a human perception perspective, perhaps they're not.



It is a veritable requirement in the social and natural sciences to justify experimental results by demonstrating that they are statistically significant at at least the 5% level. The origin of this emphasis can be traced to Sir Ronald Fisher, the father of modern statistics, who first championed the standard in 1925 in *Statistical Methods for Research Workers*. Fisher justified this standard, at least in part, because of the 'convenient' fact that, in a normal distribution, the 5% cutoff falls almost exactly at the second standard deviation. In Fisher's words, "The [standard deviation] value for which $P = 0.05$, or 1 in 20, is 1.96, or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice

the standard deviation are thus formally regarded as significant."

However, in "On the Origins of the .05 Level of Statistical Significance," published in *American Psychologist*, M. Cowles and C. David showed that "[a]n examination of the history of probability and statistical theory...indicates that [Fisher's] choice was...influenced by previous scientific conventions that, themselves, were based on the notion of 'chance' and the unlikelihood of an event occurring."

Regardless of the standard's intellectual provenance, a separate and often debated question concerns whether it is arbitrary. One answer to this question is a Bayesian functionalist critique by Dale Poirier and Justin Tobias that powerfully points out that divorcing

hypothesis testing from an explicit characterization of the loss function—what consequentially turns on type I and type II errors—is bound to be suboptimal.

But here we explore another sense in which the 5% or 10% standard for statistical significance might be non-arbitrary. We seek to test whether the 5% or 10% standard for statistical significance might hold a privileged place in our practice because it holds a privileged place cognitively. That is, we seek to test whether people are more accurate in estimating the central 90% and 95% ranges of a normal distribution than in estimating other probability ranges (centered on the mean). We hypothesize that people do a better job in translating their casual observations into estimates of these probability

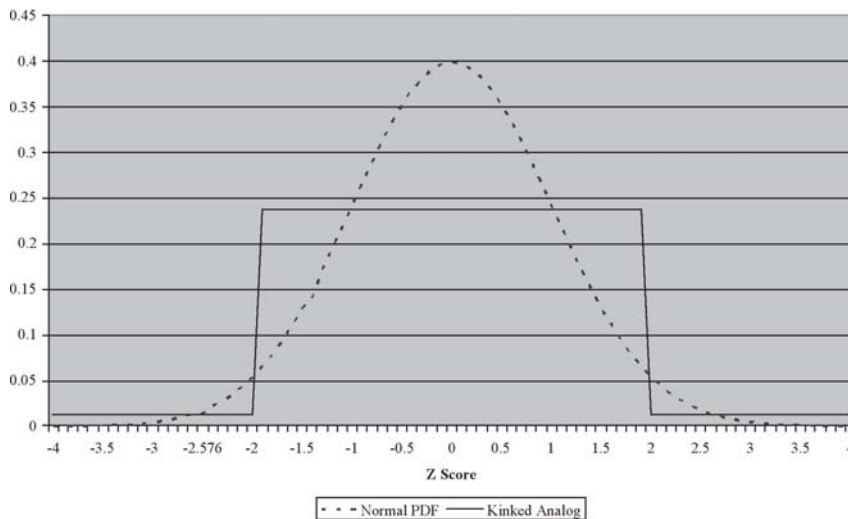


Figure 1. The Normal Distribution and a Kinked Analog

ranges, which happen to coincide with our traditions of statistical significance, than they do in estimating other probability ranges.

We conducted a very simple study. We asked college students to estimate the range of heights that would just include a particular proportion (which, depending on the survey, was 50%, 75%, 90%, 95%, or 99%) of adult women in the United States. Each of the ranges in height that the respondents estimated allowed us to infer an implicit standard deviation estimate. For example, if a respondent said that 95% of adult U.S. women are between 59 and 69 inches tall, we calculated an implicit standard deviation for that response of 2.55 inches ($=10/(2*1.96)$). We then compared the implicit standard deviations of each respondent to the population's actual standard deviation of 2.6 inches, according to the National Center for Health Statistics.

We found that implicit standard deviations of respondents who were asked the 90% question were significantly more accurate than those who were asked to estimate more extreme ranges. This result, while no more than suggestive, is consistent with the idea that our traditional standards of statistical significance may be influenced by our cognitive predispositions. Moreover, the results suggest indirect questioning may elicit more accurate information than direct questioning. If you want to elicit information about the 99% or 50% centered probability ranges, you might do

better to ask a subject to estimate the 90% or 95% ranges.

Theory

Our hypothesis for a cognitive predisposition toward more accurate estimates of the 90% or 95% ranges derives from an intuition that people may have an easier time perceiving relatively kinked points in a probability distribution. Imagine, for example, that some randomized process had 95% of its probability mass uniformly distributed within some range of its mean (which, in Figure 1, is denoted as plus or minus "2z") and its remaining probability mass uniformly distributed in its two tails within the next equally sized increments. We would predict that people who casually observed a finite number of draws from this distribution would most accurately estimate the 95% range (even if they were told nothing about the distribution's analytic structure) because their casual observations would more easily tell them where the drop off in the distribution occurred. We think something analogous occurs for people who casually observe draws from the normal distribution.

While the normal probability density function (pdf) has no kinks, it displays its greatest convex curvature at approximately 1.73 standard deviations from its mean. Falling between the two most convex points of the curve is 91.7% of the normal probability mass. In contrast, the normal curve at one standard deviation displays no curvature (the inflection

point) and asymptotes to no curvature in its tails.

Just as with the kinked distribution, we hypothesized that people can more easily perceive the end of the drop off in the normal distribution—and this place in the distribution is signaled by the convexity at the end of the drop off. The segment of the distribution with maximum curvature is the place where the distribution goes from having a relatively steep slope to a relatively flat slope. In the normal distribution, the density function has its steepest slope (approximately .24) at one standard deviation from its mean, but half that slope occurs at approximately 1.92 standard deviations from the distribution's mean, which corresponds to a 94.5% probability range. Casual observers may have an easier time perceiving not just ranges with greater curvature, but ranges where the slope of the density function is neither too steep to make fine-grained distinctions nor too flat to have much experience with the data.

While it is well known that the normal curve's inflection point occurs at one standard deviation from its mean, it is less well known that the 90% and 95% probability ranges roughly correspond to the places of the density function's maximum curvature and the midpoint of its slope. These attributes of the second derivative of the normal density function (which is the third derivative of the cumulative density function) give rise to testable implications.

Specifically, we hypothesized the following:

- The means of implicit standard deviations of the probability ranges for respondents who answer the 90% or 95% questions will be more accurate than respondents answering the 50%, 75%, or 99% questions.
- The standard deviations of the implicit standard deviations will be lower for respondents answering the 90% or 95% questions than for those answering the 50%, 75%, or 99% questions.
- The implicit standard deviations of the respondents answering the 50% and 75% questions will be too high, and the implicit standard deviations of the respondents answering the 99% question will be too low.

Table 1—Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max
Respondent Age	519	22.3	7.0	10	61
Respondent Female	519	59.3%		0	1
Respondent Height (inches)	519	67.2	3.9	54	79.9
Respondent Location Quinnipiac (0=Yale)	519	41.0%		0	1

Together, these hypotheses are consistent with the idea that people may be hard-wired to more accurately infer the 90% or 95% distribution ranges from their casual observations than other distribution ranges. The first hypothesis is that the mean responses will be more accurate, and the second hypothesis is that there will be less variation in the 90% and 95% responses. The third hypothesis pushes the idea of hard-wired responses even further by conjecturing that people have a predisposition to give the 90% and 95% answer, even when they are asked different questions. Thus, we hypothesize the ranges of heights given as responses to the 99% question will be biased downward toward the 95% answer, while the ranges of heights given as responses to the 50% and 75% questions will be biased upward toward the 90% answer.

Even if our survey's data is consistent with these hypotheses, our simple survey cannot distinguish between "nature" and "nurture." For example, it is also possible that respondents could be more successful at estimating the 95% confidence interval, not because of a hardwired cognitive predisposition concerning the shape of the Gaussian distribution, but because the historic ascendancy of the Fisherian standard of statistical significance has conditioned our respondents to more successfully distinguish the 95% central mass from its tails. But here is one place where the pervasive statistical innumeracy of our population may provide a small benefit, as we find it unlikely that many of our respondents would have more than a passing acquaintance with traditional standards of statistical significance.

Data Collection

In September 2005, we surveyed passersby at two Connecticut universities (Quinnipiac and Yale) in places that were designed to target primarily under-

graduate respondents. Potential respondents were approached at a variety of central campus locations, including the entrance to libraries, gymnasiums, and dining halls. They were asked if they would answer four quick questions in exchange for a Snickers bar. More than 90% of those asked agreed to participate. Respondents who agreed to participate were randomly assigned one of five questionnaires, which asked the following question (varying only with regard to the bracketed information):

The average height of women over 20 years old in the United States is 5'4". Using just your intuition, please give your best estimate of the range of heights that would include [50%, 75%, 90%, 95%, or 99%] of women over 20 years old in the United States. Please express your answer in feet and inches. Please make sure that the

center of your range is the average height of 5'4". Please fill in your answer in the blanks just below the figure.

This question was then followed by a figure that graphically depicted a bell curve and the cutoffs depicted in the question. For example, the figure that followed the 75% question is depicted in Figure 2. The questionnaire then asked the respondents to state their age, sex, and height.

We received 519 complete surveys. Less than 3% of the surveys were discarded because they contained incomplete or illegible responses. The resulting sample was well-balanced across treatments, with between 102 and 105 responses for each of the five treatments. Table 1 provides some summary statistics on respondent characteristics. The bulk of our respondents (64%) were of college age (18-22); a majority of the

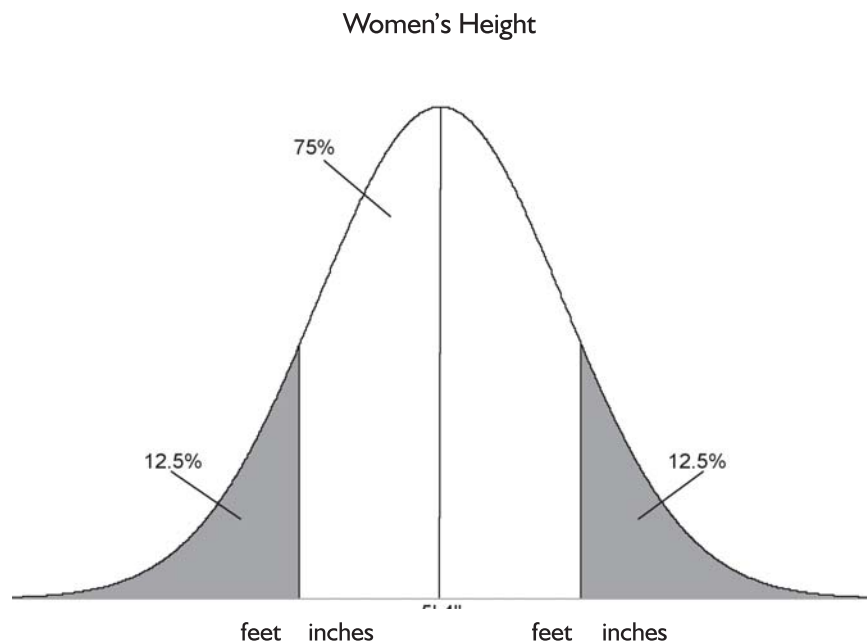


Figure 2

Table 2—Analysis of Implicit Standard Deviations from Respondent Height Ranges for Five Specific Treatment Types

Treatment	A: Mean Size of Height Range (inches)	B: Implicit Number of Standard Deviations within Requested Range	C: Mean of Implicit Standard Deviations (A/B, inches)	D: Standard Deviation of Implicit Standard Deviations (inches)	Freq.
50% Question	5.89	1.35	4.37	2.61	104
75% Question	7.49	2.30	3.26	2.54	104
90% Question	8.55	3.29	2.60	1.69	102
95% Question	8.90	3.92	2.27	1.28	104
99% Question	11.48	5.15	2.23	1.43	105
All Treatments			2.95	2.14	519

respondents were women; and a majority of the responses were recorded at Yale University.

Results

Our basic analysis of the respondents' implicit standard deviations is contained in Figure 3 and Table 2. We found support for two of our three hypotheses. As predicted by our first hypothesis, the respondents most accurately answered the 90% and 95% questions. Indeed, the respondents who answered the 90% question had an average implicit standard deviation that closely matched the standard deviation of adult women in the United States—2.6 inches.

The downward slope of the means in Figure 3 also is consistent with our

third hypothesis: the respondents' height range responses would be biased toward the 90% or 95% range. Table 2 shows the mean height ranges for the 50% and 75% questions were biased upward (toward the 90%/95% range), while the mean height-range answer to the 99% question was biased downward (toward the 90%/95% range). This is consistent with a hard-wired predisposition to answer with the 90% or 95% range, regardless of the range actually sought. This biasing in the reported ranges caused the implicit standard deviations to be too high for the 50% and 75% questions and too small for the 99% question.

The test of our second hypothesis, concerning the standard deviations of our implicit standard deviations, was a bit more mixed. The standard deviations

of respondents' implicit standard deviations were, as predicted, smaller for the 95% question than for the 50%, 75%, and 99% questions. And the standard deviation of the respondents' implicit standard deviations was smaller for the 90% question than for the 50% and 75% questions. But the standard deviation of the respondents' implicit standard deviations was, contrary to our prediction, larger for the 90% question than for the 99% question. Still, in five of six dyadic comparisons, there was smaller dispersion in the implicit standard deviations for the 90% and 95% questions than for the 50%, 75%, and 99% questions.

To explore the statistical significance of these findings, we regressed the implicit standard deviation of each response onto treatment dummies and controls for respondent characteristics. The results of two nested regression specifications are reported in Table 3. The first specification tests whether treatment means (reported in Table 2) for the 50%, 75%, 95%, and 99% questions are statistically different than the treatment mean for the implicit standard deviation of the 90% question. As shown in Table 3, the elevated implicit standard deviations for the 50% and 75% questions were statistically different from that of the 90% omitted treatment implicit standard deviation ($p < .02$). Moreover, an F test strongly rejected ($p < .01$) the hypothesis that the five treatment means were jointly equal. But the .37-inch shortfall of the implicit standard deviation for the 99% treatment question was not statistically

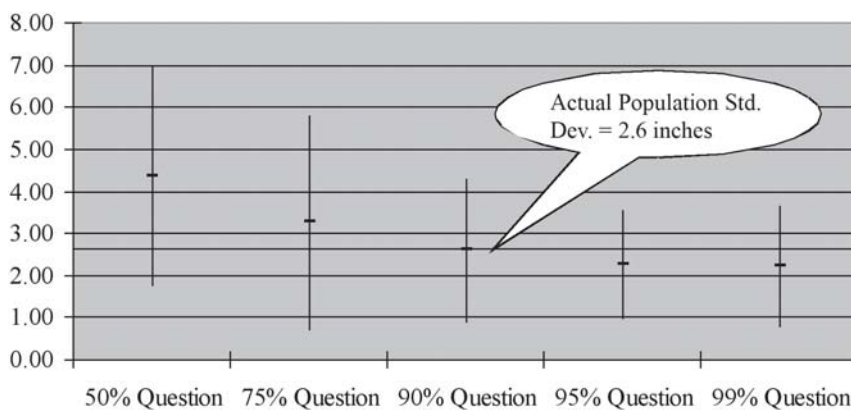


Figure 3. Mean and one standard deviation interval of implicit standard deviations for specific treatment types

different ($p = .18$) from the mean 90% treatment response.

The second specification reported in Table 3 adds controls for a variety of respondent characteristics, including sex, age, and location (Quinnipiac or Yale). We added controls for not only respondent height, but also for how near the respondent's height was to the closest true requested height. We hypothesized that a respondent whose own height happened to be at one of the two requested cutoffs for their treatment might more easily observe the queried proportion of women that was greater or lower than him or herself, and, thus, might be able to more accurately estimate the requested range.

Finally, we added two controls for both the raw and absolute difference between the midpoint of the respondent's height range and the requested midpoint of 64 inches. Even though we asked respondents to center their height ranges on the true population mean height, only 46% of respondent height ranges were exactly centered on the population mean. But 80% of the midpoints were within one inch of the population mean. Moreover, the means of the midpoints for each of the treatments were less than .7-inches from the population mean.

Our predictions about the ordinal ranking of mean treatment effects remained robust to the inclusion of these additional regressors, as the coefficients on the question-type dummies monotonically decreased from positive amounts for the 50% and 75% questions to negative amounts for the 99% question. All three coefficients were statistically significant ($p < .01$).

Conclusion

The results of a single, small study cannot persuasively establish that a cognitive predisposition exists for better 'seeing' the kinked points in normal distributions. This study tells us nothing about other distributions, especially asymmetric distributions that are prevalent in many of life's arenas. And even with regard to approximately normal distributions, this single study of adult women's height may not be robust to different respondents or questions.

Yet even when viewed as merely suggestive, this study points toward a number of implications. If our intuition

Table 3—Regression of Implicit Standard Deviations from Range of Heights on Various Characteristics of Respondent and Treatment Type (t-statistics in parentheses, $p < .05$ in **bold)**

	Spec. 1	Spec. 2
Constant	2.599 (13.18)	2.716 (1.28)
50% Question	1.772 (6.38)	1.875 (7.39)
75% Question	0.659 (2.37)	0.674 (2.65)
95% Question	-0.328 (-1.18)	-0.341 (-1.35)
99% Question	-0.371 (-1.34)	-0.680 (-2.65)
Quinnipiac (Yale=0)		-0.522 (-2.91)
Respondent Age		-0.030 (-2.42)
Respondent Female		-0.505 (-2.23)
Respondent Height (inches)		0.009 (.29)
Difference between Respondent Height and Nearest Probability Cutoff		0.031 (.78)
Difference between Mean of Reported Height Range and 64 Inches		-0.134 (-2.)
Absolute Difference between Mean of Reported Height Range and 64 Inches		0.713 (9.35)
Number of obs	519	519
Adj R-squared	13.5%	28.5%

is correct and it is easier to perceive kinks in a probability density function, then this suggests testable hypotheses for non-normal distributions. Our results also suggest that indirect questioning, properly analyzed, may elicit more accurate information from casual observers than direct questioning. For example, if a lawyer were interested in eliciting testimony from a witness about a 50% (or 99%) probability range, she might do better to ask the witness to estimate the 90% or 95% probability

range and then infer an estimate of the desired range based on the implicit standard deviation.

Indeed, indirect questioning about the 90% or 95% probability range might even improve the ability to elicit information about a distribution's mean. In this study, we arbitrarily chose the 50% probability as our smallest treatment range. But there is no reason why we couldn't have asked respondents to estimate the (centered) 40% or 10% probability range. Indeed, the 0% prob-

A Test within a Test

This study is a test of whether kids can be true collaborators on empirical social science. Two of us are children. Anna was 8 and Henry was 10 (Ian was 46) when we collected and analyzed the data. Our hypothesis was that high school students could—with the help of senior coauthors—conduct and publish serious empirical studies.

Our inspiration was the résumé study of Marianne Bertrand and Sendhil Mullainathan, in which the researchers responded to want ads in the *Boston Globe* and *Chicago Tribune* with résumés that were identical, except that the authors of some were randomly assigned “very African-American-sounding names” and others were assigned “very white-sounding names.” It struck us that many high-school students could have designed and completed this compelling project.

This paper supports our hypothesis. As grade-school children were able to pull their weight on this study, it is surely the case that high-school students can be substantively involved—at least as collaborators—in serious social science.

High-school students are not qualified to substantively contribute to most high-level academic endeavors. Only the rare student can help with a theoretical physics or English article. But high-school students with supervision can collect and analyze important data. Contributing to the production of new information need not wait until college or graduate school.

In the summers of 2004 and 2005, for 45 minutes a day, the coauthors studied statistics together (with the heavy use of Excel). In the fall of 2005, Henry and Anna, under the supervision of their parents, collected the data used in this paper at the two universities. Having children make the request almost certainly increased our response rate. Few people could resist the child’s request: “For a Snicker’s bar, would you please answer four questions to help me on a statistics project?”


Anna and Henry entered the data into an Excel spreadsheet. They double-checked the accuracy of their data, throwing out incomplete surveys and converting metric answers to feet and inches. They also analyzed the data, deriving the implicit standard deviation of each response and estimating the numbers included in all the tables. Additionally, they helped write the paper. In fact, Henry and Anna dictated part of this sidebar to their dad: “We were strongly motivated to work on this project by the prospect of getting our first family dog upon the paper’s acceptance. We have decided to name the dog (Pafnuty) Chebyshev.”

The contributions of the children—in collecting, entering, and analyzing the data—easily met the generally recognized standards of coauthorship. But Anna and Henry did not generate the hypotheses and did not contribute greatly to the design of the specific test. It thus remains an open question whether high-school students (or younger) can come up with interesting, testable hypotheses. But even here we suspect high-school students could contribute substantively to the process of identifying interesting, testable questions.

Indeed, the possibility that people know the 95% probability range better than the mean came directly from the following colloquy between father and daughter. While on a hike, Ian asked Anna how many times in her life she had climbed the Sleeping Giant Trail. Anna replied, “Six times.” Ian then asked what the standard deviation of her estimate was. Anna replied, “Two times.” But then she paused and said, “I want to revise my mean to eight.” During that pause, Anna realized she had implicitly told her father that there was a 95% chance that she had climbed Sleeping Giant between two and 10 times, yet on reflection, she knew she had climbed it more than two times.

Anna could have resolved the contradiction by lowering her standard deviation estimate, but she felt it was more accurate to increase her estimate of the mean. It was this conversation that led us to hypothesize that people may, at times, have a more accurate knowledge of the 95% range than they do of the mean. If you want to elicit more accurate estimates of a mean, you might (at least with regard to normally distributed distributions) want to ask people to identify the 95% probability range. This is an interesting, testable hypothesis that Anna and Ian jointly generated.

ability range is identical to the mean. Our intuition suggests people might be able to more accurately estimate the 95% probability range than the mean. Instead of asking people to estimate the average man’s height, we might do better to ask them to estimate the 95% probability range and infer the mean from the range’s midpoint.

We are in the process of gathering further analogous data from a self-selected group of internet users under the auspices of the Yale School of Management eLab, which will allow us to further test the robustness of these results. Data from both this study and from these further eLab experiments will be available at www.law.yale.edu/ayres. 

Further Reading

Bertrand, M. and Mullainathan, S. (2003). “Are Emily and Brendan More Employable than Latoya and Tyrone? A Field Experiment on Labor Market Discrimination.” *The American Economic Review*. 94(4).

Cochran, W. (1976). “Early Development of Techniques in Comparative Experimentation.” In *On the History of Statistics and Probability*, D.B. Owen (Ed.).

Cowles, M. and Davis, C. (1982). “On the Origins of the .05 Level of Statistical Significance.” *Am. Psychologist*. 553(37).

Fisher, R.A. (1936). *Statistical Methods for Research Workers* (6th Ed.).

National Center for Health Statistics. (2001). “Table 12: Height in Inches for Females 20 Years and Over: United States, 1988–1994.” *National Health and Nutrition Examination Survey*. www.cdc.gov/nchs/data/nhanes/t12.pdf.

Poirier, D.J. and Tobias, J.L. (2006). “Bayesian Econometrics.” *Palgrave Handbook of Econometrics, Volume 1: Econometric Theory*.

U.S. Department of Health and Human Services. (1987). *Vital and Health Statistics, Anthropometric Reference Data and Prevalence of Overweight, United States, 1976–80, Data from the National Health Survey Series 11, No. 238. Page 25, Table 14*. Available at www.cdc.gov/nchs/data/series/sr_11/sr11_238.pdf.

About the Authors



David H. Annis is vice president of forecasting and pricing at Wachovia Treasury. His current research is in developing default and prepayment models for home mortgage loans. Prior to joining the bank, he was on the operations research faculty at the Naval Postgraduate School, where he taught statistics. An avid football fan, Annis maintains www.sportsquant.com, a site devoted to objective sports analysis. david.annis@wachovia.com.



Ian Ayres, Townsend professor at Yale Law School, is a lawyer who holds a PhD in economics from MIT. He is a regular commentator for public radio's "Marketplace" and a columnist for *Forbes*. He is the author of several empirical studies, including those on car theft, concealed weapons, kidney transplantation, taxicab tipping, and reckless sex.



Antonia R. Ayres-Brown is in fourth grade at The Foote School. She enjoys drama, writing stories, and collecting dolls. She won first place in the age 8 and under breaststroke at the New Haven Summer Swim Championships. She also used Excel extensively in a corporate finance course she audited at Yale last year.



Henry J. Ayres-Brown is in sixth grade at The Foote School. He writes for his school newspaper and has played "Billy" in Sondheim's "Assassins" (Yale Dramat), "James Rodgers" in "The Will Rogers Follies" (Stagedoor manor), and both "Fritz" and the "Prince" in "The Nutcracker" (New Haven Ballet).



Jarrett Barber is assistant professor in the Department of Statistics at the University of Wyoming in Laramie. He holds degrees in forestry and mathematics and received his PhD in statistics from North Carolina State University in 2002. He completed a joint post-doctoral position at the Institute of Statistics and

Decision Sciences at Duke University in Durham and at the Geophysical Statistics Project at the National Center

for Atmospheric Research in Boulder, Colorado. Jarrett spent two years as assistant professor in the Department of Mathematical Sciences at Montana State University in Bozeman and has taught aerial photo interpretation for the USDA Forest Service. His professional interests lie in spatial statistical modeling and methodology and in applications to ecology and the environment. His most recent professional interests involve statistical modeling of map positional error. Jarrett is an active member of the American Statistical Association's Section on Statistics and the Environment. He enjoys hiking, camping, and cross-country skiing.



Kjetil K. Haugen is professor of logistics at Molde University College, Norway, and holds academic positions at the Norwegian University of Science and Technology. He graduated with an MSc in operations research in 1984 and earned a PhD in computer science in 1991 from the Norwegian University

of Science and Technology. He has held various research administrative positions at SINTEF in Trondheim and at More Research, Molde. Present interests include sports (mainly soccer) economics. Haugen, born in 1959 in Kristiansund, Norway, is married with three children, aged 6, 8, and 12.



Nicholas Horton earned his doctorate in bio-statistics from the Harvard School of Public Health in 1999. He is assistant professor of mathematics and statistics at Smith College in Northampton, Massachusetts. His research interests are longitudinal regression models and missing data methods, with applications in psychiatric epidemiology and substance abuse

research. When he's not chasing after his kids, he enjoys juggling and exploring rail-trails. nborton@email.smith.edu.



Steven J. Miller is an assistant professor of mathematics at Brown University. He earned a BS in mathematics and physics from Yale in 1996 and a PhD in mathematics from Princeton in 2002. His main research interests are in number theory and probability and applied problems in accounting, computer science, economics, marketing, and statistics. In addition to several research papers, he is the coauthor of *An Invitation to Modern Number Theory*, which introduces students with minimal prerequisites to open problems in the field.