Of beauty, sex, and power: Statistical challenges in estimating small effects[*]

Andrew Gelman[†] David Weakliem[‡]

October 27, 2008

Abstract

How do we interpret findings that are intriguing, potentially important, but not statistically significant? We discuss in the context of a series of papers in the *Journal of Theoretical Biology* that reported evidence that beautiful parents have more daughters, violent men have more sons, and other sex-ratio patterns (Kanazawa, 2005, 2006, 2007). These papers have been shown to have statistical errors, but the more general research questions remain. From a classical statistical perspective, these studies have insufficient power to detect the magnitudes of effects (on the order of 1 percentage point) that could be expected based on earlier studies of sex ratios. The anticipated small effects can also be handled using a Bayesian prior distribution. These concerns are relevant to other studies of small effects and also to the reporting of such studies.

Keywords: Bayesian inference, evolutionary psychology, power analysis, prior distribution, sex ratio, sociobiology, Trivers-Willard hypothesis, Type M (magnitude) error, Type S (sign) error

[*] We thank Kelly Rader and Lucy Robinson for data collection and analysis, David Dunson, Thomas Volscho, David Judkins, and Jeremy Freese for helpful comments, and the National Science Foundation, National Institutes of Health, and Columbia University Applied Statistics Center for financial support.

[†] Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman

[‡] Department of Sociology, University of Connecticut, weakliem@uconn.edu, web.uconn.edu/weakliem

The popular media often report on scientific studies that rely on statistical evidence. Most journalists are not equipped to evaluate the original research, and so they use peer-reviewed journals as a filter: they assume that published work is scientifically respectable and merits reporting if it seems likely to interest the public. Reviewers typically recommend publication only if results reach conventional levels of statistical significance. Not all "statistically significant" results are equivalent, however: some are just slightly beyond the minimum level, while others are significant only under certain assumptions or ways of posing the question. On the other hand, some potentially interesting results fall just short of conventional levels of significance. Here we consider the interpretation of statistical results in the gray area of significance—where there is some evidence, but not definitive evidence.

Do beautiful parents have more daughters?

As a result of the intrinsic interest of the topic and the availability of data from birth records, there have been many studies of factors affecting the probability of male and female births. Most have found little or no evidence of any effects, but a recent study by Satoshi Kanazawa published in the *Journal of Theoretical Biology* appears to be an exception. Kanazawa began with the generalized Trivers-Willard hypothesis, which holds that the probability of male and female offspring will vary in ways that increase the expected number of descendants. Kanazawa proposed that attractive parents will be more likely to have daughters, on the grounds that physical attractiveness enhances the reproductive success of women more than that of men. He tested this hypothesis using data from the National Longitudinal Survey of Adolescent Health, which included interviewers' subjective assessments of respondents' attractiveness (on a 1 to 5 scale) along with data on the sex of respondents' children. To remove potential biases arising from decisions about family size, the analysis was restricted to first-born children. Kanazawa found that 52% of the childern born to respondents in the highest category of attractiveness were girls, compared to only 44% of the children of parents in the four lower categories. This difference is statistically significant at the 5% level. These results were widely reported in the media, and Kanazawa has presented these and related claims to general audiences with an article in *Psychology Today* and a book (Miller and Kanazawa 2007a; 2007b).

The published analysis, however, comparing the parents assessed as most attractive to those in the four lower categories, is just one of many that could be performed with these data, and the use of standard significance levels is problematic under these conditions. Gelman (2007a) suggested an alternative approach that is more consistent with standard statistical practice: a regression of the proportion of girl births on the parent's attractiveness rating. If we rescale attractiveness to have a mean of 0 and standard deviation of 0.5,[1] a regression of the proportion of girl births on attractiveness yields an estimated coefficient of 0.047 with a standard error of 0.043. The challenge is to interpret this finding, which is not statistically significant but is consistent with Kanazawa's story.[2]

There also are substantive reasons why Kanazawa's hypothesis should not be dismissed out of hand, even though his results are not statistically significant. First, his findings are motivated theoretically by a well-respected model of Trivers and Willard (1973); see Fawcett et al. (2007) for a recent review. In addition, beauty has been found to have effects in areas including economics (Hamermesh and Biddle, 1994) and education (Hamermesh and Parker, 2005), and the sex of children has been associated with political attitudes of parents (Oswald and Powdthavee, 2006, Washington, 2007).

In a statistical study, observed differences are a combination of true differences and chance variation. With a binary outcome, the chance variation has a standard deviation of approximately $0.5/\sqrt{n}$, which equals 1.1% for a sample size of $n = 2000$ or 0.5% for a sample size of 10,000. The influence of chance can go in either direction, and it is impossible to be sure whether the differences observed in a given sample are larger, smaller, or equal to the true differences. In many cases, however, previous research gives us a sense of the likely size of any differences. With the sex ratio of human births, effects have almost always been small, on the order of 1 percentage point (for example, the probability of a girl birth shifting from 48.5% to 49.5%). Variation attributable to factors such as race, parental age, birth order, maternal weight, partnership status, and season of birth has been estimated at from less than 0.3 percentage points to about 2 percentage points (James, 1987, Chahnazarian, 1988, Cagnacci et al., 2003, 2004, 2005, Norberg, 2004), with larger changes (as high as 3 percentage points) arising under economic conditions of poverty and famine (Ansari-Lari and Saadat, M., 2002, Catalano, 2003, Almond et al., 2007). That extreme deprivation increases the proportion of girl births is no surprise, given that male fetuses (and also male babies and adults) are more likely than females to die under adverse conditions.[3] Consequently, it is reasonable to expect any effects of beauty on the sex ratio

[1] The new scale makes it comparable to a binary variable contrasting cases that rank above and below average; see Gelman (2007b).

[2] In other *Journal of Theoretical Biology* articles, Kanazawa (2005, 2006, 2007) and Kanazawa and Vandermassen (2005) have reported that "Big and tall parents have more sons," "Violent men have more sons," and "Engineers have more sons, nurses have more daughters." These analyses cannot be interpreted at face value, however, as they made the error of controlling for an intermediate outcome in estimating a causal effect (Gelman, 2007a).

[3] Large sex ratio differences in the other direction have been observed—an excess of boy births in parts of

to be less than 1 percentage point, which represents the range of natural variation under normal conditions.

Classical inference. Let us return to our linear regression of sex ratio on beauty. With an estimate of 4.7% and a standard error of 4.3%, the classical 95% interval for the effect of parental attractiveness on the probability of a female birth is $[-3.9\%, 13.3\%]$. To put it another way, effects as low as -3.9 or as high as $+13.3$ percentage points are roughly consistent with the data. Given that we could begin with an expectation that effects were in the range ±1%, we have essentially learned nothing from this study.

Another way to frame this point is to consider what would happen if repeated independent studies were performed with the same precision. Classical statistical theory distinguishes between two kinds of errors: obtaining statistically significant results when there is no real effect (Type 1 error) and failing to obtain statistically significance when there is a real effect (Type 2 error). The chance of Type 1 errors is fixed by the choice of a level of significance: the conventional 5% level implies a 5% chance. The chance of Type 2 errors depends on the size of the true effect and on the design of the study: in general, using a larger sample reduces the chance of Type 2 errors. It is possible to determine the sample size required to keep the risk of Type 2 errors to a low level given an effect of specified size. What constitutes a reasonable size can be decided based on a combination of relevant theory and research, the researcher's judgment, and the purposes of the study. Such "power calculations" are routinely carried out by investigators proposing experimental research, and are required by funding agencies such as the National Institutes of Health. As Morris (1987) points out, even highly statistically significant results are not particularly informative in a astudy that provides insufficient power with respect to reasonable models.

For the present example, the standard error of 4.3% means that statistical significance would only happen with an estimate of at least 8.4% in either direction: much larger than any true effect that could reasonably be expected based on previous research. Thus, even if the inference of an association between parental beauty and child's sex is valid for the general population, the magnitude of the estimate from a study of this size is likely to be much larger than the true effect. This is an example of what Gelman and Tuerlinckx (2000) call a Type M (*magnitude*) error. They also note the possibility of Type S (*sign*) errors, in which a statistically significant estimate is in the opposite direction of the true effect.

India, China, and other countries—but this is likely attributable to selective infanticide and abortion (Das Gupta, 2005, 2006, 2007) and not so relevant to the present discussion.

We may get a sense of the probabilities of these errors by considering three scenarios of studies with standard errors of 4.3 percentage points:

1. **True difference of zero.** If there is no correlation between parental beauty and sex ratio of children, then a statistically significant estimate will occur 5% of the time, and it will always be misleading—a Type 1 error.

2. **True difference of 0.3%.** If the probability of girl births is actually 0.3 percentage points higher among beautiful than among ugly parents, then there is a 3% probability of seeing a statistically significant positive result—and a 2% chance of seeing a statistically significant negative result.

   In either case, if the estimate is statistically significant, it must be at least 8.4 percentage points, over an order of magnitude higher than the true effect, and with a 40% chance of going in the wrong direction. If the result is *not* statistically significant, the chance of the estimate being in the wrong direction is 47.5%, implying that the direction of the estimate provides almost zero information on the sign of the true effect.

3. **True difference of 1%.** If the probability of girl births is actually 1 percentage point higher among beautiful than among ugly parents—which, based on the literature, is on the high end of possible effect sizes—then there is a 4% chance of a statistically significant positive result, and still over a 1% chance of a statistically significant result in the wrong direction.

   Again, the estimated effect, if statistically significant, will be over 8 times the magnitude of the true effect. If the estimate is not statistically significant, there remains a 40% chance of it having the wrong sign. Thus, again, the experiment gives little information about the sign or the magnitude of the true effect.

4. **True difference of 3%.** Even if the true difference were as high as 3 percentage points, which we find implausible from the literature review, there is still only a 10% chance of obtaining statistical significance, and such an estimate would be over twice as large as the true effect. A non-significant estimate would have a 24% chance of going in the wrong direction.

A sample of this size is not useful for estimating variation on the order of 1 or even 3 percentage points, which is why most studies of the human sex ratio use much larger samples, typically from demographic databases (e.g., Almond and Edlund, 2006, Almond et al., 2007). The example shows that if the sample is too small relative to the expected size of any differences, it is not possible to draw strong conclusions even when estimates are statistically significant: there is a good chance that the true effect is much smaller than the estimated one, or even in the opposite direction.

Bayesian inference. An alternative approach to statistical inference is based on the idea of expressing prior beliefs about the likely size of any effects in the form of a prior distribution. This distribution can be combined with the sample information to produce a posterior distribution, which represents a reasonable range of conclusions given the data and prior beliefs. Very weak prior beliefs can be represented by a "diffuse" distribution—one that is approximately uniform over a large range of values. A diffuse distribution implies that all values in the range are about equally likely. In this case, with a sufficiently diffuse prior distribution, the posterior distribution would be approximately normal with mean 4.7% and standard error 4.3%, which would imply about an 86% probability that the true effect is positive. In general, the more concentrated the prior distribution around 0 (expressing a presumption based on the sex-ratio literature that the true effect is likely to be small), the closer the posterior probability will be to 50%.

The idea that any effects are likely to be small can be represented by using a Cauchy distribution with center 0 and scale 0.3%. This distribution implies that the true difference in the proportion of girl births, comparing beautiful and ugly parents, is most likely to be near zero, with a 50% chance of being in the range $[-0.3\%, 0.3\%]$, a 90% chance of being in the range $[-1\%, 1\%]$, and a 94% chance of being less than 3 percentage points in absolute value. We center the prior distribution at zero because, ahead of time, we have no reason to believe that the effect will be in one direction rather than the other.

Compared to alternatives such as the normal distribution, the Cauchy family gives a relatively high probability of very large effects. That is, the Cauchy distribution gives more room to the possibility that our prior belief in a small effect is badly mistaken.

We combine the prior distribution with the normal likelihood for the regression coefficient on the standardized beauty coefficient (that is, the likelihood corresponding to the normal distribution with mean 4.7% and standard deviation 4.3%). The resulting posterior distribution gives a probability of only 58% that the difference is positive—that beautiful parents actually have more daughters—and even if the effect is positive, there is a 78% chance it is less than 1 percentage point. This analysis depends on the prior distribution but not to an extreme amount; for example, if we broaden the Cauchy prior and give it a scale of 1%, the posterior probability that the true difference is positive is still only 65%. Switching from the Cauchy to another family such as the normal distribution has little effect on the results. The key is that effects are likely to be small, and in fact the data are consistent with small results.

Additional data: children of the 50 most beautiful people

One way to calibrate our thinking about these results is to collect more data. Every year, *People* magazine publishes a list of the "fifty most beautiful people," and, because they are celebrities, it is not difficult to track down the sexes of their children, which we did for the years 1995–2000.[4]

As of 2007, the 50 most beautiful people of 1995 had 32 girls and 24 boys, or 57.1% girls, which is 8.6 percentage points higher than the population frequency of 48.5%. This sounds like good news for the hypothesis. But the standard error is $0.5/\sqrt{56} = 6.7\%$, so the discrepancy is not statistically significant. Let's get more data.

The 50 most beautiful people of 1996 had 45 girls and 35 boys: 56.2% girls, or 7.8% more than in the general population. Good news! Combining with 1995 yields 56.6% girls—8.1% more than expected—with a standard error of 4.3%, tantalizingly close to statistical significance. Let's continue to get some confirming evidence.

The 50 most beautiful people of 1997 had 24 girls and 35 boys—no, this goes in the wrong direction, let's keep going ... For 1998, we have 21 girls and 25 boys, for 1999 we have 23 girls and 30 boys, and the class of 2000 has had 29 girls and 25 boys.

Putting all the years together and removing the duplicates, such as Brad Pitt, *People*'s most beautiful people from 1995 to 2000 have had 157 girls out of 329 children, or 47.7% girls (with standard error 2.8%), a statistically insignificant 0.8% percentage points lower than the population frequency. So nothing much seems to be going on here. But if statistically insignificant effects from a sample size sufficient for a standard error of 4.3% were considered acceptable, we could publish a paper every two years with the data from the latest "most beautiful people."

[4] Data were collected from Wikipedia, the Internet Movie Database, and celebrities' personal webpages in August, 2007. Information was missing for 2 beautiful people in 1995, 2 in 1996, 3 in 1997, 6 in 1998, 3 in 1999, and 2 in 2000. The data are available at `www.stat.columbia.edu/~gelman/research/beautiful/`.

Two other examples of the difficulties of inference with small sample sizes

We briefly illustrate the omnipresent difficulty of statistical significance and sample size with two recent examples, the first also involving a recent study of human sex ratios. Mathews, Johnson, and Neil (2008) found an association between maternal diet and child's sex. After constructing an index of general energy and nutritional intake before conception, they found that 56% of the women in the highest third had sons, compared to 45% in the lowest third. As in Kanazawa's studies, the estimated effects were much larger than those found in previous studies of the sex ratio. Statistically, the results were stronger than Kanazawa's: the differences were significant at the 1% level, a value that is sometimes labeled "highly significant."[5] Assuming a true difference of 1 percentage point in the probability of girl births, it turns out that a study of this size has about a 1% chance yielding of a positive result that is significant at the 1% level and a 0.2% chance of a negative result that is significant at that level. Overall, there is about a 16% chance of a Type S error. The reason that even "highly significant" results leave a substantial chance of error is that the sample is small—about 720 women. With a sample of this size, the magnitude of sampling error is large relative to the magnitude of the anticipated effects.

The second example involves an important outcome for which the number of cases in the samples is necessarily small: presidential elections. A number of observers have noticed an intriguing pattern in the results of previous elections: in most cases, the taller major-party candidate has been the winner (Open N.Y., 2008). It is possible to make a case that height is a real advantage: voters may unconsciously associate height with dominance or leadership. It is also plausible to suppose that any influence of height should be more dramatic in since the introduction of television broadcasting, which made voters more aware of the candidates' appearance. In this period, the taller candidate won the popular vote in ten elections and lost in only three.[6] Despite the small sample, this rate is statistically significantly different from 50%. If height had no influence, the taller candidate would have a 50% of winning in any given election, and the chance that the taller candidate would win 10 or more or 13 elections would be only about 1.1%.

Once again, however, the results should be assessed in light of the likely size of any effect. In contrast to the examples involving the sex ratio, there is not a large body of previous research. However, general knowledge of presidential elections suggests that any

[5] Mathews et al. report a significance of .00095, but this appears to be a misprint; the significance of the reported statistic is .0095.

[6] Al Gore lost the election in 2000, but led in the popular vote.

influence of height is likely to be fairly small, because there are many other factors that can affect success. Suppose that the taller candidate actually has a 55% chance of winning (averaging over all other factors). In that case, the chance that the taller candidate would win 10 or more of 13 elections would be only about 2.7%. Even with a 60% chance of winning a given election, the chance of doing this well would be only about 5.8%. That is, the observed result is unusual under any reasonable assumptions about the likely size of the effect. Consequently, it is mostly a matter of chance and provides only weak evidence about the possible influence of height on a candidate's success.

Why is this important?

The point of these examples is not to single out particular pieces of research for criticism. They are illustrations of a more general problem: that the structure of scientific publication and media attention seem to produce a bias toward overstating the magnitude and certainty of effects found in small studies. Journalists are naturally attracted to startling results and can downplay qualifications or uncertainties.

Before reaching *Psychology Today* and book publication, Kanazawa's findings received broad attention in the news media. For example, the popular Freakanomics blog (Dubner, 2006) reported,

> "a new study by Satoshi Kanazawa, an evolutionary psychologist at the London School of Economics, suggests . . . there are more beautiful women in the world than there are handsome men. Why? Kanazawa argues its because good-looking parents are 36% more likely to have a baby daughter as their first child than a baby son—which suggests, evolutionarily speaking, that beauty is a trait more valuable for women than for men. The study was conducted with data from 3,000 Americans, derived from the National Longitudinal Study of Adolescent Health, and was published in the Journal of Theoretical Biology."

Publication in a peer-reviewed journal seemed to have removed all skepticism, which is noteworthy given that the authors of *Freakanomics* are themselves well qualified to judge social science research. The blog entry mentioned that the study was party of a series of studies about male and female births, including one that found that scientists and engineers were more likely to have sons. It concluded with an ironic comment: "It is good that Kanazawa is only a researcher and not, say, the president of Harvard. If he were, that last finding about scientists may have gotten him fired."

But there was no mention of the substantial range of uncertainty in the estimates. Moreover, the estimated effect grew during the reporting. As noted above, the 4.7 percentage point difference in the data (which was not actually statistically significant) became 8 points in Kanazawa's choice of the largest comparison. Kanazawa reported this difference as an odds ratio, $\frac{52/48}{44/56}$, which is approximately equal to 1.36. The 36% figure thus referred to the change in the odds ratio, not the percentages of male and famale births. To a reader familiar with the research literature, the size of the estimated effect could be an indication that something was wrong, since it was so much larger than those found in other studies of the sex ratio (James, 1987, Chahnazarian, 1988). To a journalist or casual reader, however, a claimed 36% effect could be taken as evidence that the differences were real and large enough to be of practical importance.

This problem will continue to occur and is worth thinking about now. Most of the low-hanging fruit in social science have presumably been plucked, so most contemporary research focuses on small effects. As discussed earlier in this article, the influence of chance variation means that studies with insufficient statistical power will occasionally produce results that are not only statistically significant, but that appear to be large. In fact, large differences are more likely to occur in small samples. This is an example of a general phenomenon discussed by Wainer (2007): the smaller the sample, the greater the relative influence of random variation. When such results bear on questions of public interest, it is tempting to emphasize their potential importance and understate the degree of uncertainty. Miller and Kanazawa's billing of their inference as a "politically incorrect truth" hints at the connection to live political issues such as abortion, parental leave policies, and comparable-worth laws that turn upon judgments of the appropriate roles for men and women in society.

The dangers of mistaking essentially random findings for real results is particularly high when a theory can yield predictions of effects in either direction depending on various auxiliary assumptions. For example, in Kanazawa's beauty-and-births study, evolutionary psychology could be used to predict a result in the opposite direction, using the following sort of argument: persons judged to be beautiful are more likely to be healthy, affluent, and from dominant ethnic groups, more generally having traits that are valued in the society at large. (Consider, for example, Miss Americas, who until recent decades were all white.) Such groups are more likely to exercise power, a trait that, in some sociobiological arguments, are more beneficial for men than women—thus it would be natural for more attractive parents to be more likely to have boys. We are not claiming this is true; we are just noting that the argument could go in either direction, which puts a particular burden on the data analysis.[7] Freese (2007) describes this sort of argument as "more 'vampirical' than 'empirical'—unable to be killed by mere evidence" (see also Freese and Powell, 2001, and Volscho, 2005). There seems to be a human desire to find more than pure randomness in sex ratios, despite that there is no convincing evidence that boys or girls run in families or that sex ratios vary much at all except under extraordinary conditions.[8] The problem is not in the underlying theory, but in the approach of offering a variety of intriguing predictions and undertaking small-scale analyses in the hope of confirming them.

In statistics, you can't prove a negative. "Beautiful parents have more daughters" is a compelling headline, but "There is no compelling evidence that beautiful parents are more or less likely to have daughters" is not, and "Weak evidence that beautiful parents have more daughters; more research suggested" is not much better. The result is a sort of asymmetrical warfare, with proponents of sex differences and other "politically incorrect" results producing a series of empirical papers that, for reasons of statistical power, give essentially random clues about true population patterns, and opponents of this line of research being reduced to statements such as "the data are insufficient" or claims that such research is dangerous and out of bounds, an attitude rightly deplored by Pinker (2007). Such debates could lead to a general public skepticism that would, in boy-who-cried-wolf fashion, unfairly discredit entire areas of research.

What should be done?

It is well known that with a large enough sample size, even a very small effect will be statistically significant. Many textbooks contain warnings about mistaking statistical significance in a large sample for practical importance. It is also well known that it is difficult

[7] Another argument that the effect could go in the other direction is based on the connection of the generalized Trivers-Willard hypothesis to the pattern that men show more variation than women in many traits. The theory, as Kanazawa describes it, could be expressed by saying that the probability of a male child is positively related to $\mathrm{E}(N|M_q) - \mathrm{E}(N|F_q)$, where $\mathrm{E}(N|M_q)$ is the expected number of children that a male child born to parents with a particular quality would have, and $\mathrm{E}(N|F_q)$ is the expected number of children a female would have. Even if the correlation between beauty and the number of children is stronger for women (which could be plausible), the variance of the number of children is larger for men, so the slope of a regression of number of children on beauty might be larger for men. So in terms of evaluating the theory, getting empirical evidence on actual variation (if any) in $\mathrm{E}(N|M) - \mathrm{E}(N|F)$ is important. The direct analysis of parental beauty and children's sex amounts to taking a variety of possibly weak instruments as substitutes for the theoretically important variable.

[8] For example, recently there has been discussion of timing of intercourse and the sex ratio, with some studies reporting large effects, and various plausible stories to explain these effects. However, careful analyses have found that the true impact of intercourse timing is very small if anything (David Dunson, private communication).

to obtain statistically significant results in a small sample. Consequently, when results are significant despite the handicap of a small sample, it is natural to think that they are real and important. The examples discussed above show that this is not necessarily the case.

If the estimated effects in the sample are much larger than those that might reasonably be expected in the population, even statistically significant results provide only weak evidence of any effect. Yet one cannot simply ask researchers to avoid using small samples. There are cases in which it is difficult or impossible to obtain more data, and researchers must make do with what is available. We offer two practical recommendations for such situations:

First, whenever possible, researchers should determine plausible effect sizes based on previous research or theory, and carry out power calculations based on the observed test statistics. Conventional significance levels tell us how often the observed test statistic would be obtained if there were no effect, but one should also ask how often the observed test statistic would be obtained under a reasonable assumption about the size of effects. Sample effects that are much larger than expected might indicate that our assumptions were wrong and that the population effects are much larger than previously imagined. Often, however, large sample estimates will merely reflect the influence of random variation. These calculations may be disappointing to researchers, since they may indicate that even "significant" findings do not provide strong evidence. Accurately identifying findings that are suggestive rather than definitive, however, will benefit both the scientific community and the general public.

Second, researchers should exercise more ingenuity in efforts to obtain additional data from outside of their primary sample. There may be records of cases that are recognized as having extreme values on the variable of interest: for example, the "50 most beautiful people" discussed above. It may also be reasonable to broaden the scope of the investigation—for example, a quality that provides an advantage in elections for one office might be expected to provide a similar advantage in elections for others. After Todorov, Mandisodza, Goren, and Hall (2005) found evidence that facial appearance affected success in several rounds of elections for the U.S. Congress, Ballew and Todorov (2007) obtained similar results for gubernatorial elections. Although the statistical significance in some of the individual studies was marginal, the combined results provided much stronger evidence.[9]

In the sex ratio example, a researcher has to make two arguments: a statistical case that observed patterns represent real population effects and cannot be explained simply

[9] That said, the findings of Todorov et al. are associational rather than causal: facial appearance is not an assigned treatment and is associated with political variables; see Atkinson, Enos, and HIll (2008).

by sampling variability, and a biological argument that effects on the order of 1% are substantively important. A claimed effect size of 36% should arouse suspicion in comparison to the literature on the human sex ratio. Beyond this, though, it should be recognized that with these sample sizes, *only* very large estimated effects could make it through the statistical-significance filter. The result is almost a machine for producing exaggerated claims, which of course become only more exaggerated when they hit the credulous news media with the seal of scientific approval.

Statisticians should take some of the blame for this situation. Classical significance calculations do not make use of prior knowledge of effect sizes, and Bayesian analyses are often not much better. Textbook treatments of Bayesian inference (e.g., Carlin and Louis, 2001, Gelman et al., 2003) almost entirely use noninformative prior distributions and essentially ignore issues of statistical power. Conversely, power calculations are commonly used in designing studies but are rarely used to enlighten data analyses, and theoretical concepts such as Type S and Type M errors (Gelman and Tuerlinckx, 2000) have not been integrated into common practice.

The solution to difficulties of statistical communication is to have more open exchange of methods and ideas. Paying more systematic attention to the size of estimated effects in small samples would provide a clearer framework for open communication.

References

Almond, D., and Edlund, L. (2006). Trivers-Willard at birth and one year: evidence from U.S. natality data 1983–2001. *Proceedings of the Royal Society B.*

Almond, D., Edlund, L., Li, H., and Zhang, J. (2007). Long-term effects of the 1959–1961 China famine: mainland China and Hong Kong.

Ansari-Lari, M., and Saadat, M. (2002). Changing sex ratio in Iran, 1976–2000. *Journal of Epidemiology and Community Health* **56**, 622–623.

Atkinson, M., Enos, R. D., and Hill, S. J. (2008). Candidate faces and election outcomes. Technical report, Department of Political Science, University of California, Los Angeles.

Ballew, C. C., and Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences* 104, 17948–17953.

Cagnacci, A., Renzi, A., Arangino, S., Alessandrini, C., and Volpe, A. (2003). The male disadvantage and the seasonal rhythm of sex ratio at the time of conception. *Human*

Reproduction **18**, 885–887.

Cagnacci, A., Renzi, A., Arangino, S., Alessandrini, C., and Volpe, A. (2004). Influences of maternal weight on the secondary sex ratio of human offspring. *Human Reproduction* **19**, 442–444.

Cagnacci, A., Renzi, A., Arangino, S., Alessandrini, C., and Volpe, A. (2005). Interplay between maternal weight and seasons in determining the secondary sex ratio of human offspring. *Fertility and Sterility* **84**, 246–248.

Carlin, B. P., and Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*, second edition. London: CRC Press.

Catalano, R. A. (2003). Sex ratios in the two Germanies: a test of the economic stress hypothesis. *Human Reproduction* **18**, 1972–1975.

Chahnazarian, A. (1988). Determinants of the sex ratio at birth: review of recent literature. *Social Biology* **35**, 214–235.

Das Gupta, M. (2005). Explaining Asia's "missing women": a new look at the data. *Population and Development Review* **31**, 529–535.

Das Gupta, M. (2006). Cultural versus biological factors in explaining Asia's "missing women": response to Oster. *Population and Development Review* **32**, 328–332.

Das Gupta, M. (2007). China's "missing girls"—son preference or hepatitis B infections? Research brief, Human Development and Public Services Research, World Bank.

Dubner, S. J. (2006). Why do beautiful women sometimes marry unattractive men? Freakanomics blog, 2 August, 9:44 am.

Fawcett, T. W., Kuijper, B., Pen, I., and Weissing, F. J. (2007). Should attractive males have more sons? *Behavioral Ecology* **18**, 71–80.

Freese, J. (2007). The problem of predictive promiscuity in deductive applications of evolutionary reasoning to intergenerational transfers: three cautionary tales. *Caring and Exchange Within and Across Generations*, ed. A. Booth et al. Washington, D.C.: Urban Institute Press.

Freese, J., and Powell, B. (2001). Making love out of nothing at all? Null findings and the Trivers-Willard hypothesis. *American Journal of Sociology* **106**, 1776–1788.

Gelman, A. (2007a). Letter to the editor regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology* **245**, 597–599.

Gelman, A. (2007b). Scaling regression inputs by dividing by two standard deviations.

Technical report, Department of Statistics, Columbia University.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A., and Jakulin, A. (2007). Bayes: liberal, radical, or conservative? *Statistica Sinica* **17**, 422–426.

Gelman, A., and Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* **15**, 373–390.

Hamermesh, D. S., and Biddle, J. E. (1994). Beauty and the labor market. *American Economic Review* **84**, 1174–1194.

Hamermesh, D. S., and Parker, A. M. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review* **24**, 369–376.

James, W. H. (1987). The human sex ratio. I: a review of the literature. *Human Biology* **59**, 721–752.

Kanazawa, S. (2005). Big and tall parents have more sons: further generalizations of the Trivers-Willard hypothesis. *Journal of Theoretical Biology* **233**, 583–590.

Kanazawa, S. (2006). Violent men have more sons: further evidence for the generalized the Trivers-Willard hypothesis. *Journal of Theoretical Biology* **239**, 450–459.

Kanazawa, S. (2007). Beautiful parents have more daughters: a further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology.*

Kanazawa, S., and Vandermassen, G. (2005). Engineers have more sons, nurses have more daughters: an evolutionary psychological extension of Baron-Cohen's extreme male brain theory of autism. *Journal of Theoretical Biology* **233**, 589–599.

Mathews, F., Johnson, P. J., and Neil, A. (2008). You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. *Proceedings of the Royal Society B* **275**, 1661–1668.

Miller, A. S., and Kanazawa, S. (2007a). Ten politically incorrect truths about human nature. *Psychology Today*, July/August.

Miller, A. S., and Kanazawa, S. (2007b). *Why Beautiful People Have More Daughters.* New York: Perigee.

Morris, C. (1987). Comment on "Testing a point null hypothesis: the irreconcilability of p values and evidence," by J. O. Berger and T. Sellke. *Journal of the American Statistical*

Association **82**, 131–133.

Norberg, K. (2004). Partnership status and the human sex ratio at birth. *Proceedings of the Royal Society B* **271**, 2403–2410.

Open N.Y. (2008). The measure of a president. *New York Times* op-ed page, 6 Oct. `http://www.nytimes.com/interactive/2008/10/06/opinion/06opchart.html`

Oswald, A., and Powdthavee, N. (2006). Daughters and left-wing voting. Technical report, Department of Economics, University of Warwick, U.K.

Pinker, S. (2007). In defense of dangerous ideas. *Chicago Sun-Times*, 15 July.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science* **308**, 1623–1626.

Trivers, R. L, and Willard, D. E. (1973). Natural selection of parental ability to vary the sex ratio of offspring. *Science* **179**, 90–92.

Volscho, T. W. (2005). Money and sex, the illusory universal sex difference: comment on Kanazawa. *Sociological Quarterly* **46**, 716–736.

Wainer, H. (2007). The most dangerous equation. *American Scientist* **95**, 249–56.

Washington, E. L. (2007). Female socialization: how daughters affect their legislator fathers' voting on women's issues. *American Economic Review.*