# 1

# The Most Dangerous Equation

## 1.1. INTRODUCTION

What constitutes a dangerous equation? There are two obvious interpretations: some equations are dangerous if you know the equation and others are dangerous if you do not. In this chapter I will not explore the dangers an equation might hold because the secrets within its bounds open doors behind which lie terrible peril. Few would disagree that the obvious winner in this contest would be Einstein's iconic equation

$$E = MC^2 \tag{1.1}$$

for it provides a measure of the enormous energy hidden within ordinary matter. Its destructive capability was recognized by Leo Szilard, who then instigated the sequence of events that culminated in the construction of atomic bombs.

This is not, however, the direction I wish to pursue. Instead, I am interested in equations that unleash their danger, not when we know about them, but rather when we do not; equations that allow us to understand things clearly, but whose absence leaves us dangerously ignorant.* There are many plausible candidates that seem to fill the bill. But I feel that there is one that surpasses all others in the havoc wrought by ignorance of it over many centuries. It is the equation that provides us with the standard deviation of the sampling distribution of the mean; what might be called De Moivre's equation:

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}\,. \tag{1.2}$$

For those unfamiliar with De Moivre's equation let me offer a brief aside to explain it.

**An Aside Explaining De Moivre's Equation**

The Greek letter $\sigma$, with no subscript, represents a measure of the variability of a data set (its standard deviation). So if we measure, for example, the heights of, say, 1000 students at a particular high school, we might find that the average height is 67 inches, but heights might range from perhaps as little as 55 inches to as much as 80 inches. A number that characterizes this variation is the standard deviation. Under normal circumstances we would find that about two-thirds of all children in the sample would be within one standard deviation of the average. But now suppose we randomly grouped the 1000 children into 10 groups of 100 and calculated the average within each group. The variation of these 10 averages would likely be

(*Continued*)

* One way to conceive of the concept "danger" is as a danger function

$$P(Y) = P(Y|x=1)\,P(x=1) + P(Y|x=0)\,P(x=0),$$

where

$Y$ = the event of looking like an idiot,
$x = 1$ is the event of knowing the equation in question,
$x = 0$ is the event of not knowing the equation.
$|$ is mathematical notation meaning "given that" so that the expression $P(Y|x=1)$ is read as "the probability of looking like an idiot given that you know the equation."

This equation makes explicit the two aspects of what constitutes a dangerous equation. It also indicates that danger is a product of the inherent danger and the probability of that situation occurring. Thus it may be very dangerous not to know that the mean is the sum of the observations divided by $n$, but since most people know it, it is overall not a very dangerous equation.

(*Continued*)
much smaller than $\sigma$ because it is likely that a very tall person in the group would be counterbalanced by a shorter person. De Moivre showed a relationship between the variability in the original data and the variability in the averages. He also showed how one could calculate the variability of the averages ($\sigma_{\bar{x}}$) by simply dividing the original variability by the square root of the number of individuals ($n$) that were combined to calculate the averages. And so the variability of the average of groups of 100 would be one-tenth that of the original group. Similarly, if we wanted to reduce the variability in half, we would need groups of four; to cut it to one-fifth we would need groups of 25, and so forth. That is the idea behind De Moivre's equation.

Why have I decided to choose this simple equation as the most dangerous? There are three reasons, related to

(1) the extreme length of time during which ignorance of it has caused confusion,
(2) the wide breadth of areas that have been misled, and
(3) the seriousness of the consequences that such ignorance has caused.

In the balance of this chapter I will describe five very different situations in which ignorance of De Moivre's equation has led to billions of dollars of loss over centuries, yielding untold hardship; these are but a small sampling; there are many more.

### 1.2. THE TRIAL OF THE PYX: SIX CENTURIES OF MISUNDERSTANDING

In 1150, a century after the Battle of Hastings, it was recognized that the king could not just print money and assign to it any value he chose. Instead the coinage's value must be intrinsic, based on the amount of precious materials in its makeup. And so standards were set for the weight of gold in coins—a guinea should weigh 128 grains (there are 360 grains in an ounce). It was recognized, even then, that coinage methods were too imprecise to insist that all coins be exactly equal in weight, so instead the king and the barons, who supplied the London Mint (an independent organization) with gold, insisted that coins when tested* in the aggregate [say one hundred at a time] conform to the regulated size plus or minus some allowance for variability [1/400 of the weight] which for one guinea would be 0.32 grains and so, for the aggregate, 32 grains). Obviously, they assumed that variability decreased proportionally to the number of coins and not to its square root.

This deeper understanding lay almost six hundred years in the future with De Moivre's 1730 exploration of the binomial distribution.[1,†] The costs of making errors are of two types. If the average of all the coins was too light, the barons were being cheated, for there would be extra gold left over after minting the agreed number of coins. This kind of error would easily have been detected and, if found, the director of the Mint would suffer grievous punishment. But if the variability were too great, it would mean that there would be an unacceptably large number of too heavy coins produced that could be collected, melted down, and recast with the extra gold going into the pockets of the minter. By erroneously allowing too much variability, the Mint could stay within the bounds specified and still make extra money by collecting the heavier-than-average coins and reprocessing them. The fact that this error was able to continue for almost six hundred years provides strong support for De Moivre's equation to be considered a strong candidate for the title of most dangerous equation.

### 1.3. LIFE IN THE COUNTRY: A HAVEN OR A THREAT

Figure 1.1 is a map of age-adjusted kidney cancer rates. The counties shaded are those counties that are in the lowest decile of the cancer distribution. We note that these healthy counties tend to be very rural, midwestern, southern, and western counties. It is both easy and tempting to infer that this outcome is directly due to the clean living of the rural lifestyle—no air pollution, no water pollution, access to fresh food without additives, etc.

Figure 1.2 is another map of age adjusted kidney cancer rates.

* The box the coins were kept in was called the Pyx, and so each periodic test was termed the Trial of the Pyx.

† "De Moivre (1730) knew that root $n$ described the spread for the binomial, but did not comment more generally than that. Certainly by Laplace (1810) the more general version could be said to be known." (Stigler, personal communication, January 17, 2007.)

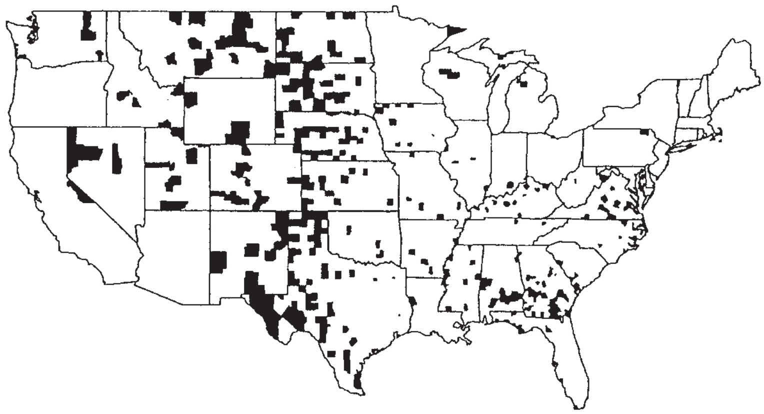

Figure 1.1.
Lowest kidney cancer death rates. The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/urethra for U.S. males, 1980–1989 (from Gelman and Nolan, 2002, p. 15, reprinted with permission).

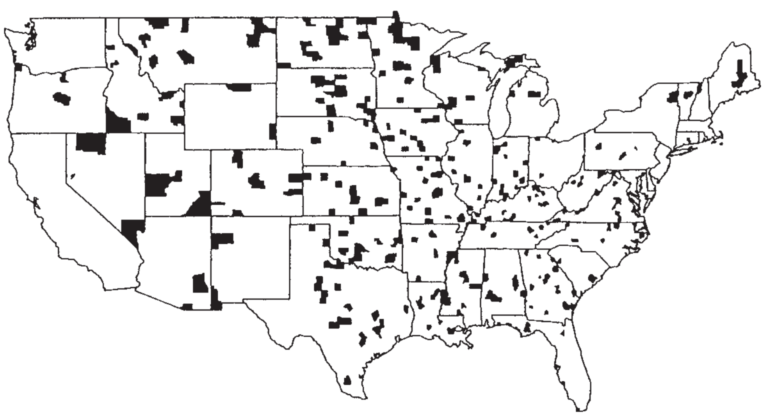

Figure 1.2.
Highest kidney cancer death rates. The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/urethra for U.S. males, 1980–1989 (from Gelman and Nolan, 2002, p. 14, reprinted with permission).

While it looks very much like figure 1.1, it differs in one important detail—the counties shaded are those counties that are in the *highest* decile of the cancer distribution. We note that these ailing counties tend to be very rural, midwestern, southern, and western counties. It is easy to infer that this outcome might be directly due to the poverty of the rural lifestyle—no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco.

If we were to plot figure 1.1 on top of figure 1.2 we would see that many of the shaded counties on one map are right next to the shaded counties in the other. What is going on? What we are seeing is De

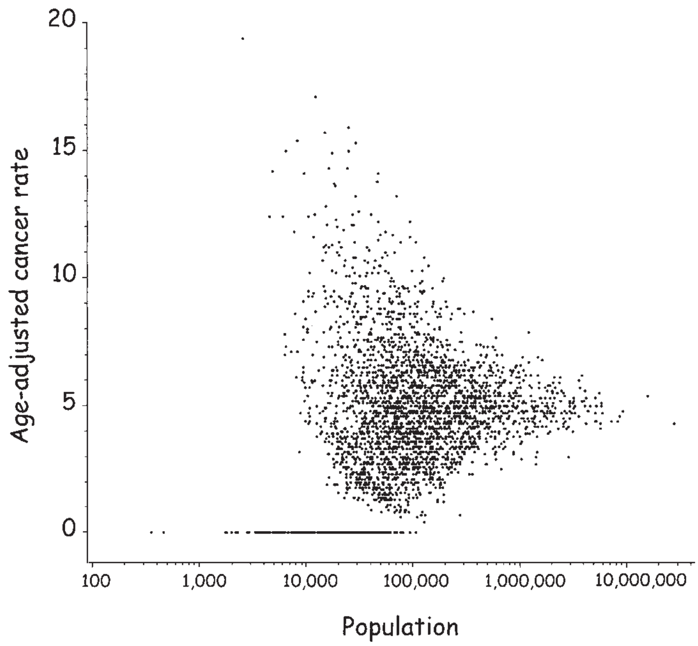


Figure 1.3.
Age-adjusted kidney cancer rates for all U.S. counties in 1980–1989 shown as a function of the log of the county population.

Moivre's equation in action. The variation of the mean is inversely proportional to the square root of the sample size, and so small counties have much larger variation than large counties. A county with, say one hundred inhabitants that has no cancer deaths would be in the lowest category. But if it has one cancer death it would be among the highest. Counties like New York or Los Angeles or Miami/Dade with millions of inhabitants do not bounce around like that.

If we plot the age-adjusted cancer rates against county population, this result becomes clearer still (figure 1.3). We see the typical triangular-shaped bivariate distribution in which when the population is small (left side of the graph) there is wide variation in cancer rates, from twenty per hundred thousand to zero. When county populations are large (right side of graph), there is very little variation, with all counties at about five cases per hundred thousand of population.

This is a compelling example of how someone who looked only at, say figure 1.1, and not knowing about De Moivre's equation might draw incorrect inferences and give incorrect advice (e.g., if you are at risk of kidney cancer, you should move to the wide open spaces of rural America). This would be dangerous advice, and yet this is precisely what was done in my third example.

### 1.4. THE SMALL SCHOOLS MOVEMENT: BILLIONS FOR INCREASING VARIANCE[2]

The urbanization that characterized the twentieth century yielded abandonment of the rural lifestyle and, with it, an increase in the size of schools. The time of one-room school houses ended; they were replaced by large schools, often with more than a thousand students, dozens of teachers of many specialties, and facilities that would not have been practical without the enormous increase in scale. Yet during the last quarter of the twentieth century[3] there were the beginnings of dissatisfaction with large schools, and the suggestion that smaller schools could provide better-quality education. Then in the late 1990s the Bill and Melinda Gates Foundation began supporting small schools on a broad-ranging, intensive, national basis. By 2001, the Foundation had given grants to education projects totaling approximately $1.7 billion. They have since been joined in support for smaller schools by the Annenberg Foundation, the Carnegie Corporation, the Center for Collaborative Education, the Center for School Change, Harvard's Change Leadership Group, Open Society Institute, Pew Charitable Trusts, and the U.S. Department of Education's Smaller Learning Communities Program. The availability of such large amounts of money to implement a smaller schools policy yielded a concomitant increase in the pressure to do so, with programs to splinter large schools into smaller ones being proposed and implemented broadly (e.g., New York City, Los Angeles, Chicago, and Seattle).

What is the evidence in support of such a change? There are many claims made about the advantages of smaller schools, but we will focus here on just one—that when schools are smaller, students' achievement improves. That is, the expected achievement in schools, given that they are small [$E(\text{achievement}|\text{small})$] is greater than what is expected if they are big [$E(\text{achievement}|\text{big})$]. Using this convenient mathematical notation,

$$E(\text{achievement}|\text{small}) > E(\text{achievement}|\text{big}), \quad (1.3)$$

all else being equal. But the supporting evidence for this is that, when one looks at high-performing schools, one is apt to see an unrepresentatively large proportion of smaller schools. Or, stated mathematically, that

$$P(\text{small}|\text{high achievement}) > P(\text{large}|\text{high achievement}). \quad (1.4)$$

Note that expression (1.4) does not imply (1.3).

In an effort to see the relationship between small schools and achievement we looked at the performance of all of Pennsylvania's public schools, as a function of their size, on the Pennsylvania testing program (PSSA), which is very broad and yields scores in a variety of subjects and over the entire range of precollegiate school years. If we examine the mean scores of the 1662 separate schools that provide fifth grade reading scores, we find that of the top-scoring fifty schools (the top 3%), six of them were among the smallest 3% of the schools. This is an overrepresentation by a factor of four. If size of school was unrelated to achievement, we would expect 3% of small schools to be in this select group, and we found 12%. The bivariate distribution of enrollment and test score is shown in figure 1.4. The top fifty schools are marked by a square.

We also identified the fifty lowest scoring schools, marked by an "o" in figure 1.4. Nine of these (18%) were among the fifty smallest schools. This result is completely consonant with what is expected from De Moivre's equation—the smaller schools are expected to have higher variance and hence should be overrepresented at both extremes. Note that the regression line shown on figure 1.4 is essentially flat, indicating that, overall, there is no apparent relationship between school size and performance. But this is not always true.

Figure 1.5 is a plot of eleventh grade scores. We find a similar overrepresentation of small schools on both extremes, but this time the regression line shows a significant positive slope; overall, students at bigger schools do better.

The small schools movement seems to have arrived at one of its recommendations through the examination of only one tail of the performance distribution. Small schools are overrepresented at both tails, exactly as expected, since smaller schools will show greater variation in performance and empirically will show up wherever we look. Our

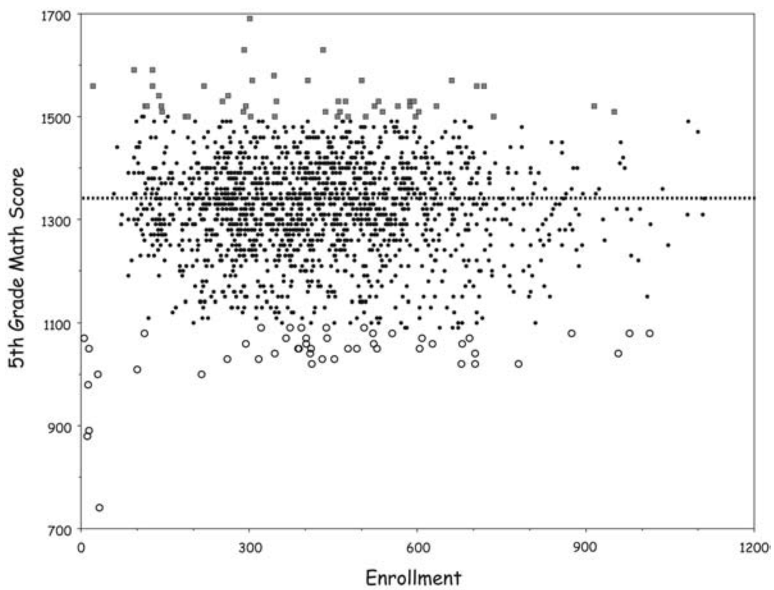


Figure 1.4.
Average score of fifth grade classes in mathematics shown as a function of school size.

examination of fifth grade performance suggests that school size alone seems to have no bearing on student achievement. This is not true at the high-school level, where larger schools show better performance. This too is not unexpected, since very small high schools cannot provide as broad a curriculum or as many highly specialized teachers as large schools. This was discussed anecdotally in a July 20, 2005, article in the *Seattle Weekly* by Bob Geballe. The article describes the conversion of Mountlake Terrace High School in Seattle from a large suburban school with an enrollment of 1,800 students into five smaller schools. The conversion was greased with a Gates Foundation grant of almost a million dollars. Although class sizes remained the same, each of the five schools had fewer teachers. Students complained, "There's just one English teacher and one math teacher . . . teachers ended up teaching things they don't really know." Perhaps this helps to explain the regression line in figure 1.5.

On October 26, 2005, Lynn Thompson, in an article in the *Seattle Times* reported that "The Gates Foundation announced last week it is moving away from its emphasis on converting large high schools into smaller ones and instead giving grants to specially selected school

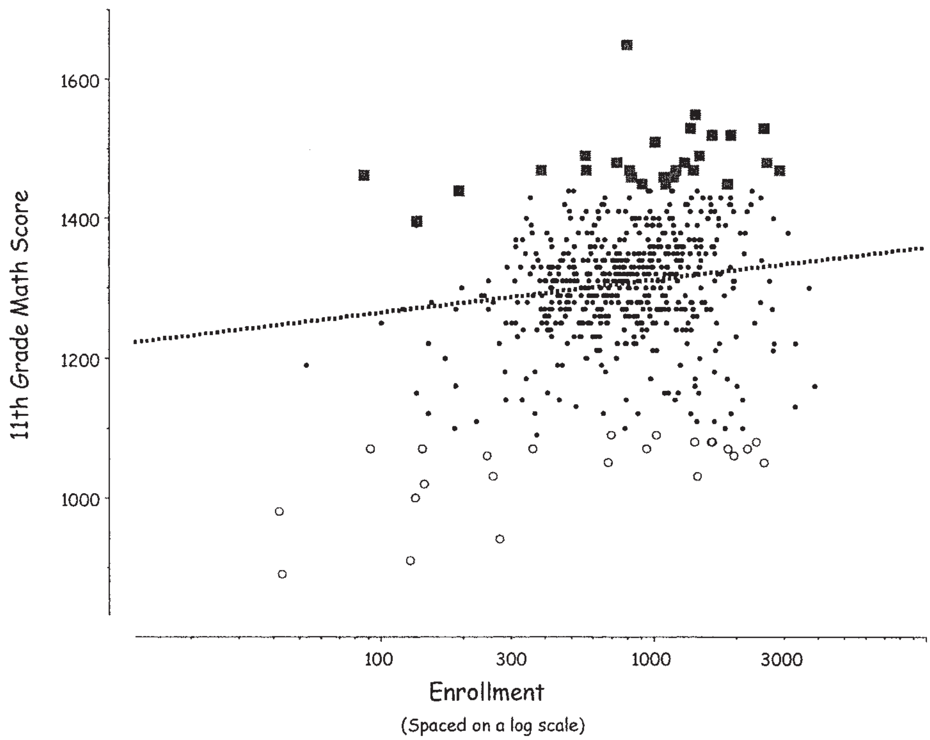


Figure 1.5.
Eleventh grade student scores on a statewide math test shown as a function of school size.

districts with a track record of academic improvement and effective leadership. Education leaders at the Foundation said they concluded that improving classroom instruction and mobilizing the resources of an entire district were more important first steps to improving high schools than breaking down the size." This point of view was amplified in a study that carefully analyzed matched students in schools of varying sizes. The lead author concluded, "I'm afraid we have done a terrible disservice to kids."[4]

Expending more than a billion dollars on a theory based on ignorance of De Moivre's equation suggests just how dangerous that ignorance can be.

### 1.5. THE SAFEST CITIES

The *New York Times* recently reported[5] the ten safest U.S. cities and the ten most unsafe based on an automobile insurance company statistic, the "average number of years between accidents." The cities were drawn from the two hundred largest cities in the U.S. It should come as no surprise that a list of the ten safest cities, the ten most dangerous cities, and the ten largest cities have no overlap (see table 1.1).

Exactly which cities are the safest, and which the most dangerous, must surely depend on many things. But it would be difficult (because of De Moivre's equation) for the largest cities to be at the extremes. Thus we should not be surprised that the ends of the safety distribution are anchored by smaller cities.

### 1.6. SEX DIFFERENCES

For many years it has been well established that there is an overabundance of boys at the high end of test score distributions. This has meant that about twice as many boys as girls received Merit Scholarships and other highly competitive awards. Historically, some observers have used such results to make inferences about differences in intelligence between the sexes. Over the last few decades, however, most enlightened investigators have seen that it is not necessarily a difference in level but a difference in variance that separates the sexes. The public observation of this fact has not been greeted gently; witness the recent outcry when Harvard's (now ex-)president Lawrence Summers pointed this out:

> It does appear that on many, many, different human attributes—height, weight, propensity for criminality, overall IQ, mathematical ability, scientific ability—there is relatively clear evidence that whatever the difference in means—which can be debated—there is a difference in standard deviation/variability of a male and female population. And it is true with respect to attributes that are and are not plausibly, culturally determined.* (Summers [2005])

The boys' score distributions are almost always characterized by greater variance than the girls'. Thus, while there are more boys at the high end, there are also more at the low end.

* Informal remarks made by Lawrence Summers in July 2005 at the National Bureau of Economic Research Conference on Diversifying the Science and Engineering Workforce.

Table 1.1

**Information on Automobile Accident Rates in 20 Cities**

| City | State | Population rank | Population | Number of years between accidents |
|---|---|---|---|---|
| **Ten safest** | | | | |
| Sioux Falls | South Dakota | 170 | 133,834 | 14.3 |
| Fort Collins | Colorado | 182 | 125,740 | 13.2 |
| Cedar Rapids | Iowa | 190 | 122,542 | 13.2 |
| Huntsville | Alabama | 129 | 164,237 | 12.8 |
| Chattanooga | Tennessee | 138 | 154,887 | 12.7 |
| Knoxville | Tennessee | 124 | 173,278 | 12.6 |
| Des Moines | Iowa | 103 | 196,093 | 12.6 |
| Milwaukee | Wisconsin | 19 | 586,941 | 12.5 |
| Colorado Springs | Colorado | 48 | 370,448 | 12.3 |
| Warren | Michigan | 169 | 136,016 | 12.3 |
| **Ten least safe** | | | | |
| Newark | New Jersey | 64 | 277,911 | 5.0 |
| Washington | DC | 25 | 563,384 | 5.1 |
| Elizabeth | New Jersey | 189 | 123,215 | 5.4 |
| Alexandria | Virginia | 174 | 128,923 | 5.7 |
| Arlington | Virginia | 114 | 187,873 | 6.0 |
| Glendale | California | 92 | 200,499 | 6.1 |
| Jersey City | New Jersey | 74 | 239,097 | 6.2 |
| Paterson | New Jersey | 148 | 150,782 | 6.5 |
| San Francisco | California | 14 | 751,682 | 6.5 |
| Baltimore | Maryland | 18 | 628,670 | 6.5 |
| **Ten biggest** | | | | |
| New York | New York | 1 | 8,085,742 | 8.4 |
| Los Angeles | California | 2 | 3,819,951 | 7.0 |
| Chicago | Illinois | 3 | 2,869,121 | 7.5 |
| Houston | Texas | 4 | 2,009,690 | 8.0 |
| Philadelphia | Pennsylvania | 5 | 1,479,339 | 6.6 |
| Phoenix | Arizona | 6 | 1,388,416 | 9.7 |
| San Diego | California | 7 | 1,266,753 | 8.9 |
| San Antonio | Texas | 8 | 1,214,725 | 8.0 |
| Dallas | Texas | 9 | 1,208,318 | 7.3 |
| Detroit | Michigan | 10 | 911,402 | 10.4 |

Source: Data from the *NY Times* and www.allstate.com/media/newsheadlines. From Wainer, 2007d.

Table 1.2

**Eighth Grade NAEP National Results: Summary of Some Outcomes, by Sex, from National Assessment of Educational Progress**

| | | *Mean scale scores* | | *Standard deviations* | | *Male/female* |
|---|---|---|---|---|---|---|
| *Subject* | *Year* | *Male* | *Female* | *Male* | *Female* | *ratio* |
| Math | 1990 | 263 | 262 | 37 | 35 | 1.06 |
| | 1992 | 268 | 269 | 37 | 36 | 1.03 |
| | 1996 | 271 | 269 | 38 | 37 | 1.03 |
| | 2000 | 274 | 272 | 39 | 37 | 1.05 |
| | 2003 | 278 | 277 | 37 | 35 | 1.06 |
| | 2005 | 280 | 278 | 37 | 35 | 1.06 |
| Science | 1996 | 150 | 148 | 36 | 33 | 1.09 |
| | 2000 | 153 | 146 | 37 | 35 | 1.06 |
| | 2005 | 150 | 147 | 36 | 34 | 1.06 |
| Reading | 1992 | 254 | 267 | 36 | 35 | 1.03 |
| | 1994 | 252 | 267 | 37 | 35 | 1.06 |
| | 1998 | 256 | 270 | 36 | 33 | 1.09 |
| | 2002 | 260 | 269 | 34 | 33 | 1.03 |
| | 2003 | 258 | 269 | 36 | 34 | 1.06 |
| | 2005 | 257 | 267 | 35 | 34 | 1.03 |
| Geography | 1994 | 262 | 258 | 35 | 34 | 1.03 |
| | 2001 | 264 | 260 | 34 | 32 | 1.06 |
| US History | 1994 | 259 | 259 | 33 | 31 | 1.06 |
| | 2001 | 264 | 261 | 33 | 31 | 1.06 |

Source: http://nces.ed.gov/nationsreportcard/nde/

An example, chosen from the National Assessment of Educational Progress (NAEP), is shown in table 1.2. NAEP is a true survey and so problems of self-selection (rife in college entrance exams, licensing exams, etc.) are substantially reduced. The data summarized in table 1.2 were accumulated over fifteen years and five subjects. In all instances the standard deviation of males is from 3% to 9% greater than that of females. This is true when males score higher on average (math, science, geography) or lower (reading).

Both inferences, the incorrect one about differences in level and the correct one about differences in variability, cry out for explanation.

The old cry would have been "why do boys score higher than girls?"; the newer one, "why do boys show more variability?" If one did not know about De Moivre's result and tried to answer only the first question, it would be a wild goose chase, trying to find an explanation for a phenomenon that did not exist. But, if we focus on why variability is greater in males, we may find pay dirt. Obviously the answer to the causal question "why?" will have many parts. Surely socialization and differential expectations must be major components—especially in the past, before the realization grew that a society cannot compete effectively in a global economy with only half its workforce fully mobilized. But there is another component that is key—and especially related to the topic of this chapter, De Moivre's equation.

In discussing Lawrence Summers' remarks about sex differences in scientific ability, Christiane Nüsslein-Volhard, the 1995 Nobel Laureate in Physiology/Medicine,* said

> He missed the point. In mathematics and science, there is no difference in the intelligence of men and women. The difference in genes between men and women is simply the Y chromosome, which has nothing to do with intelligence. (Dreifus [July 4, 2006])

But perhaps it is Professor Nüsslein-Volhard who missed the point here. The *Y* chromosome is not the only difference between the sexes. Summers' point was that, when we look at either extreme of an ability distribution, we will see more of the group that has greater variation. Any mental trait that is conveyed on the *X* chromosome will have larger variability among males than females, for females have two *X* chromosomes versus only one for males. Thus, from De Moivre's equation, we would expect, *ceteris paribus*, about 40% more variability† among males than females. The fact that we see less than 10% greater variation in NAEP demands the existence of a deeper explanation. First, De Moivre's equation requires independence of the two *X*'s, and

* Dr. Nüsslein-Volhard shared the prize with Eric F. Wieschaus for their work that showed how the genes in a fertilized egg direct the formation of an embryo.

† Actually the square root of two (1.414. . .) is more, hence my approximation of 40% is a tad low; it would be more accurate to have said 41.4%, but "40%" makes the point.

with assortative mating this is not going to be true. Additionally, both $X$ chromosomes are not expressed in every cell. Moreover, there must be major causes of high-level performance that are not carried on the $X$ chromosome, and indeed are not genetic. But it suggests that for some skills between 10% and 25% of the increased variability is likely to have had its genesis on the $X$ chromosome. This observation would be invisible to those, even those with Nobel prizes for work in genetics, who are in ignorance of De Moivre's equation.

It is well established that there is evolutionary pressure toward greater variation within species—within the constraints of genetic stability. This is evidenced by the dominance of sexual over asexual reproduction among mammals. But this leaves us with a puzzle. Why was our genetic structure built to yield greater variation among males than females? And not just among humans, but virtually all mammals. The pattern of mating suggests an answer. In most mammalian species that reproduce sexually, essentially all adult females reproduce, whereas only a small proportion of males do so (modern humans excepted). Think of the alpha-male lion surrounded by a pride of females, with lesser males wandering aimlessly and alone in the forest roaring in frustration. One way to increase the likelihood of off-spring being selected to reproduce is to have large variance among them. Thus evolutionary pressure would reward larger variation for males relative to females.

This view gained further support in studies by Arthur Arnold and Eric Vilain of UCLA that were reported by Nicholas Wade of the *New York Times* on April 10, 2007. He wrote,

> It so happens that an unusually large number of brain-related genes are situated on the $X$ chromosome. The sudden emergence of the $X$ and $Y$ chromosomes in brain function has caught the attention of evolutionary biologists. Since men have only one $X$ chromosome, natural selection can speedily promote any advantageous mutation that arises in one of the $X$'s genes. So if those picky women should be looking for smartness in prospective male partners, that might explain why so many brain-related genes ended up on the $X$.

He goes on to conclude,

> Greater male variance means that although average IQ is identical in men and women, there are fewer average men and more at both extremes. Women's care in selecting mates, combined with the fast selection made possible by men's lack of backup copies of *X*-related genes, may have driven the divergence between male and female brains.

## 1.7. CONCLUSION

It is no revelation that humans don't fully comprehend the effect that variation, and especially differential variation, has on what we observe. Daniel Kahneman's 2002 Nobel Prize was for his studies on intuitive judgment (which occupies a middle ground "between the automatic operations of perception and the deliberate operations of reasoning" [6]. Kahneman showed that humans don't intuitively "know" that smaller hospitals will have greater variability in the proportion of male to female births. But such inability is not limited to humans making judgments in psychology experiments. Small hospitals are routinely singled out for special accolades because of their exemplary performance, only to slip toward average in subsequent years. Explanations typically abound that discuss how their notoriety has overloaded their capacity. Similarly, small mutual funds are recommended, *post hoc*, by Wall Street analysts, only to have their subsequent performance disappoint investors. The list goes on and on, adding evidence and support to my nomination of De Moivre's equation as the most dangerous of them all.

This chapter has been aimed at reducing the peril that accompanies ignorance of De Moivre's equation and also at showing why an understanding of variability is critical if we are to avoid serious errors.