# MODERN SCIENCE AND THE BAYESIAN-FREQUENTIST CONTROVERSY

Bradley Efron

**Abstract**

The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having important practical consequences. Whenever noisy data is a major concern, scientists depend on statistical inference to pursue nature's mysteries. 19th Century science was broadly Bayesian in its statistical methodology, while frequentism dominated 20th Century scientific practice. This brings up a pointed question: which philosophy will predominate in the 21st Century? One thing is already clear – statistical inference will pay an increased role in scientific progress as scientists attack bigger, messier problems in biology, medicine, neuroscience, the environment, and other fields that have resisted traditional deterministic analyses. This talk argues that a combination of frequentist and Bayesian thinking will be needed to deal with the massive data sets scientists are now bringing us. Three examples are given to suggest how such combinations might look in practice. A large portion of the talk is based on my presidential address to the American Statistical Association and a related column in Amstat News.

Autumn in Stanford brings with it two notable natural phenomena: the days get short and it starts to rain a lot. These are both "scientific facts" but they involve quite different kinds of science. Shorter days exemplify hard-edged science, precise and so predictable that you can sell an almanac saying exactly how short each day will be, down to the nearest second. The almanacs try to predict rainfall too but they're not nearly so successful. Rainfall is a famously random phenomenon, as centuries of unhappy farmers can testify. (My father's almanac went further, predicting good or bad fishing weather for each day, indicated by a full fish icon, an empty fish, or a half fish for borderline days.)

Hard-edged science still dominates public perceptions, but the attention of modern scientists has swung heavily toward rainfall-like subjects, the kind where random behavior plays a major role. A cartoon history of western thought might recognize three eras: an unpredictable pre-scientific world ruled by willful gods and magic; the precise clockwork universe of Newton and Laplace; and the modern scientific perspective of an understandable world, but one where predictability is tempered by a heavy dose of randomness. Deterministic Newtonian science is majestic, and the basis of modern science too, but a few hundred years of it pretty much exhausted nature's storehouse of precisely predictable events. Subjects like biology, medicine, and economics require a more flexible scientific world view, the kind we statisticians are trained to understand.

These thoughts were very much in my mind at "Phystat2003", a conference of particle physicists and statisticians held at the Stanford Linear Accelerator Center las year. It's at least slightly amazing to me that the physicists, who were the convening force, were eager to confer with us. One can't imagine "Phystat1903", back when the physics world disdained statistics. "If your experiment needs statistics you ought to have done a better experiment" in Lord Rutherford's infamous words.

Rutherford lived in a rich man's world of scientific experimentation, where nature generously provided boatloads of data, enough for the law of large numbers to squelch any noise. Nature has gotten more tight-fisted with modern physicists. They are asking harder questions, ones where the data is thin on the ground, and where efficient inference becomes a necessity. In short, they have started playing in our ball park.

The question of greatest interest at Phystat2003 concerned the mass of the neutrino, a famously elusive particle that is much lighter than an electron, and may weigh almost nothing at all. Heroic experiments, involving house-sized vats of cleaning fluid in abandoned mine shafts, yielded only a few dozen or a few hundred neutrinos. This left lots of room for experimental noise, and in fact the best unbiased estimate of neutrino mass turned out to be

negative. Mass itself can't be negative of course. Given a negative estimate, the physicists wished to establish a statistical upper bound for the true mass, the smaller the better from the point of view of further experimentation. As a result the particle physics literature now contains a healthy debate on Bayesian versus frequentist ways of setting the bound. The current favorite is the "Feldman-Cousins" method, developed by two prominent physicists, a likelihood-ratio based system of one-sided confidence intervals.

The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there's only one way of doing physics but there seems to be at least two ways to do statistics, and they don't always give the same answers. This says something about the special nature of our field. Most scientists study some aspect of nature, rocks, stars, particles; we study scientists, or at least scientific data. Statistics is an information science, the first and most fully developed information science. Maybe it's not surprising then that there is more than one way to think about an abstract subject like "information".

The Bayesian-Frequentist debate reflects two different attitudes to the process of doing science, both quite legitimate. Bayesian statistics is well-suited to individual researchers, or a research group, trying to use all the information at its disposal to make the quickest possible progress. In pursuing progress, Bayesians tend to be aggressive and optimistic with their modeling assumptions. Frequentist statisticians are more cautious and defensive. One definition says that a frequentist is a Bayesian trying to do well, or at least not too badly, against any possible prior distribution. The frequentist aims for universally acceptable conclusions, ones that will stand up to adversarial scrutiny. The FDA for example doesn't care about Pfizer's prior opinion of how well it's new drug will work, it wants objective proof. Pfizer, on the other hand may care very much about its own opinions in planning future drug development.

Bayesians excel at combining information from different sources, "coherence" being the technical word for correct combination. On the other hand, a common frequentist tactic is to pull problems apart, focusing for the sake of objectivity on a subset of the data that can be analyzed optimally. I'll show examples of both tactics soon.

Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist. What's going to happen in the 21st Century? One thing that's already happening is that scientists are bringing statisticians much bigger data sets to analyze, with millions of data points and thousands of parameters to consider all at once. Microarrays, a favorite technology of modern bioscience, are the poster boy for scientific giganticism.

Classical statistics was fashioned for small problems, a few hundred data points at most, a few parameters. Some new thinking is definitely called for on our part. I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning. I'm going to argue that in some ways huge data sets are actually easier to handle for both schools of thought.

Here's a real-life example I used to illustrate Bayesian virtues to the physicists. A physicist friend of mine and her husband found out, thanks to the miracle of sonograms, that they were going to have twin boys. One day at breakfast in the student union she suddenly asked me what was the probability that the twins would be identical rather than fraternal. This seemed like a tough question, especially at breakfast. Stalling for time, I asked if the doctor had given her any more information. "Yes", she said, "he told me that the proportion of identical twins was one third". This is the population proportion of course, and my friend wanted to know the probability that *her* twins would be identical.

Bayes would have lived in vain if I didn't answer my friend using Bayes' rule. According to the doctor the prior odds ratio of identical to nonidentical is one-third to two-thirds, or one half. Because identical twins are always the same sex but fraternal twins are random, the likelihood ratio for seeing "both boys" in the sonogram is a factor of two in favor of Identical. Bayes' rule says to multiply the prior odds by the likelihood ratio to get the current odds: in this case 1/2 times 2 equals 1; in other words, equal odds on identical or nonidentical given the sonogram results. So I told my friend that her odds were 50-50 (wishing the answer had come out something else, like 63-37, to make me seem more clever.) Incidentally, the twins are a couple of years old now, and "couldn't be more nonidentical" according to their mom.

Now Bayes rule is a very attractive way of reasoning, and fun to use, but using Bayes rule doesn't make one a Bayesian. *Always* using Bayes rule does, and that's where the practical difficulties begin: the kind of expert opinion that gave us the prior odds one-third to two-thirds usually doesn't exist, or may be controversial or even wrong. The likelihood ratio can cause troubles too. Typically the numerator is easy enough, being the probability of seeing the data at hand given our theory of interest; but the denominator refers to probabilities under Other theories, which may not be clearly defined in our minds. This is why Bayesians have to be such aggressive math modelers. Frequentism took center stage in the 20th Century in order to avoid all this model specification.

Figure 1 concerns a more typical scientific inference problem, of the sort that is almost always handled frequentistically these days. It involves a breast cancer study that attracted national attention when it appeared in the New England Journal of Medicine in 2001. Dr.

Hedenfalk and his associates were studying two genetic mutations that each lead to increased breast cancer risk, called "BRCA1" and "BRCA2" by geneticists. These are different mutations on different chromesomes. Hedenfalk et al wondered if the tumors resulting from the two different mutations were themselves genetically different.

• BRCA1 (7 Tumors)

**-1.29 -1.41 -0.55 -1.04 1.28 -0.27 -0.57**

• BRACA2 (8 Tumors)

**-0.70 1.33 1.14 4.67 0.21 0.65 1.02 0.16**

**Figure 1**: *Expression data for the first of 3226 genes, microarray study of breast cancer; Hedenfalk et al. (2001).*

To answer the question they took tumor material from 15 breast cancer patients, seven from women with the BRCA1 mutation and eight with BRCA2. A separate microarray was developed for each of the 15 tumors, each microarray having the same 3226 genes on it. Here we only see the data for the first gene: 7 genetic activity numbers for the BRCA1 cases and 8 activity numbers for the BRCA2's. These numbers don't have much meaning individually, even for microbiologists, but they can be compared with each other statistically. The question of interest is "are the expression levels different for BRCA1 compared to BRCA2?" It looks like this gene might be more active in the BRCA2 tumors, since those 8 numbers are mostly positive , while 6 of the 7 BRAC1's are negative.

A standard frequentist answer to this question uses Wilcoxon's nonparametric two-sample test. (which amounts to the usual $t$-test except with ranks replacing the original numbers.) We order the 15 expression values from smallest to largest and compute "$W$", the sum of ranks for the BRCA2 values. The biggest W could be is 92, if all 8 BRCA2 numbers were larger than all 7 BRCA1's; at the opposite end of the scale, if the 8 BRCA2's were all smaller than the 7 BRCA1's, we'd get $W = 36$. For the gene 1 data we actually get $W = 83$, which looks pretty big. It *is* big, by the usual frequentist criterion. Its two-sided $p$-value, the probability of getting a W at least this extreme, is only .024 under the null hypothesis that there is no real expression difference. We'd usually put a star next to .024 to indicate significance, according to Fisher's famous .05 cutoff point. Notice that this analysis requires very little from the statistician, no prior probabilities or likelihoods, only the specification of a null hypothesis. It's no wonder that hypothesis testing is wildly popular with scientists, and has been for a hundred years.

The .05 significance cutoff has been used literally millions of times since Fisher proposed it in the early Nineteen Hundreds. It has become a standard of objective comparison in all areas of science. I don't think that .05 could stand up to such intense use if it wasn't producing basically correct scientific inferences most of the time. But .05 was intended to apply to a single comparison, not 3226 comparisons at once.

I computed W for each of the 3226 genes in the BRCA microarray data. The histogram in Figure 2 shows the results, which range from 8 genes with the smallest possible $W$, $W = 36$, to 7 with $W = 92$, the largest possible, and with all intermediate values represented many times over. There's more about the analysis of this data set in Efron (2004).

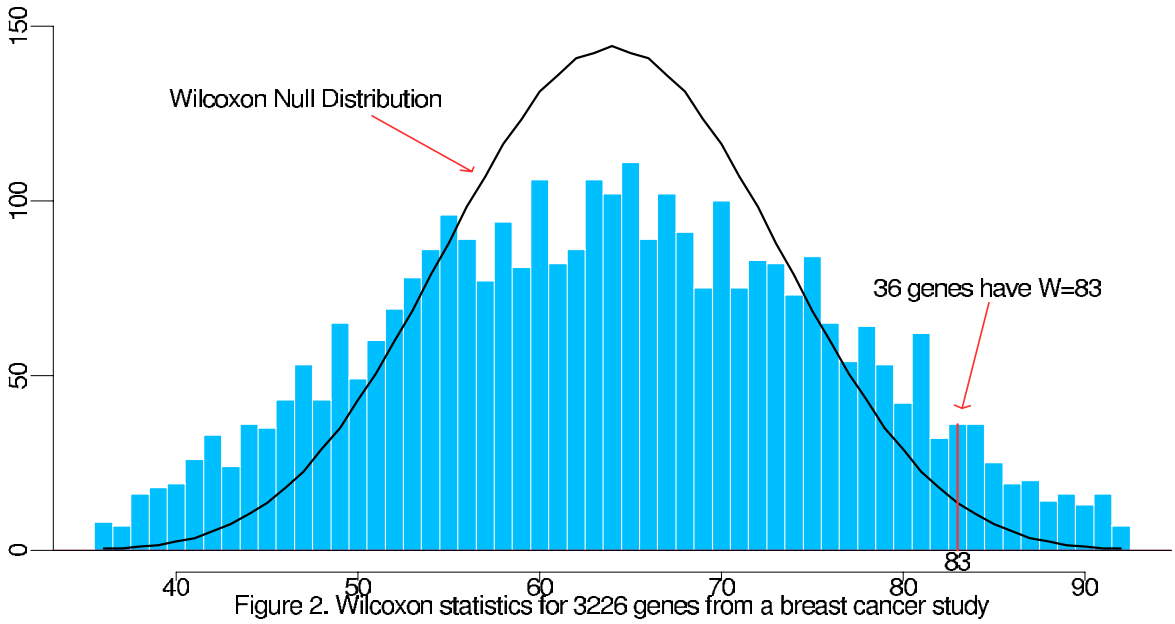


**Figure 2**: *Wilcoxon statistics for 3226 genes from a breast cancer study*

It looks like something is definitely going on here. The histogram is much wider than the theoretical Wilcoxon null density (the smooth curve) that would apply if none of the genes behaved differently for BRCA1 vs BRCA2. 580 of the genes, 18% of them, achieve significance according to the usual one-at-a-time .05 criterion. That's a lot more than the null hypothesis 5%, but now it isn't so clear how to assess significance for any one gene given so many candidates. Does the $W = 83$ we saw for gene 1 really indicate significance?

I was saying earlier that huge data sets are in some ways easier to analyze than the small one's we've been used to. Here is a big-data-set kind of answer to assessing the significance of the observed Wilcoxon value $W = 83$ : 36 of the 3226 genes, gene 1 and 35 others, have $W = 83$; under the null hypothesis that there's no real difference between BRCA1

and BRCA2 expressions, we would expect to see only 9 genes with $W = 83$. Therefore the expected false discovery rate is 9 out of 36, or 25%. If Hedenfalk decides to investigate gene 1 further he has a 25% chance of wasting his time. Investigator time is usually precious, so he might prefer to focus attention on those genes with more extreme W values, having smaller false discovery rates. For example there are 13 genes with $W = 90$, and these have a false discovery rate of only 8%.

The "9 out of 36" calculation looks definitely frequentist. In fact the original false discovery rate theory developed by Benjamini and Hochberg in 1995 was phrased entirely in frequentist terms, not very much different philosophically than Fisher's .05 cutoff or Neyman-Pearson testing. Their work is a good example of the kind of "new theory" that I hope statisticians will be developing in response to the challenge of massive data sets.

It turns out that the false discovery rate calculations also have a very nice Bayesian rationale. We assume that a priori a proportion $p_0$ of the genes are null, and that these genes have W's following the null Wilcoxon density $f_0(w)$. In a usual microarray experiment we'd expect most of the genes to be null, with $p_0$ no smaller than say 90%. The remainder of the genes are non-null, and follow some other density, let's call it $f_1(w)$, for their Wilcoxon scores. These are the "interesting genes", the ones we want to identify and report back to the investigators. If we know $p_0, f_0$, and $f_1$, then Bayes rule tells us right away what the probability is of a gene being null or non-null, given it's Wilcoxon score W.

The catch is that to actually carry out Bayes rule we need to know the prior quantities $p_0, f_0(w)$, and $f_1(w)$. This looks pretty hopeless without an alarming amount of prior modeling and guesswork. But an interesting thing happens with a large data set like this one: we can use the data to estimate the prior quantities, and then use these estimates to approximate Bayes rule. When we do so the answer turns out much the same as before, for example null probability 9 out of 36 given $W = 83$.

This is properly called an "empirical Bayes" approach. Empirical Bayes estimates combine the the two statistical philosophies: the prior quantities are estimated frequentistically in order to carry out Bayesian calculations. Empirical Bayes analysis goes back to Robbins and Stein in the 1950's, but they were way ahead of their time. The kind of massively parallel data sets that really benefit from empirical Bayes analysis seem to be much more a 21st Century phenomenon.

The BRCA data set is big by classical standards, but it is big in an interesting way: it repeats the same "small" data structure again and again, so we are presented with 3226 similar two-sample comparisons. This kind of parallel structure gives the statistician a terrific

advantage, it's just what we need to bring empirical Bayes methods to bear. Statisticians are not passive observers of the scientific scene. The fact that we can successfully analyze ANOVA problems leads scientists to plan their experiments in ANOVA style. In the same way we can influence the design of big data sets by demonstrating impressively successful analyses of parallel structures.

We have a natural advantage here: it's a lot easier to manufacture high-throughput devices if they have a parallel design. The familiar medical breakthrough story on TV, showing what looks like a hundred eyedroppers squirting at once, illustrates parallel design in action. Microarrays, flow cytometry, proteomics, time-of-flight spectroscopy, all refer to machines of this sort that are going to provide us with huge data sets nicely suited for empirical Bayes methods.

Figure 3 shows another example. It concerns an experiment comparing 7 normal children with 7 dyslexic kids. A diffusion tensor imaging scan (related to fMRI scanning) was done for each child, providing measurements of activity at 16000 locations in the brain. At each of these locations a two-sample $t$-test was performed comparing the normal and dyslexic kids. The figure shows the signs of the $t$-statistics for 580 of the positions on one horizontal slice of the brain scan. (There are 40 other slices with similar pictures.) Squares indicate positive $t$-statistics, x's negative, with filled in squares indicating values exceeding two; these are positions that would be considered significantly different between the two groups by the standard .05 one-at-a-time criterion.

We can use the same false discovery rate empirical Bayes analysis here, with one important difference: the geometry of the brain scan lets us see the large amount of spatial correlation. Better results are obtained by averaging the data over small contiguous blocks of brain position, better in the sense of giving more cases with small false discovery rates. The best way of doing so is one of those interesting questions raised by the new technology.

There's one last thing to say about my false discovery rate calculations for the BRCA data: they may not be right! At first glance the "9 out of 36 equals 25% false discoveries" argument looks too simple to be wrong. The "9" in the numerator, which comes from Wilcoxon's null hypothesis distribution, is the only place where any theory is involved. But that's where potential trouble lies. If we only had one gene's data, say for gene 1 as before, we would *have* to use the Wilcoxon null, but with thousands of genes to consider at once, most of which are probably null, we can empirically estimate the null distribution itself. Doing so gives far fewer significant genes in this case, as you can read about in Efron (2004). Estimating the null hypothesis itself from the data sounds a little crazy, but that's what I

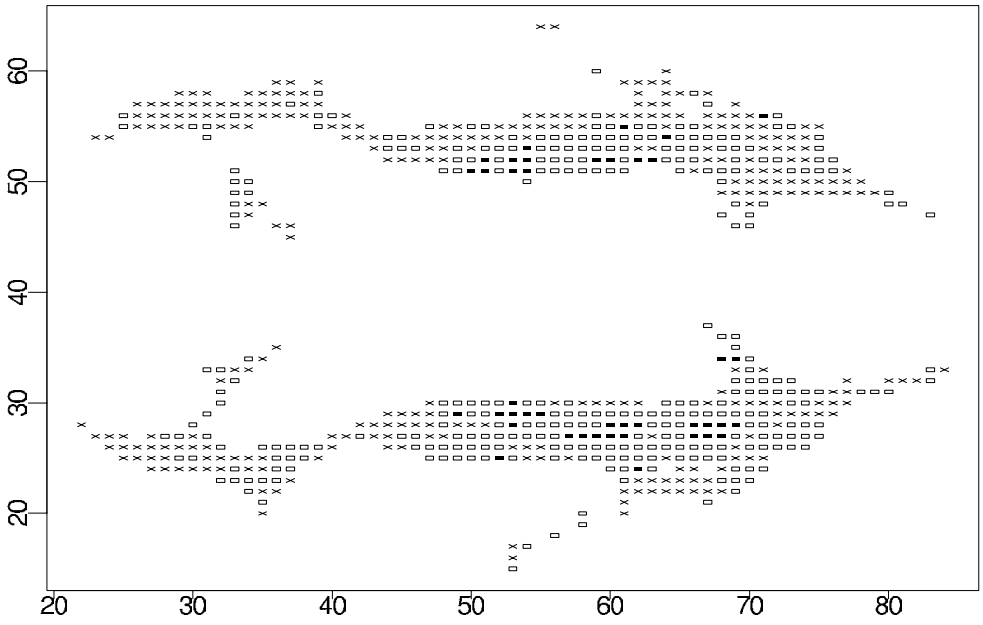


**Figure 3**: *580 t-statistics from a brain imaging study of dyslexia; solid squares $t \geq 2$; empty squares $t \geq 0$; x's $t < 0$. From data in Schwartzman, Dougherty, and Taylor (2004).*

meant about huge data sets presenting new opportunities as well as difficulties.

How could Wilcoxon's null hypothesis possibly go astray? There is a simple answer in this case: There was dependence *across* microarrays. The empirical Bayes analysis doesn't need independence *within* a microarray, but the validity of Wilcoxon's null requires independence across any one gene's 15 measurements. For some reason this wasn't time for the BRCA experiment, particularly the BRCA2 data. Table 1 shows that the BRCA2 measurements were substantially correlated in groups of four.

I have to apologize for going on so long about empirical Bayes, which has always been one of my favorite topics, and now at last seems to be going from ugly duckling to swan in the world of statistical applications. Here is another example of Bayesian-frequentist convergence, equally dear to my heart.

Figure 4 tells the unhappy story of how people's kidneys get worse as they grow older. The 157 dots represent 157 healthy volunteers, with the horizontal axis their age and the vertical axis a measure of total kidney function. I've used the "lowess" curve-fitter, a complicated sort of robust moving average, to summarize the decline of kidney function with age. The fitted curve goes steadily downward except for a plateau in the 20's.

How accurate is the lowess fit? This is one of those questions that has gone from hopeless

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.02 | 0.02 | 0.23 | -0.36 | -0.35 | -0.39 | -0.34 |
| 2 | 0.02 | 1.00 | 0.10 | -0.08 | -0.30 | -0.30 | -0.23 | -0.33 |
| 3 | 0.02 | 0.10 | 1.00 | -0.17 | -0.21 | -0.26 | -0.31 | -0.27 |
| 4 | 0.23 | -0.08 | -0.17 | 1.00 | -0.30 | -0.23 | -0.27 | -0.32 |
| 5 | -0.36 | -0.30 | -0.21 | -0.30 | 1.00 | -0.02 | 0.11 | 0.22 |
| 6 | -0.35 | -0.30 | -0.26 | -0.23 | -0.02 | 1.00 | 0.15 | 0.13 |
| 7 | -0.39 | -0.23 | -0.31 | -0.27 | 0.11 | 0.15 | 1.00 | 0.07 |
| 8 | -0.34 | -0.33 | -0.27 | -0.32 | 0.22 | 0.13 | 0.07 | 1.00 |

**Table 1**: Correlations of the 8 BRCA2 microarray measurements (after first subtracting each gene's mean value).

to easy with the advent of high-speed computation. A simple bootstrap analysis gives the answer in, literally, seconds. We resample the 157 points, that is, take a random sample of 157 points with replacement from the original 157 (so some of the original points appear once, twice, three times or more, and others don't appear at all in the resample.) Then the lowess curve-fitter is applied to the resampled data set, giving a bootstrap version of the original curve.

In Figure 5 I've repeated the whole process 100 times, yielding 100 bootstrap lowess curves. Their spread gives a quick and dependable picture of the statistical variability in the original curve. For instance we can see that the variability is much greater near the high end of the age scale, at the far right, than it is in the plateau.

The bootstrap was originally developed as a purely frequentist device. Nevertheless the bootstrap picture has a Bayesian interpretation: if we could put an "uninformative" prior on the collection of possible age-kidney curves, that is, a prior that reflects a lack of specific opinions, then the resulting Bayes analysis would tend to agree with the bootstrap distribution. The bootstrap-objective Bayes relationship is pursued in Efron and Tibshirani (1998).

This brings up an important trend in Bayesian statistics. Objectivity is one of the principal reasons that frequentism dominated 20th century applications; a frequentist method like Wilcoxon's test, which is completely devoid of prior opinion, has a clear claim to being objective – a crucial fact when scientists communicate with their skeptical colleagues. Uninformative priors, the kind that also have a claim to objectivity, are the Bayesian response.

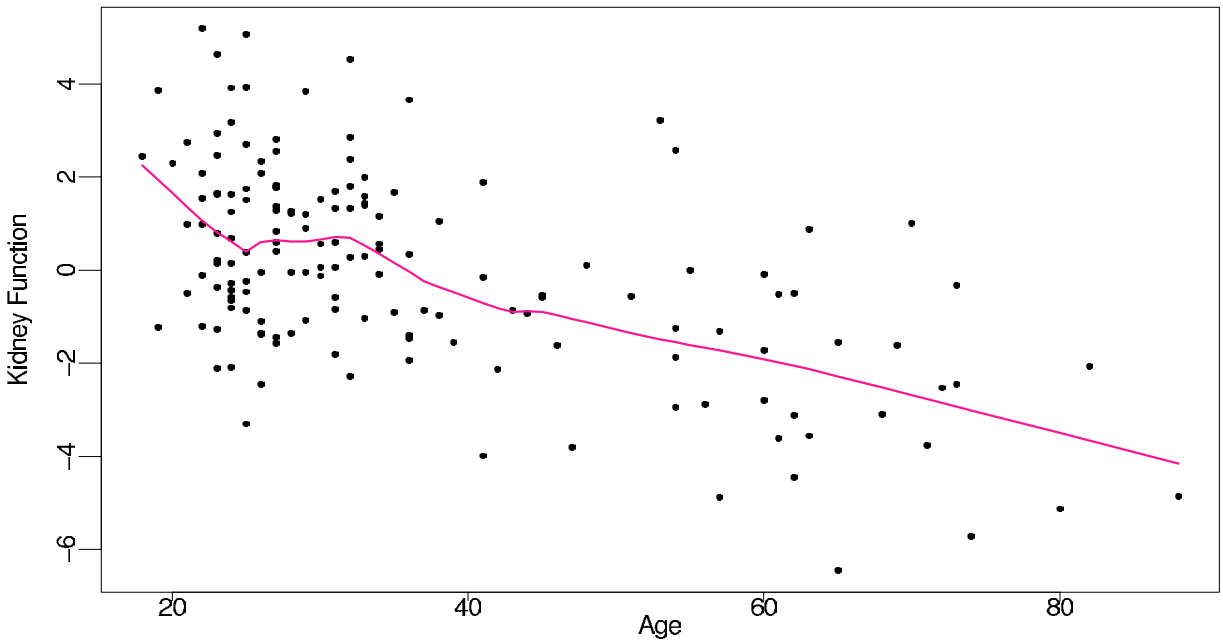


**Figure 4**: *Kidney function versus age for 157 normal volunteers, and lowess fit*

Bayesian statistics has seen a strong movement away from subjectivity and towards objective uninformative priors in the past 20 years.

Technical improvements, the computer implementation of Markov Chain Monte Carlo methods, have facilitated this trend, but the main reason I believe is a desire to compete with frequentism in the domain of real-world applications. Whatever the reason, the effect has been to bring Bayesian and frequentist practice closer together.

In practice it isn't easy to specify an uninformative prior, especially in messy-looking problems like choosing a possibly jagged regression curve. What looks uninformative enough often turns out to subtly force answers in one direction or another. The bootstrap connection is intriguing because it suggest a simple way of carrying out a genuinely objective Bayesian analysis, but this is only a suggestion so far.

Perhaps I've let my enthusiasm for empirical Bayes and the bootstrap run away with the main point I started out to make. The bottom line is that we have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of classical statistics. In the struggle to find these answers the statistics profession needs to use both frequenstist and Bayesian ideas, and new combinations of the two. Moreover I think this is already beginning to happen... which was the real point of the examples.

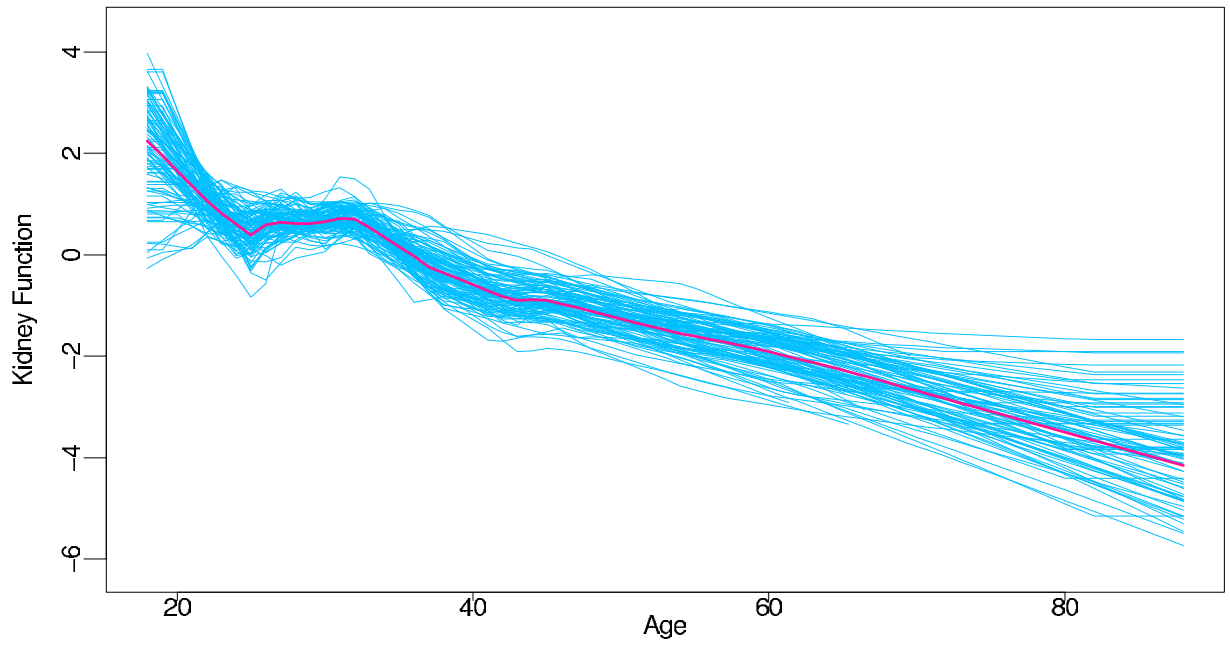


**Figure 5**: *100 Bootstrap replication of lowess fit*

It took enormously long for the statistical point of view to develop. Two thousand years separate Aristotelian logic from Bayes theorem, its natural probabilistic extension. Another 150 years went past before Fisher, Neyman, and other frequentists developed a statistical theory satisfactory for general scientific use in situations where Bayes theorem is difficult to apply. The truth is that statistical reasoning does not come naturally to the human brain. We are cause and effect thinkers, ideal perhaps for avoiding the jaws of the sabre-toothed tiger, but less effective in dealing with partial correlation or regression to the mean.

Once it caught on though, statistical reasoning proved to be a scientific success story. Starting from just about zero in 1900, statistics spread from one field to another, becoming the dominant mode of quantitative thinking in literally dozens of fields, from agriculture, education, and psychology to medicine, biology, and economics, with the hard sciences knocking on our door now. A graph of statistics during the 20th Century shows a steadily rising curve of activity and influence. Statisticians, a naturally modest bunch, tend to think of their field as a small one, but it is a discipline with a long arm, reaching into almost every area of science and social science these days.

Our curve has taken a bend upwards in the 21st Century. A new generation of scientific devices, typified by microarrays, produce data on a gargantuan scale – with millions of data points and thousands of parameters to consider at the same time. These experiments are "deeply statistical". Common sense, and even good scientific intuition, won't do the job

by themselves. Careful statistical reasoning is the only way to see through the haze of randomness to the structure underneath. Massive data collection, in astronomy, psychology, biology, medicine, and commerce, is a fact of 21st Century science, and a good reason to buy statistics futures if they are ever offered on the NASDAQ.

I find the microarray story particularly encouraging for statistics. The first fact is that the biologists *did* come to us for answers to the inference problems raised by their avalanche of microarray data. This is our payoff for being helpful colleagues in the past, doing all those ANOVAs, $t$-tests, and randomized clinical trials that have become a standard part of biomedical research. And indeed we seem to be helping again, providing a solid set of new analytic tools for microarray experiments. The benefit goes both ways. Microarrays are helping out inside our field too, raising difficult new problems in large-scale simultaneous inference, stimulating a new burst of methodology and theory, and refocusing our attention on underdeveloped areas like empirical Bayes.

Ken Alder's 2002 book, "The Measure of All Things", brilliantly relates the story of the meter, one ten-millionth the distance from the equator to the pole, and how its length was determined in post-revolutionary France. Most of the book concerns the difficulties of the "savants" in carrying out their arduous astronomical-geographical measurements. One savant, Pierre Mechain, couldn't quite reconcile his readings, and wound up fudging the answers, driving himself to near-madness and death.

Near the conclusion of "Measure" Alder suddenly springs his main point, forgiving Mechain as laboring under an obsolete, overly precise, notion of scientific reality:

"Approach the world instead through the veil of uncertainty and science would never be the same. And nor would savants. During the course of the next century science learned to manage uncertainty. The field of statistics that would one day emerge from the insights to Legendre, Laplace, and Gauss would transform the physical sciences, inspire the biological sciences, and give birth to the social sciences. In the process 'savants' became scientists."

Right on, Ken! Alder's new world of science has been a long time emerging but there is no doubt that 21st Century scientists are committed to the statistical point of view. This puts the pressure on us, the statisticians, to fulfill our end of the bargain. We have been up to the task in the past and I suspect we will succeed again, though it may take a couple more Fishers, Neymans, and Walds to do the trick.

## References


Adler, K. (2002). "The measure of all things: the seven-year odyssey that transformed the world", *Little-Brown*, New York.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Jour. Royal Stat. Soc. B* **57**, 289-300.

Efron, B. (2005). "Bayesians, Frequentists, and Scientists". To appear *Journ. American Statistical Association* **100**.

Efron, B. (2004). "Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis", *JASA* **99**, 96-104.

Efron, B. (2004). "Life in a random universe", *Amstat News* **330**, 2-3.

Efron, B. (2003). "Robbins, Empirical Bayes, and Microarrays", *Annals Stat.* **24**, 366-378.

Hedenfalk, I., Duggen, D., Chen, Y. et al. (2001). "Gene Expression Profiles in Hereditary Breast Cancer", *New England Journal of Medicine* **344**, 539-48.

Schwartzman, A., Dougherty, R., Taylor, J. (2005), "Cross-Subject Comparison of Principal Diffusion Direction Maps". To appear *Magnetic Resonance in Medicine*, armins@stanford.edu.

However, both Bayesian and frequentist statisticians have expanded their epistemological basis away from a singular focus on the null hypothesis, to a broader perspective involving the development and comparison of competing statistical/mathematical models. For frequentists, statistical developments such as structural equation modeling and multilevel modeling have facilitated thisâ€¦ CONTINUE READING. View on Taylor & Francis.Â  Modern Science And The Bayesian-Frequentist Controversy. A Frequentist-leaning data scientist might say a Bayesian approach is too subjective because a prior must be selected, while a Bayesian-leaning data scientist might say a Frequentist approach is foolish to ignore our knowledge about the world and focus solely on counts. At Glossier, we hoped to settle this internal debate by simulating a series of experiments to test how using Bayesian and Frequentist A/B test evaluation methods affected our ability to detect small improvements in metrics. All else equal â€” sample size, baseline metric, difference in variants etc. â€” which method would best dete The Bayesian-Frequentist argument is more applicable regarding the choice of the variables to be tested in the A/B paradigm, but even there most A/B testers violate the hell out of research hypotheses, probability, and confidence intervals.â€ Further reading: Bayesian A/B Testing by Evan Miller.Â  Though you could dig forever and find strong arguments for and against each side, it comes down to this: Weâ€™re solving the same problem in two ways. I like the analogy that Optimizely gave using bridges: Just like a suspension and arch bridges both successfully get cars across a gap, both Bayesian and Frequentist statistical methods provide to an answer to the question: which variation performed best in an A/B test? Anderson also had a fun way of looking at it