Module 6: Probabilistic Reasoning Through the Basic Sampling Model

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

– Sir Francis Galton (*Natural Inheritance*, 1889)

Abstract: One mechanism for assisting in various tasks encountered in probabilistic reasoning is to adopt a simple sampling model. A population of interest is first posited, characterized by some random variable, say X. This random variable has a population distribution (often assumed to be normal), characterized by (unknown) parameters. The sampling model posits n independent observations on X, denoted by $X_1, \ldots, X_n$, and which constitutes the sample. Various functions of the sample can then be constructed (that is, various statistics can be computed such as the sample mean and sample variance); in turn, statistics have their own sampling distributions. The general problem of statistical inference is to ask what sample statistics tell us about their population counterparts; for example, how can we construct a confidence interval for a population parameter such as the population mean from the sampling distribution for the sample mean.

Under the framework of a basic sampling model, a number of topics

are discussed: confidence interval construction for a population mean where the length of the interval is determined by the square root of the sample size; the Central Limit Theorem and the Law of Large Numbers; the influence that sample size and variability have on our probabilistic reasoning skills; the massive fraud case involving the Dutch social psychologist, Diederik Stapel, and the role that lack of variability played in his exposure; the ubiquitous phenomenon of regression toward the mean and the importance it has for many of our probabilistic misunderstandings; how reliability corrections can be incorporated into prediction; the dichotomy and controversy encountered every ten years about complete enumeration versus sampling (to correct for, say, an undercount) in the United States Census.

Contents

1 The Basic Sampling Model and Associated Topics 3

2 Regression Toward the Mean 12

3 Incorporating Reliability Corrections in Prediction 16

4 Complete Enumeration versus Sampling in the Census 24

5 Appendix: Brief for American Statistical Association as Amicus Curiae, Department of Commerce v. United States House of Representatives 26

6 Appendix: Department of Commerce v. United States House of Representatives 27

1 The Basic Sampling Model and Associated Topics

We begin by refreshing our memories about the distinctions between *population* and *sample*, *parameters* and *statistics*, and *population distributions* and *sampling distributions*. Someone who has successfully completed a first course in statistics should know these distinctions well. Here, only a simple univariate framework is considered explicitly, but an obvious and straightforward generalization exists for the multivariate context as well.

A *population* of interest is posited, and operationalized by some random variable, say X. In this *Theory World* framework, X is characterized by *parameters*, such as the expectation of X, $\mu = \mathrm{E}(X)$, or its variance, $\sigma^2 = \mathrm{V}(X)$. The random variable X has a *(population) distribution*, which is often assumed normal. A *sample* is generated by taking observations on X, say, $X_1, \ldots, X_n$, considered independent and identically distributed as X; that is, they are exact copies of X. In this *Data World* context, statistics are functions of the sample and therefore characterize the sample: the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$; the sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$, with some possible variation in dividing by $n-1$ to generate an unbiased estimator for σ^2. The statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, are *point estimators* of μ and σ^2. They are random variables by themselves, so they have distributions referred to as *sampling distributions*. The general problem of statistical inference is to ask what sample statistics, such as $\hat{\mu}$ and $\hat{\sigma}^2$, tell us about their population counterparts,

μ and σ^2. In other words, can we obtain a measure of accuracy for estimation from the sampling distributions through, for example, confidence intervals?

Assuming that the population distribution is normally distributed, the sampling distribution of $\hat{\mu}$ is itself normal with expectation μ and variance σ^2/n. Based on this result, an approximate 95% confidence interval for the unknown parameter μ can be given by

$$\hat{\mu} \pm 2.0\frac{\hat{\sigma}}{\sqrt{n}} .$$

Note that it is the square root of the sample size that determines the length of the interval (and not the sample size per se). This is both good news and bad. Bad, because if you want to double precision, you need a fourfold increase in sample size; good, because sample size can be cut by four with only a halving of precision.

Even when the population distribution is not originally normally distributed, the central limit theorem (CLT) (that is, the "Law of Frequency of Error," as noted by the opening epigram for this module) says that $\hat{\mu}$ is approximately normal in form and becomes exactly so as n goes to infinity. Thus, the approximate confidence interval statement remains valid even when the underlying distribution is not normal. Such a result is the basis for many claims of robustness; that is, when a procedure remains valid even if the assumptions under which it was derived may not be true, as long as some particular condition is satisfied; here, the condition is that the sample size be reasonably large.

Besides the robustness of the confidence interval calculations for μ, the CLT also encompasses the law of large numbers (LLN). As the

sample size increases, the estimator, $\hat{\mu}$, gets closer to μ, and converges to μ at the limit as n goes to infinity. This is seen most directly in the variance of the sampling distribution for $\hat{\mu}$, which becomes smaller as the sample size gets larger.

The basic results obtainable from the CLT and LLN that averages are both less variable and more normal in distribution than individual observations, and that averages based on larger sample sizes will show less variability than those based on smaller sample sizes, have far-ranging and sometimes subtle influences on our probabilistic reasoning skills. For example, suppose we would like to study organizations, such as schools, health care units, or governmental agencies, and have a measure of performance for the individuals in the units, and the average for each unit. To identify those units exhibiting best performance (or, in the current jargon, "best practice"), the top 10%, say, of units in terms of performance are identified; a determination is then made of what common factors might characterize these top-performing units. We are pleased when we are able to isolate one very salient feature that most units in this top tier are small. We proceed on this observation and advise the breaking up of larger units. Is such a policy really justified based on these data? Probably not, if one also observes that the bottom 10% are also small units. That smaller entities tend to be more variable than the larger entities seems to vitiate a recommendation of breaking up the larger units for performance improvement. Evidence that the now-defunct "small schools movement," funded heavily by the Gates Foundation, was a victim of the "square root of n law" was presented by Wainer (2009, pp. 11–14).

Sports is an area in which there is a great misunderstanding and lack of appreciation for the effects of randomness. A reasonable model for sports performance is one of "observed performance" being the sum of "intrinsic ability" (or true performance) and "error," leading to a natural variability in outcome either at the individual or the team level. Somehow it appears necessary for sports writers, announcers, and other pundits to give reasons for what is most likely just random variability. We hear of team "chemistry," good or bad, being present or not; individuals having a "hot hand" (or a "cold hand," for that matter); someone needing to "pull out of a slump"; why there might be many .400 hitters early in the season but not later; a player being "due" for a hit; free-throw failure because of "pressure"; and so on. Making decisions based on natural variation being somehow "predictive" or "descriptive" of the truth, is not very smart, to say the least. But it is done all the time—sports managers are fired and CEOs replaced for what may be just the traces of natural variability.

People who are asked to generate random sequences of numbers tend to underestimate the amount of variation that should be present; for example, there are not enough longer runs and a tendency to produce too many short alternations. In a similar way, we do not see the naturalness in regression toward the mean (discussed in the next section of this module), where extremes are followed by less extreme observations just because of fallibility in observed performance. Again, causes are sought. We hear about multi-round golf tournaments where a good performance on the first day is followed by a less adequate score the second (due probably to "pressure"); or a bad performance on the first day followed by an improved perfor-

mance the next (the golfer must have been able to "play loose"). Or in baseball, at the start of a season an underperforming Derek Jeter might be under "pressure" or too much "media scrutiny," or subject to the difficulties of performing in a "New York market." When individuals start off well but then appear to fade, it must be because people are trying to stop them ("gunning" for someone is a common expression). One should always remember that in estimating intrinsic ability, individuals are unlikely to be as good (or as bad) as the pace they are on. It is always a better bet to vote against someone eventually breaking a record, even when they are "on a pace" to so do early in the season. This may be one origin for the phrase "sucker bet"—a gambling wager where your expected return is significantly lower than your bet.

Another area where one expects to see a lot of anomalous results is when the dataset is split into ever-finer categorizations that end up having few observations in them, and thus subject to much greater variability. For example, should we be overly surprised if Albert Pujols doesn't seem to bat well in domed stadiums at night when batting second against left-handed pitching? The pundits look for "causes" for these kinds of extremes when they should just be marveling at the beauty of natural variation and the effects of sample size. A similar and probably more important misleading effect occurs when our data are on the effectiveness of some medical treatment, and we try to attribute positive or negative results to ever-finer-grained classifications of the clinical subjects.

Random processes are a fundamental part of nature and ubiquitous in our day-to-day lives. Most people do not understand them,

or worse, fall under an "illusion of control" and believe they have influence over how events progress. Thus, there is an almost mystical belief in the ability of a new coach, CEO, or president to "turn things around." Part of these strong beliefs may result from the operation of regression toward the mean or the natural unfolding of any random process. We continue to get our erroneous beliefs reconfirmed when cause is attributed when none may actually be present. As humans we all wish to believe we can affect our future, but when events have dominating stochastic components, we are obviously not in complete control. There appears to be a fundamental clash between our ability to recognize the operation of randomness and the need for control in our lives.

An appreciation for how random processes might operate can be helpful in navigating the uncertain world we live in. When investments with Bernie Madoff give perfect 12% returns, year after year, with no exceptions and no variability, alarms should go off. If we see a supposed scatterplot of two fallible variables with a least-squares line imposed but where the actual data points have been withdrawn, remember that the relationship is not perfect. Or when we monitor error in quality assurance and control for various manufacturing or diagnostic processes (for example, application of radiation in medicine), and the tolerances become consistently beyond the region where we should generally expect the process to vary, a need to stop and recalibrate may be necessary. It is generally important to recognize that data interpretation may be a long-term process, with a need to appreciate variation appearing around a trend line. Thus, the immediacy of some major storms does not vitiate a longer-term perspective on global climate change. Remember the old meteorological adage:

climate is what you expect; weather is what you get. Relatedly, it is important to monitor processes we have some personal responsibility for (such as our own lipid panels when we go for physicals), and to assess when unacceptable variation appears outside of our normative values.

Besides having an appreciation for randomness in our day-to-day lives, there is also a flip side: if you don't see randomness when you probably should, something is amiss. The Bernie Madoff example noted above is a salient example, but there are many such deterministic traps awaiting the gullible. When something seems just too good to be true, most likely it isn't. A recent ongoing case in point involves the Dutch social psychologist, Diederik Stapel, and the massive fraud he committed in the very best psychology journals in the field. A news item by G. Vogel in *Science* (2011, *334*, 579) has the title, "Psychologist Accused of Fraud on 'Astonishing Scale'." Basically, in dozens of published articles and doctoral dissertations he supervised, Stapel never failed to obtain data showing the clean results he expected to see at the outset. As any practicing researcher in the behavioral sciences knows, this is just too good to be true. We give a short quotation from the *Science* news item (October 31, 2011) commenting on the Tilberg University report on the Stapel affair (authored by a committee headed by the well-known Dutch psycholinguist, Willem Levelt):

> Stapel was "absolute lord of the data" in his collaborations ... many of Stapel's datasets have improbable effect sizes and other statistical irregularities, the report says. Among Stapel's colleagues, the description of data as too good to be true "was a heartfelt compliment to his skill and creativity."

The report discusses the presence of consistently large effects being found; few missing data and outliers; hypotheses rarely refuted. Journals publishing Stapel's articles did not question the omission of details about the source of the data. As understated by Levelt, "We see that the scientific checks and balances process has failed at several levels." In a related article in the *New York Times* by Benedict Carey (November 2, 2011), "Fraud Case Seen as a Red Flag for Psychology Research," the whole field of psychology is now taken to task, appropriately we might add, in how research has generally been done and evaluated in the field. Part of the Levelt Committee report that deals explicitly with data and statistical analysis is redacted below:

The data were too good to be true; the hypotheses were almost always confirmed; the effects were improbably large; missing data, or impossible, out-of-range data, are rare or absent.

This is possibly the most precarious point of the entire data fraud. Scientific criticism and approach failed on all fronts in this respect. The falsification of hypotheses is a fundamental principle of science, but was hardly a part of the research culture surrounding Mr. Stapel. The only thing that counted was verification. However, anyone with any research experience, certainly in this sector, will be aware that most hypotheses that people entertain do not survive. And if they do, the effect often vanishes with replication. The fact that Mr. Stapel's hypotheses were always confirmed should have caused concern, certainly when in most cases the very large "effect sizes" found were clearly out of line with the literature. Rather than concluding that this was all improbable, instead Mr. Stapel's experimental skills were taken to be phenomenal. "Too good to be true" was meant as a genuine compliment to his skill and creativity. Whereas all these excessively neat findings should have provoked thought, they were embraced. If other researchers had failed, that was assumed to be because of a lack of preparation, insight, or experimental skill. Mr. Stapel became the model: the standard. Evidently only Mr. Stapel

was in a position to achieve the precise manipulations needed to make the subtle effects visible. People accepted, if they even attempted to replicate the results for themselves, that they had failed because they lacked Mr. Stapel's skill. However, there was usually no attempt to replicate, and certainly not independently. The few occasions when this did happen, and failed, were never revealed, because the findings were not publishable.

In other words, scientific criticism has not performed satisfactorily on this point. Replication and the falsification of hypotheses are cornerstones of science. Mr. Stapel's verification factory should have aroused great mistrust among colleagues, peers and journals.

As a supervisor and dissertation advisor, Mr. Stapel should have been expected to promote this critical attitude among his students. Instead, the opposite happened. A student who performed his own replications with no result was abandoned to his fate rather than praised and helped.

Strange, improbable, or impossible data patterns; strange correlations; identical averages and standard deviations; strange univariate distributions of variables.

The actual data displayed several strange patterns that should have been picked up. The patterns are related to the poor statistical foundation of Mr. Stapel's data fabrication approach (he also tended to make denigrating remarks about statistical methods). It has emerged that some of the fabrication involved simply "blindly" entering numbers based on the desired bivariate relationships, and by cutting and pasting data columns. This approach sometimes gave rise to strange data patterns. Reordering the data matrix by size of a given variable sometimes produces a matrix in which one column is identical to another, which is therefore the simple result of cutting and pasting certain scores. It was also possible for a variable that would normally score only a couple of per cent "antisocial," for no reason and unexpectedly suddenly to show "antisocial" most of the time. Independent replication yielded exactly the same averages and standard deviations. Two independent variables that always correlated positively, conceptually and in other research, now each had the right expected effects on the dependent

variable, but correlated negatively with each other. There was no consistent checking of data by means of simple correlation matrices and univariate distributions. It is to the credit of the whistle blowers that they did discover the improbabilities mentioned above.

Finally, a lamentable element of the culture in social psychology and psychology research is for everyone to keep their own data and not make them available to a public archive. This is a problem on a much larger scale, as has recently become apparent. Even where a journal demands data accessibility, authors usually do not comply ... Archiving and public access to research data not only makes this kind of data fabrication more visible, it is also a condition for worthwhile replication and meta-analysis. (pp. 13-15)

2 Regression Toward the Mean

Regression toward the mean is a phenomenon that will occur whenever dealing with fallible measures with a less-than-perfect correlation. The word "regression" was first used by Galton in his 1886 article, "Regression Towards Mediocrity in Hereditary Stature." Galton showed that heights of children from very tall or short parents regress toward mediocrity (that is, toward the mean) and exceptional scores on one variable (parental height) are not matched with such exceptionality on the second (child height). This observation is purely due to the fallibility for the various measures and the concomitant lack of a perfect correlation between the heights of parents and their children.

Regression toward the mean is a ubiquitous phenomenon, and given the name "regressive fallacy" whenever cause is ascribed where none exists. Generally, interventions are undertaken if processes are at an extreme (for example, a crackdown on speeding or drunk driv-

ing as fatalities spike, treatment groups formed from individuals who are seriously depressed, or individuals selected because of extreme good or bad behaviors). In all such instances, whatever remediation is carried out will be followed by some lessened value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened; I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (also known as the "sophomore jinx"); although he hadn't had a hit in some time, he was "due," and the coach played him; and so on. More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness of fit or prediction calculated need to be cross-validated either on new data or by a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures. The degree of "shrinkage" seen in our measures based on this cross-validation is an indication of the fallibility of our measures and the (in)adequacy of the given sample sizes.

The misleading interpretive effects engendered by regression toward the mean are legion, particularly when we wish to interpret observational studies for some indication of causality. There is a continual violation of the traditional adage that "the rich get richer and the poor get poorer," in favor of "when you are at the top, the only way is down." Extreme scores are never quite as extreme as they first appear. Many of these regression artifacts are discussed in

the cautionary source, *A Primer on Regression Artifacts* (Campbell & Kenny, 1999), including the various difficulties encountered in trying to equate intact groups by matching or analysis of covariance. Statistical equating creates the illusion but not the reality of equivalence. As summarized by Campbell and Kenny, "the failure to understand the likely direction of bias when statistical equating is used is one of the most serious difficulties in contemporary data analysis" (p. 85).

The historical prevalence of the regression fallacy is considered by Stephen Stigler in his 1997 article entitled "Regression Towards the Mean, Historically Considered" (*Statistical Methods in Medical Research*, *6*, 103–114). Stigler labels it "a trap waiting for the unwary, who were legion" (p. 112). He relates a story that we excerpt below about a Northwestern University statistician falling into the trap in 1933:

> The most spectacular instance of a statistician falling into the trap was in 1933, when a Northwestern University professor named Horace Secrist unwittingly wrote a whole book on the subject, *The Triumph of Mediocrity in Business*. In over 200 charts and tables, Secrist "demonstrated" what he took to be an important economic phenomenon, one that likely lay at the root of the great depression: a tendency for firms to grow more mediocre over time. Secrist was aware of Galton's work; he cited it and used Galton's terminology. The preface even acknowledged "helpful criticism" from such statistical luminaries as HC Carver (the editor of the *Annals of Mathematical Statistics*), Raymond Pearl, EB Wilson, AL Bowley, John Wishart and Udny Yule. How thoroughly these statisticians were informed of Secrist's work is unclear, but there is no evidence that they were successful in alerting him to the magnitude of his folly (or even if they noticed it). Most of the reviews of the book applauded it. But there was one dramatic exception: in late 1933 Harold Hotelling wrote a devastating review, noting among other things that

"The seeming convergence is a statistical fallacy, resulting from the method of grouping. These diagrams really prove nothing more than that the ratios in question have a tendency to wander about." (p. 112)

Stigler goes on to comment about the impact of the Secrist-Hotelling episode for the recognition of the importance of regression toward the mean:

One would think that so public a flogging as Secrist received for his blunder would wake up a generation of social scientists to the dangers implicit in this phenomenon, but that did not happen. Textbooks did not change their treatment of the topic, and if there was any increased awareness of it, the signs are hard to find. In the more than two decades between the Secrist-Hotelling exchange in 1933 and the publication in 1956 of a perceptively clear exposition in a textbook by W Allen Wallis and Harry Roberts, I have only encountered the briefest acknowledgements. (p. 113)

A variety of phrases seem to get attached whenever regression toward the mean is probably operative. We have the "winner's curse," where someone is chosen from a large pool (such as of job candidates), who then doesn't live up to expectations; or when we attribute some observed change to the operation of "spontaneous remission." As Campbell and Kenny noted, "many a quack has made a good living from regression toward the mean." Or, when a change of diagnostic classification results upon repeat testing for an individual given subsequent one-on-one tutoring (after being placed, for example, in a remedial context). More personally, there is "editorial burn-out" when someone is chosen to manage a prestigious journal at the apex of a career, and things go quickly downhill from that point.

3 Incorporating Reliability Corrections in Prediction

As discussed in the previous section, a recognition of when regression toward the mean might be operative can assist in avoiding the "regressive fallacy." In addition to this cautionary usage, the same regression-toward-the-mean phenomenon can make a positive contribution to the task of prediction with fallible information, and particularly in how such prediction can be made more accurate by correcting for the unreliability of the available variables. To make the argument a bit more formal, we assume an implicit underlying model for how any observed score, X, might be constructed additively from a true score, T_X, and an error score, E_X, where E_X is typically considered uncorrelated with T_X: $X = T_X + E_X$. The distribution of the observed variable over, say, a population of individuals, involves two sources of variability in the true and the error scores. If interests center on structural models among true scores, some correction should be made to the observed variables because the common regression models implicitly assume that all variables are measured without error. But before "errors-in-variables" models are briefly discussed, our immediate concern will be with how best to predict a true score from the observed score.[1]

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920s (Kelley,

[1]When an observed score is directly used as a prediction for the true score, the prediction is referred to as "non-regressive" and reflects an over-confidence in the fallible observed score as a direct reflection of the true score. One commonly used baseball example is to consider an "early-in-the-season" batting average (an "observed" score) as a direct prediction of an "end-of-the-season batting average (a presumed "true" score). As given by Kelley's equation in the text, better estimates of the true scores would regress the observed scores toward the average of the observed scores.

1947), with a reliance on the algebraic equivalence that the squared correlation between observed and true score is the reliability. If we let $\hat{\rho}$ be the estimated reliability, Kelley's equation can be written as

$$\hat{T}_X = \hat{\rho}X + (1 - \hat{\rho})\bar{X} \ ,$$

where $\bar{X}$ is the mean of the group to which the individual belongs. In other words, depending on the size of $\hat{\rho}$, a person's estimate is partly due to where the person is in relation to the group—upward if below the mean, downward if above. The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive that this equation has been labeled "Kelley's Paradox" (Wainer, 2005, pp. 67–70).

In addition to obtaining a true score estimate from an obtained score, Kelly's regression model also provides a standard error of estimation (which in this case is now referred to as the standard error of measurement). An approximate 95% confidence interval on an examinee's true score is given by

$$\hat{T}_X \pm 2\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}}) \ ,$$

where $\hat{\sigma}_X$ is the (estimated) standard deviation of the observed scores. By itself, the term $\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}})$ is the standard error of measurement, and is generated from the usual regression formula for the standard error of estimation but applied to Kelly's model that predicts true scores. The standard error of measurement most commonly used in the literature is not Kelly's but rather $\hat{\sigma}_X\sqrt{1 - \hat{\rho}}$, and a 95% confidence interval taken as the observed score plus or minus twice this standard error. An argument can be made that this latter procedure leads to "reasonable limits" (after Gulliksen, 1950) whenever

$\hat{\rho}$ is reasonably high, and the obtained score is not extremely deviant from the reference group mean. Why we should assume these latter preconditions and not use the more appropriate procedure to begin with, reminds us of a Bertrand Russell quotation (1919, p. 71): "The method of postulating what we want has many advantages; they are the same as the advantages of theft over honest toil."[2]

[2]The standard error of measurement (SEM) can play a significant role in the legal system as to who is eligible for execution. The recent Supreme Court case of *Hall v. Florida* (2014) found unconstitutional a "bright-line" Florida rule about requiring an I.Q. score of 70 or below to forestall execution due to intellectual disability. We redact part of this ruling as it pertains to the SEM of an I.Q. test:

FREDDIE LEE HALL, PETITIONER v. FLORIDA

ON WRIT OF CERTIORARI TO THE SUPREME COURT OF FLORIDA

[May 27, 2014]

JUSTICE KENNEDY delivered the opinion of the Court.

This Court has held that the Eighth and Fourteenth Amendments to the Constitution forbid the execution of persons with intellectual disability (*Atkins v. Virginia*). Florida law defines intellectual disability to require an IQ test score of 70 or less. If, from test scores, a prisoner is deemed to have an IQ above 70, all further exploration of intellectual disability is foreclosed. This rigid rule, the Court now holds, creates an unacceptable risk that persons with intellectual disability will be executed, and thus is unconstitutional.

...

On its face, the Florida statute could be consistent with the views of the medical community noted and discussed in *Atkins*. Florida's statute defines intellectual disability for purposes of an *Atkins* proceeding as "significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior and manifested during the period from conception to age 18." ... The statute further defines "significantly subaverage general intellectual functioning" as "performance that is two or more standard deviations from the mean score on a standardized intelligence test." ... The mean IQ test score is 100. The concept of standard deviation describes how scores are dispersed in a population. Standard deviation is distinct from standard error of measurement, a concept which describes the reliability of a test and is discussed further below. The standard deviation on an IQ test is approximately 15 points, and so two standard deviations is approximately 30 points. Thus a test taker who performs "two or more standard deviations from the mean" will score approximately 30 points below the mean on an IQ test, i.e., a score of approximately 70 points.

On its face this statute could be interpreted consistently with *Atkins* and with the conclusions this Court reaches in the instant case. Nothing in the statute precludes Florida from

There are several remarkable connections between Kelley's work

taking into account the IQ test's standard error of measurement, and as discussed below there is evidence that Florida's Legislature intended to include the measurement error in the calculation. But the Florida Supreme Court has interpreted the provisions more narrowly. It has held that a person whose test score is above 70, including a score within the margin for measurement error, does not have an intellectual disability and is barred from presenting other evidence that would show his faculties are limited. ... That strict IQ test score cutoff of 70 is the issue in this case.

Pursuant to this mandatory cutoff, sentencing courts cannot consider even substantial and weighty evidence of intellectual disability as measured and made manifest by the defendant's failure or inability to adapt to his social and cultural environment, including medical histories, behavioral records, school tests and reports, and testimony regarding past behavior and family circumstances. This is so even though the medical community accepts that all of this evidence can be probative of intellectual disability, including for individuals who have an IQ test score above 70. ... ("[T]he relevant clinical authorities all agree that an individual with an IQ score above 70 may properly be diagnosed with intellectual disability if significant limitations in adaptive functioning also exist"); ... ("[A] person with an IQ score above 70 may have such severe adaptive behavior problems ... that the person's actual functioning is comparable to that of individuals with a lower IQ score").

Florida's rule disregards established medical practice in two interrelated ways. It takes an IQ score as final and conclusive evidence of a defendant's intellectual capacity, when experts in the field would consider other evidence. It also relies on a purportedly scientific measurement of the defendant's abilities, his IQ score, while refusing to recognize that the score is, on its own terms, imprecise.

The professionals who design, administer, and interpret IQ tests have agreed, for years now, that IQ test scores should be read not as a single fixed number but as a range. ... Each IQ test has a "standard error of measurement," ... often referred to by the abbreviation "SEM." A test's SEM is a statistical fact, a reflection of the inherent imprecision of the test itself. ... An individual's IQ test score on any given exam may fluctuate for a variety of reasons. These include the test-taker's health; practice from earlier tests; the environment or location of the test; the examiner's demeanor; the subjective judgment involved in scoring certain questions on the exam; and simple lucky guessing.

The SEM reflects the reality that an individual's intellectual functioning cannot be reduced to a single numerical score. For purposes of most IQ tests, the SEM means that an individual's score is best understood as a range of scores on either side of the recorded score. The SEM allows clinicians to calculate a range within which one may say an individual's true IQ score lies. ... In addition, because the test itself may be flawed or administered in a consistently flawed manner, multiple examinations may result in repeated similar scores, so that even a consistent score is not conclusive evidence of intellectual functioning.

Despite these professional explanations, Florida law used the test score as a fixed number, thus barring further consideration of other evidence bearing on the question of intellectual

in the first third of the twentieth century and the modern theory of statistical estimation developed in the last half of the century. In considering the model for an observed score, X, to be a sum of a true score, T, and an error score, E, plot the observed test scores on the x-axis and their true scores on the y-axis. As noted by Galton in the 1880s (Galton, 1886), any such scatterplot suggests two regression lines. One is of true score regressed on observed score (generating Kelley's true score estimation equation given in the text); the second is the regression of observed score being regressed on true score (generating the use of an observed score to directly estimate the observed score). Kelley clearly knew the importance for measurement theory of this distinction between two possible regression lines in a true-score versus observed-score scatterplot. The quotation given below is from his 1927 text, *Interpretation of Educational Measurements*. The reference to the "last section" is where the true score was estimated directly by the observed score; the "present section" refers to his true score regression estimator:

This tendency of the estimated true score to lie closer to the mean than the obtained score is the principle of regression. It was first discovered by Francis Galton and is a universal phenomenon in correlated data. We may now characterize the procedure of the last and present sections by saying that in the last section regression was not allowed for and in the present it is. If the reliability is very high, then there is little difference between [the two methods], so that this second technique, which is slightly the more laborious, is not demanded, but if the reliability is low, there is much difference in individual outcome, and the refined procedure is always to be used in making

disability. For professionals to diagnose – and for the law then to determine – whether an intellectual disability exists once the SEM applies and the individual's IQ score is 75 or below the inquiry would consider factors indicating whether the person had deficits in adaptive functioning. These include evidence of past performance, environment, and upbringing.

individual diagnoses. (p. 177)

Kelley's preference for the refined procedure when reliability is low (that is, for the regression estimate of true score) is due to the standard error of measurement being smaller (unless reliability is perfect); this is observable directly from the formulas given earlier. There is a trade-off in moving to the regression estimator of the true score in that a smaller error in estimation is paid for by using an estimator that is now biased. Such trade-offs are common in modern statistics in the use of "shrinkage" estimators (for example, ridge regression, empirical Bayes methods, James–Stein estimators). Other psychometricians, however, apparently just don't buy the trade-off; for example, see Gulliksen (*Theory of Mental Tests*; 1950); Gulliksen wrote that "no practical advantage is gained from using the regression equation to estimate true scores" (p. 45). We disagree—who really cares about bias when a generally more accurate prediction strategy can be defined?

What may be most remarkable about Kelley's regression estimate of true score is that it predates the work in the 1950s on "Stein's Paradox" that shook the foundations of mathematical statistics. A readable general introduction to this whole statistical kerfuffle is the 1977 *Scientific American* article by Bradley Efron and Carl Morris, "Stein's Paradox in Statistics" (*236*(5), 119-127). When reading this popular source, keep in mind that the class referred to as James–Stein estimators (where bias is traded off for lower estimation error) includes Kelley's regression estimate of the true score. We give an excerpt below from Stephen Stigler's 1988 Neyman Memorial Lecture, "A Galtonian Perspective on Shrinkage Estimators" (*Statisti-*

cal Science, 1990, *5*, 147-155), that makes this historical connection explicit:

> The use of least squares estimators for the adjustment of data of course goes back well into the previous century, as does Galton's more subtle idea that there are two regression lines. ... Earlier in this century, regression was employed in educational psychology in a setting quite like that considered here. Truman Kelley developed models for ability which hypothesized that individuals had true scores ... measured by fallible testing instruments to give observed scores ... ; the observed scores could be improved as estimates of the true scores by allowing for the regression effect and shrinking toward the average, by a procedure quite similar to the Efron–Morris estimator. (p. 152)

Before we leave the topic of true score estimation by regression, we might also note what it does not imply. When considering an action for an individual where the goal is to help make, for example, the right level of placement in a course or the best medical treatment and diagnosis, then using group membership information to obtain more accurate estimates is the appropriate course to follow. But if we are facing a contest, such as awarding scholarships, or offering admission or a job, then it is inappropriate (and ethically questionable) to search for identifiable subgroups that a particular person might belong to and then adjust that person's score accordingly. Shrinkage estimators are "group blind." Their use is justified for whatever population is being observed; it is generally best for accuracy of estimation to discount extremes and "pull them in" toward the (estimated) mean of the population.

In the topic of errors-in-variables regression, we try to compensate for the tacit assumption in regression that all variables are measured

without error. Measurement error in a response variable does not bias the regression coefficients per se, but it does increase standard errors and thereby reduces power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms. Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient. Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables. Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

Milton Friedman's 1992 article entitled "Do Old Fallacies Ever Die?" (*Journal of Economic Literature*, *30*, 2129-2132), gives a downbeat conclusion regarding errors-in-variables modeling:

> Similarly, in academic studies, the common practice is to regress a variable Y on a vector of variables X and then accept the regression coefficients as supposedly unbiased estimates of structural parameters, without recognizing that all variables are only proxies for the variables of real interest, if only because of measurement error, though generally also because of transitory factors that are peripheral to the subject under consideration. I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data, alleviated only occasionally by consideration of the bias introduced when "all variables are subject to error." (p. 2131)

4 Complete Enumeration versus Sampling in the Census

The basic sampling model implies that when the size of the population is effectively infinite, this does not affect the accuracy of our estimate, which is driven solely by sample size. Thus, if we want a more precise estimate, we need only draw a larger sample.[3] For some reason, this confusion resurfaces and is reiterated every ten years when the United States Census is planned, where the issue of complete enumeration, as demanded by the Constitution, and the problems of undercount are revisited. We begin with a short excerpt from a *New York Times* article by David Stout (April 2, 2009), "Obama's Census Choice Unsettles Republicans." The quotation it contains from John Boehner in relation to the 2010 census is a good instance of the "resurfacing confusion"; also, the level of Boehner's statistical reasoning skills should be fairly clear.

> Mr. Boehner, recalling that controversy [from the early 1990s when Mr. Groves pushed for statistically adjusting the 1990 census to make up for an undercount], said Thursday that "we will have to watch closely to ensure the 2010 census is conducted without attempting similar statistical sleight of hand."

There has been a continuing and decades-long debate about the efficacy of using surveys to correct the census for an undercount. The

[3]Courts have been distrustful of sampling versus complete enumeration, and have been so for a long time. A case in 1955, for example, involved Sears, Roebuck, and Company and the City of Inglewood (California). The Court ruled that a sample of receipts was inadequate to estimate the amount of taxes that Sears had overpaid. Instead, a costly complete audit or enumeration was required. For a further discussion of this case, see R. Clay Sprowls, "The Admissibility of Sample Data into a Court of Law: A Case History," *UCLA Law Review*, *4*, 222–232, 1956–1957.

arguments against surveys are based on a combination of partisan goals and ignorance. Why? First, the census is a big, costly, and complicated procedure. And like all such procedures, it will have errors. For example, there will be errors where some people are counted more than once, such as an affluent couple with two homes being visited by census workers in May in one and by different workers in July at the other, or they are missed entirely. Some people are easier to count than others. Someone who has lived at the same address with the same job for decades, and who faithfully and promptly returns census forms, is easy to count. Someone else who moves often, is a migrant laborer or homeless and unemployed, is much harder to count. There is likely to be an undercount of people in the latter category. Republicans believe those who are undercounted are more likely to vote Democratic, and so if counted, the districts they live in will get increased representation that is more likely to be Democratic. The fact of an undercount can be arrived at through just logical considerations, but its size must be estimated through surveys. Why is it we can get a better estimate from a smallish survey than from an exhaustive census? The answer is that surveys are, in fact, small. Thus, their budgets allow them to be done carefully and everyone in the sampling frame can be tracked down and included (or almost everyone).[4] A complete enumeration is a big deal, and even though census workers try hard, they have a limited (although large) budget that does not allow the same level of precision. Because of the enormous size of the census task, increasing the budget to any plausible level will still not be enough to get everyone. A number of well-designed surveys will do a better job at a fraction of the cost.

[4]A sampling frame is the list of all those in the population that can be sampled.

The Supreme Court ruling in *Department of Commerce v. United States House of Representatives* (1999) seems to have resolved the issue of sampling versus complete enumeration in a Solomon-like manner. For purposes of House of Representatives apportionment, complete enumeration is required with all its problems of "undercount." For other uses of the Census, however, "undercount" corrections that make the demographic information more accurate are permissable And these corrected estimates could be used in differential resource allocation to the states. Two items are given in an appendix below: a short excerpt from the American Statistical Association *amicus* brief for this case, and the syllabus from the Supreme Court ruling.

5 Appendix: Brief for American Statistical Association as Amicus Curiae, Department of Commerce v. United States House of Representatives

Friend of the Court brief from the American Statistical Association —

ASA takes no position on the appropriate disposition of this case or on the legality or constitutionality of any aspect of the 2000 census. ASA also takes no position in this brief on the details of any proposed use of statistical sampling in the 2000 census.

ASA is, however, concerned to defend statistically designed sampling as a valid, important, and generally accepted scientific method for gaining accurate knowledge about widely dispersed human populations. Indeed, for reasons explained in this brief, properly designed sampling is often a better and more accurate method of gaining such knowledge than an inevitably incomplete attempt to survey all members of such a population. Therefore, in principle, statistical sampling applied to the census "has the potential to increase the quality and accuracy of the count and to reduce costs." ... There are no sound scientific grounds for rejecting all use of statistical sampling in

the 2000 census.

As its argument in this brief, ASA submits the statement of its Blue Ribbon Panel that addresses the relevant statistical issues. ASA respectfully submits this brief in hopes that its explanation of these points will be helpful to the Court.

6 Appendix: Department of Commerce v. United States House of Representatives

Syllabus from the Supreme Court ruling: The Constitution's Census Clause authorizes Congress to direct an "actual Enumeration" of the American public every 10 years to provide a basis for apportioning congressional representation among the States. Pursuant to this authority, Congress has enacted the Census Act (Act), ... delegating the authority to conduct the decennial census to the Secretary of Commerce (Secretary). The Census Bureau (Bureau), which is part of the Department of Commerce, announced a plan to use two forms of statistical sampling in the 2000 Decennial Census to address a chronic and apparently growing problem of "undercounting" of some identifiable groups, including certain minorities, children, and renters. In early 1998, two sets of plaintiffs filed separate suits challenging the legality and constitutionality of the plan. The suit in No. 98-564 was filed in the District Court for the Eastern District of Virginia by four counties and residents of 13 States. The suit in No. 98-404 was filed by the United States House of Representatives in the District Court for the District of Columbia. Each of the courts held that the plaintiffs satisfied the requirements for Article III standing, ruled that the Bureau's plan for the 2000 census violated the Census Act, granted the plaintiffs' motion for summary judgment, and permanently enjoined the planned use of statistical sampling to determine the population for congressional apportionment purposes. On direct appeal, this Court consolidated the cases for oral argument.

Held:

1. Appellees in No. 98-564 satisfy the requirements of Article III standing. In order to establish such standing, a plaintiff must allege personal injury fairly traceable to the defendant's allegedly unlawful conduct and likely to

be redressed by the requested relief. . . . A plaintiff must establish that there exists no genuine issue of material fact as to justiciability or the merits in order to prevail on a summary judgment motion. . . . The present controversy is justiciable because several of the appellees have met their burden of proof regarding their standing to bring this suit. In support of their summary judgment motion, appellees submitted an affidavit that demonstrates that it is a virtual certainty that Indiana, where appellee Hofmeister resides, will lose a House seat under the proposed census 2000 plan. That loss undoubtedly satisfies the injury-in-fact requirement for standing, since Indiana residents' votes will be diluted by the loss of a Representative. . . . Hofmeister also meets the second and third standing requirements: There is undoubtedly a "traceable" connection between the use of sampling in the decennial census and Indiana's expected loss of a Representative, and there is a substantial likelihood that the requested relief—a permanent injunction against the proposed uses of sampling in the census—will redress the alleged injury. Appellees have also established standing on the basis of the expected effects of the use of sampling in the 2000 census on intrastate redistricting. Appellees have demonstrated that voters in nine counties, including several of the appellees, are substantially likely to suffer intrastate vote dilution as a result of the Bureau's plan. Several of the States in which the counties are located require use of federal decennial census population numbers for their state legislative redistricting, and States use the population numbers generated by the federal decennial census for federal congressional redistricting. Appellees living in the nine counties therefore have a strong claim that they will be injured because their votes will be diluted vis-à-vis residents of counties with larger undercount rates. The expected intrastate vote dilution satisfies the injury-in-fact, causation, and redressibility requirements.

2. The Census Act prohibits the proposed uses of statistical sampling to determine the population for congressional apportionment purposes. In 1976, the provisions here at issue took their present form. Congress revised 13 U. S. C. §141(a), which authorizes the Secretary to "take a decennial census . . . in such form and content as he may determine, including the use of sampling procedures." This broad grant of authority is informed, however, by the narrower and more specific §195. As amended in 1976, §195 provides: "Except

for the determination of population for purposes of [congressional] apportionment ... the Secretary shall, if he considers it feasible, authorize the use of ... statistical ... 'sampling' in carrying out the provisions of this title." Section 195 requires the Secretary to use sampling in assembling the myriad demographic data that are collected in connection with the decennial census, but it maintains the longstanding prohibition on the use of such sampling in calculating the population for congressional apportionment. Absent any historical context, the "except/shall" sentence structure in the amended §195 might reasonably be read as either permissive or prohibitive. However, the section's interpretation depends primarily on the broader context in which that structure appears. Here, that context is provided by over 200 years during which federal census statutes have uniformly prohibited using statistical sampling for congressional apportionment. The Executive Branch accepted, and even advocated, this interpretation of the Act until 1994.

3. Because the Court concludes that the Census Act prohibits the proposed uses of statistical sampling in calculating the population for purposes of apportionment, the Court need not reach the constitutional question presented.

References

[1] Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

[2] Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

[3] Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

[4] Levelt Committee. (2011, October 31). *Interim report regarding the breach of scientific integrity committed by Prof. D. A. Stapel*. Tilburg, The Netherlands: Tilburg University.

[5] Russell, B. (1919). *Introduction to mathematical philosophy.* New York: Macmillan.

[6] Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures.* Princeton, NJ: Princeton University Press.

[7] Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate, and control uncertainty through graphical display.* Princeton, NJ: Princeton University Press.