Individual Risk Prediction: Or, Confidence Intervals for Dummies

Joseph Lockhart, Department of State Hospitals, California
Direct correspondence to: Joseph J. Lockhart, PhD, ABPP, 1305 North "H" Street, #117, Lompoc, CA, 93436
Email: josephjlockhart@gmail.com

Experts in many fields, including forensic psychology and psychiatry, are asked to provide risk assessments regarding future events. These risk estimates are used in individual decision-making, often with life-changing consequences. How should the expert conceptualize individual risk? And, how is such risk best communicated to the trier-of-fact? Understanding and conveying the level of uncertainty involved in the use of actuarial risk assessment instruments (ARAIs), and their application in "individual" risk has generated substantial recent controversy. Significant disagreement centers on a rather humble statistic: the confidence interval, and how to best calculate and use group confidence intervals for proportions when making decisions about individual cases. This article will discuss some of the foundational assumptions of statistics related to confidence intervals, to allow the forensic expert to understand the basic controversy surrounding the issue, and provide suggestions in how to convey risk estimates while reducing potential confusion.

Keywords: confidence intervals, risk assessment

"Life is made up of a series of judgments on insufficient data, and if we waited to run down all our doubts, it would flow past us."

-Judge Learned Hand (1939), On Receiving an Honorary Degree.

What does it mean when a physician tells their patient they have a 30% chance of having a heart attack in the next 10 years based on their Framingham risk score (http://framinghamheartstudy.org)? Or, suppose a cancer patient asks the doctor what her chances are of surviving for the next five years based on the PREDICT tool (www.predict.nhs.uk)? More to the point, what does the forensic psychologist mean when they say an offender has, for example, a 30% chance of reoffending sexually in the next five years, based on their Static-99R score?

In all these examples, the parties involved are attempting to communicate information which was developed by studying larger groups (i.e., patients, offenders) and apply it to the individual situation. As forensic experts, we recognize that the information we are asked to provide to the trier-of-fact will be used in making decisions about individuals, rather than groups. The serious consequences of these decisions make it imperative to fully understand the nature (and limitations) of the tools we use in this endeavor. Clear answers to these questions have significant ramifications, particularly in the clinical prediction of violent outcomes.

Lockhart, J., (2017). Individual risk prediction: Or, confidence intervals for dummies. *Open Access Journal of Forensic Psychology*, *9*, 19-34.

Typically, in violence risk assessments, the expert will provide a judgment of risk, either in the form of a risk category or proportional estimate (i.e., a percentage). It is generally understood there is some level of uncertainty around this prediction, and that the point estimate (e.g., “30%”) does not precisely correspond to the individual's (or group’s) exact risk. Many of the issues discussed in this article focus on the level of this uncertainty when moving from group to individual prediction, and whether such uncertainty is so great (as some have suggested) to preclude the entire enterprise of individual risk prediction.

There are three broad categories or distinct perspectives regarding group to individual inference: First, some experts believe the prediction estimate refers to an intrinsic probability within the individual (the “propensity” perspective; e.g., Gillies, 2000; Stern, 2012): “In spite of this, it is assumed that a unique probability, a true risk, of an undesirable outcome for an individual exists” (Stern, 2012, p.262). A second, opposing viewpoint proposes discarding individual risk as a concept, stating, "Contrary to what is thought, this [group] risk level is not that person’s personal risk, because probability is not meaningful in an individual context” (Sniderman, D’Agostino Sr, & Pencina, 2015, p.25). In a third approach, a group of psychologists has suggested that uncertainty about predictions can be reliably measured, but this uncertainty increases as the group size decreases (Cooke & Michie, 2010; Hart & Cooke, 2013; Hart, Michie, & Cooke, 2007). At the level of the individual, they suggest the uncertainty is so great as to make individual prediction meaningless. Somewhat pessimistically, they state, “Clinicians should consider whether it is best to give up altogether on the idea of calculating probability estimates of risk for future violence” (Hart & Cooke, 2013).

Recently, the level of uncertainty involved in the use of actuarial risk assessment instruments (ARAIs) has generated significant controversy. Many researchers have suggested that ARAIs are the most empirically supported approach to risk assessment (e.g., Hanson & Howard, 2010; Harris, Rice, Quinsey, & Cormier, 2015). Others, however (e.g., Cooke & Michie, 2010; Hart & Cooke, 2013; Hart et al., 2007) suggest that ARAIs produce unacceptably wide intervals of risk at the group and “individual” levels, provide a false sense of certainty, and do not meaningfully assist the trier of fact. They question their basic usefulness within the entire risk assessment enterprise: “At the individual level, the margins of error [of ARAIs] were so high as to render the test results virtually meaningless.”(Hart et al., 2007. p.63).

Recently, differences between these approaches led to a flurry of articles (with an entire issue of *Behavioral Sciences and the Law* devoted to the topic). These papers often contain complex-looking statistics and formulas, with exotic-sounding names such as "Wilson's formula" (for calculating confidence intervals for proportions) or "Jeffreys prior" (from Bayesian analysis; Mossman, 2015). Surprisingly, much disagreement centers on a rather humble statistic: the confidence interval. How to best calculate confidence intervals, and apply them in assessing risk, is a seemingly simple procedure. Yet, it has turned out to be fraught with mathematical, epistemological, and real-life consequences.

The average forensic psychologist or psychiatrist (and I count myself as definitely "average" when it comes to statistics) may understandably find themselves somewhat confused by this debate. It threatens their previously settled assumptions, and leads to even more uncertainty when making predictions in their daily professional life.

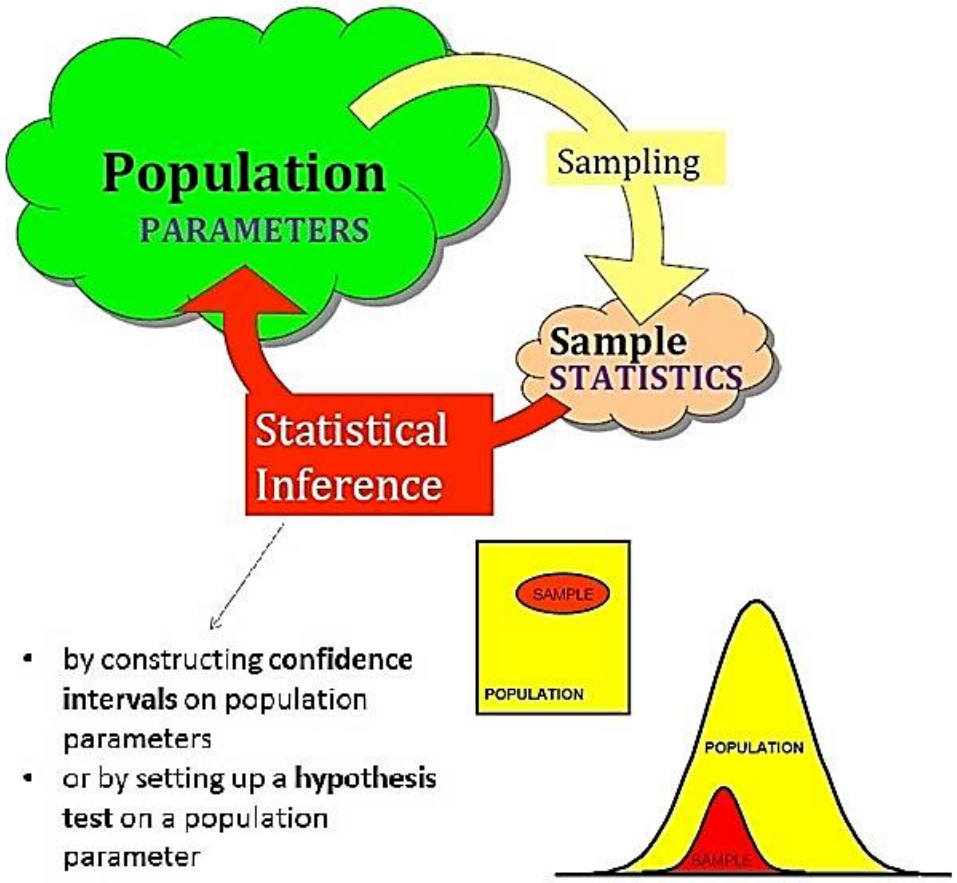
Forensic experts are not alone in this struggle, however. Other professions (particularly medicine) also use models to predict important events (such as heart disease or cancer survival), generating their own controversies around the nature of individual risk (e.g., McEvoy et al., 2014; Sniderman et al., 2015; Stern, 2012). A recent paper by Imrey and Dawid (Imrey & Dawid, 2015) has attempted to address this issue from the statistician's point of view. As the "dummy" mentioned in the title, it took me some time to review the basic principles and statistics that Imrey and Dawid present. By sharing foundational information about this important area, it is my hope that brighter forensic practitioners can access the arguments presented in Imrey and Dawid's article without the remedial work I had to do. In addition, I hope to provide some general direction and ideas for how experts may wish to conceptualize and apply these concepts in practice, including recent court rulings that have begun to take this approach into account in their judicial reasoning.

To that end, this article reviews some of the basic axioms and foundations of statistics - those that we as pragmatic forensic experts often forget, or perhaps missed in our introductory statistics class. This will hopefully provide an outline of the way in which statisticians apply the concept of prediction, especially to new individuals who may come to the evaluator's attention (the typical situation in risk assessment). This article does not discuss other aspects of test validity or accuracy, including indices of accuracy such as the AUC (Cook, 2007) or balancing "true positives" versus "false positives." (Neller & Frederick, 2013). Of course, the interested reader is encouraged to read the original sources for themselves for a more complete understanding of these issues (e.g., Gillies, 2000; Hanson & Howard, 2010; Hart et al., 2007; Imrey & Dawid, 2015).

Populations and parameters, samples and statistics

If Statistics is a way of learning about the world, what does the statistician say are its primary assumptions? Although these terms are basic (remember, I'm the dummy here), it is helpful to review first principles.

Figure 1. Populations and parameters, Samples and Statistics. Samples are drawn from populations, and their statistics are calculated. From these, the parameters of the population are estimated. (Reprinted with permission from Mahrita Harahap)



Inferential statistics is concerned with describing aspects of **populations** (see Figure 1). Populations (such as the heights of everyone in the US, the number of recidivists with a score of 6 on the STATIC-99) are typically not directly observable; they must be inferred or estimated by examining **samples**.

According to *The Philosophy of Statistics:*

We know by now that in statistical inference, we make inferences about a population based on information in a sample which is a subset of the population. Mostly we do not know the details of the population and we will use the information in the sample to estimate the unknown population quantities of interest...We call such characteristics parameters and these are constants. (Gabbay, Thagard, & Woods, 2011, p. 80)

Here is our first foundational assumption: Populations are not directly measurable; only samples drawn from them are. **Parameters** are fixed characteristics of populations that describe essential aspects of the population distribution, and we use sample **statistics** to estimate them. Typical

population parameters include the **average** and **standard deviation**. Another parameter of great interest to forensics experts, is the **proportion** or **percentage** of individuals within a population with a particular characteristic.

Statisticians introduced the following familiar notation (see Table 1) to distinguish between the parameters of a population and their corresponding sample statistics:

Table 1. *Population parameters and their sample counterparts.*

Population Parameter	**Sample Statistic**
μ ($pronounced$ "mu," for the average)	$\bar{x}$
σ ("sigma," the standard deviation)	s
P ($for\ proportion$)	p

We draw our sample from the population, and calculate sample "statistics," such as the mean, standard deviation, and proportion estimates obtained using just the data in the sample. From these sample characteristics, we "infer" the corresponding population parameters.

Orthodox statistical inference (Gabbay et al., 2011) attempts to move **from** the sample to the larger population; the only exception would be when truly **individual** data is available (e.g., batting averages for an individual baseball player). This approach seems somewhat contrary to the particular concerns of forensic psychology and psychiatry, which are tasked with providing useful information regarding the individual case.

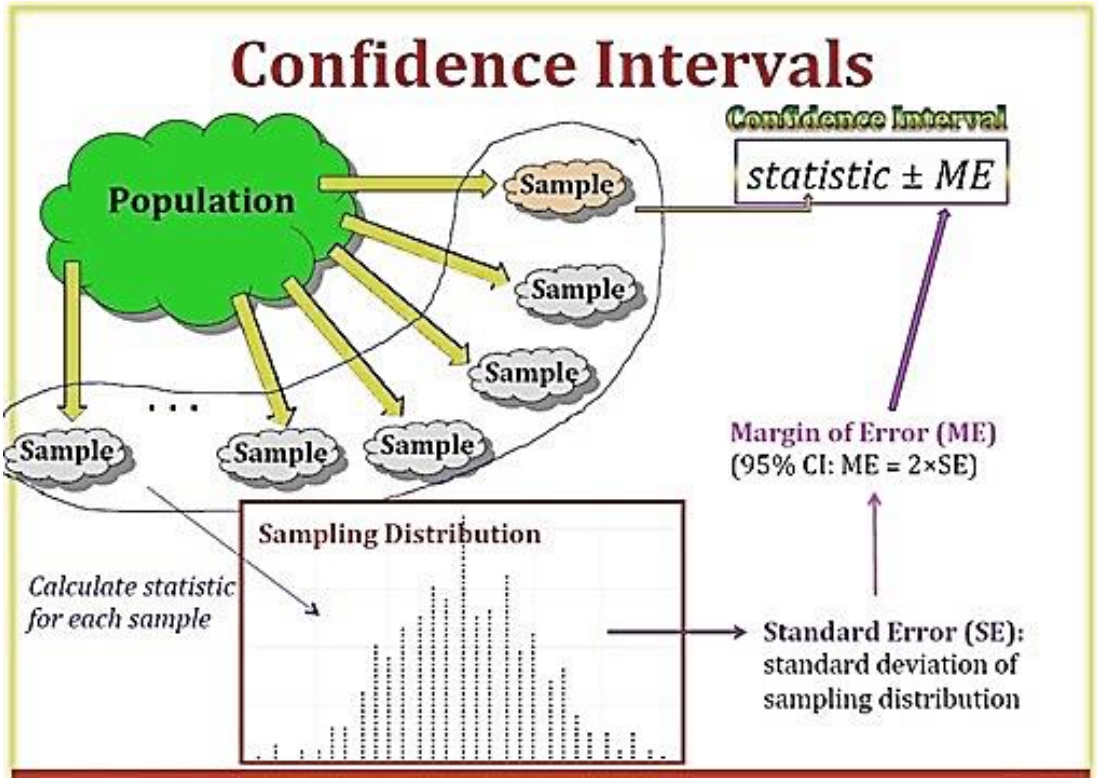
Indeed, not only are our typical statistical formulas apparently inapplicable to individuals, the task of risk assessment statistics is nearly always geared toward describing the population, rather than the single case. In order to develop a clearer understanding of how group predictions can be useful for individual cases, it is helpful to better understand the basic formulas that describe our uncertainty.

The sampling distribution: A first step in the direction of understanding confidence intervals

As noted above, samples are drawn from populations, and the sample statistics are calculated from those samples. The sample mean $\bar{x}$, for example, is obviously not exactly the same number as the parameter in the corresponding population μ. How then, do we begin to describe our uncertainty when we infer the population mean? This brings us to the topic of **sampling error**. The terms "sampling distribution," "standard error of the mean," and "confidence interval of the statistic" are all related. They attempt to describe how likely it is that our estimate of the parameter will actually

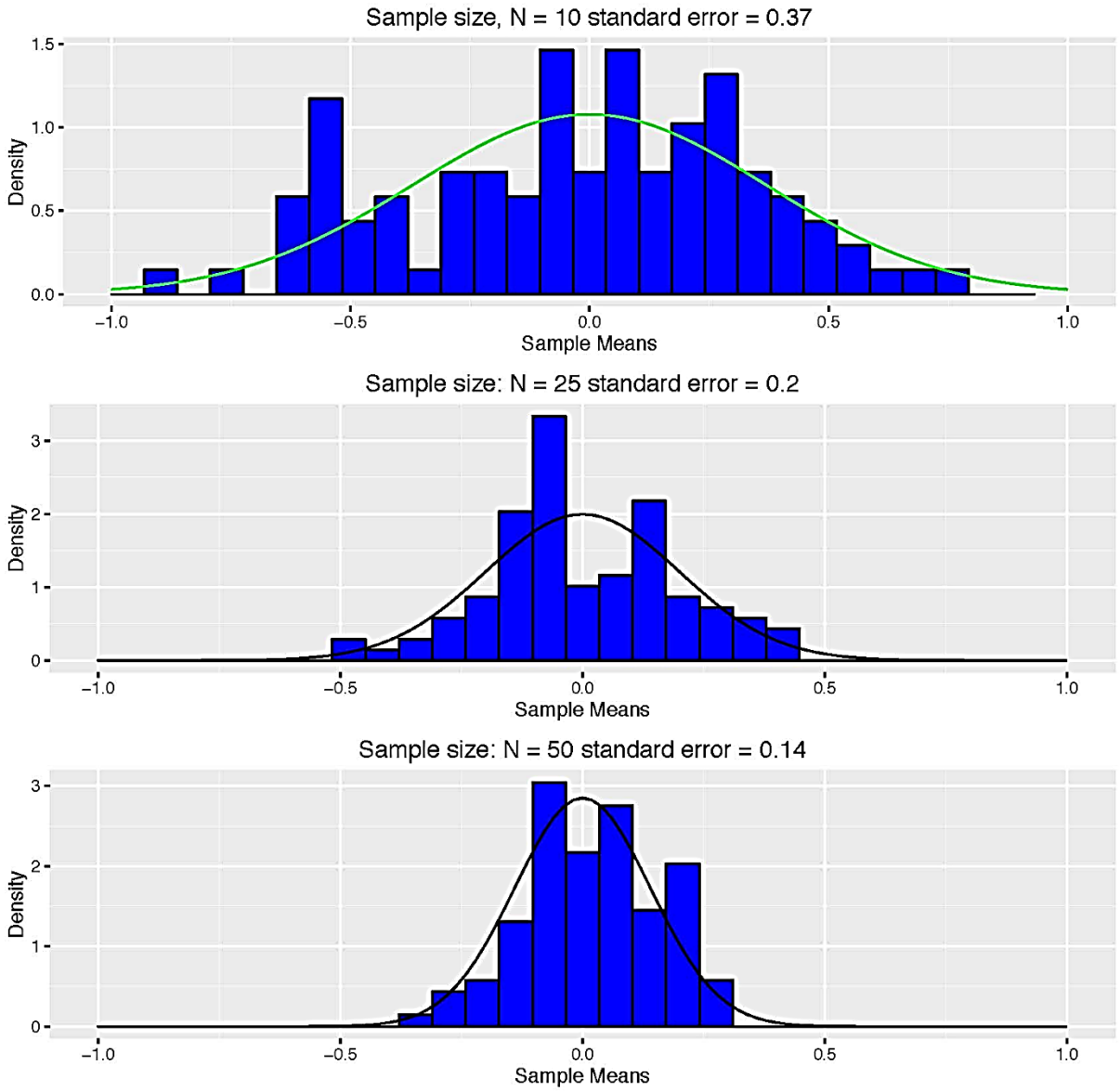
contain the "true" number. We can imagine this from Figure 2, in which the first sample is drawn from the population and its mean is calculated. A second sample is drawn, with a slightly different mean. You repeat this process a number of times (as many as you like), and generate a frequency table (i.e., a distribution) made up of the means of each of the samples.

Figure 2. The Sampling Distribution. Random samples of size *n* are drawn from the population, generating a sampling distribution that clusters around the population mean. (Reprinted with permission from Mahrita Harahap)



These sample means will cluster around the mean (μ) of the population, but will have its ow, much smaller standard deviation. This standard deviation of the mean is a measure of the **sampling error**. Typically, we don't have the luxury of multiple samples-usually there is just one. Let us begin to introduce some of the formulas that at first appear so daunting and complex in the articles debating this topic. In reality, all of these formulas have one primary goal: to estimate the sampling error based on one sample.

Figure 3. The effect of sample size on the sampling distribution. As the size of the sample drawn from the population increases, the sampling error gets narrower.



As can be seen from Figure 3, the larger the sample is, the closer it will be in estimating the parameter of interest (the population mean, in this case). All of the sampling error formulas for our statistics of interest (mean, standard deviation, or proportion) include the sample size as an essential part of their formula. As the sample size increases, the sampling error (also called the "standard error") gets smaller. This is a consequence of the Central Limit Theorem ("Central limit theorem," 2016).

The formula for the sampling error/standard error of the mean, which is perhaps most familiar to psychologists is:

$$SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

Equation 1. Standard error of the mean

We obtain an interval estimate the standard error of the mean, known as a "confidence interval" from two variables: the standard deviation (s_x) of the sample and the sample size (n). s_x is the standard deviation of our variable of interest, x. Crucially, n is the size of the sample we have drawn from the population. By increasing the sample size, the overall standard error becomes correspondingly smaller. In contrast, several researchers (Cooke & Michie, 2010; Hart & Cooke, 2013; Hart et al., 2007) attempted to apply this formula to the individual observation (i.e., using n=1) rather than the sample to generate confidence intervals regarding their risk; it was not surprising that the "confidence intervals" they obtained were extremely wide. Imrey and Dawid (2015) are clear this cannot be done within the basic confines of statistical inference, stating: "the confidence interval […] is valid only when this sample arises from the group whose true proportion, here group risk, is the inferential target"(Imrey & Dawid, 2015).

Next, we construct a 95% confidence interval for the parameter, using the standard error plus our knowledge of the area under the normal curve (i.e., roughly two standard deviations on either side of the mean of the "bell-shaped" curve, found in the back of any stats book):

$$95\%\ Confidence\ Interval\ of\ the\ Mean = \bar{X} \pm 1.96\frac{s_x}{\sqrt{n}}$$

Equation 2. Confidence interval of the mean.

This formula tells us that an interval constructed in this manner from a sample will contain the population parameter, μ (mu), for 95% of the potential random samples that might be drawn from the population. As discussed further below, it is not a prediction of individual characteristics within the population or sample.

The standard error applied to percentages

The discussion above reviewed the use of the standard error with a continuous variable, which is how many psychologists learned the concept in their statistics courses. However, most predictive risk assessments (and most ARAIs) are usually interested in the **proportion** of individuals from the larger group who either recidivate or reoffend. Variables such as recidivism are called **dichotomous** because they can only take on two values: rearrested or not, "0" or "1." When we measure the proportion of recidivists in a population of, say, sex offenders, we are in effect drawing a sample from the population of all sex offenders that we might be interested in (e.g., all sex offenders in North America, etc.).

One of the simpler formulas for the standard error of proportion is:

$$Standard\ error\ of\ proportion, \qquad SE_p = \sqrt{\frac{p(1-p)}{n}}$$

Equation 3. Standard error of the proportion (Wald's formula)

and, by extension:

$$Confidence\ Interval\ of\ the\ proportion = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

Equation 4. Confidence Interval of the proportion

This formula is quite similar to our formula for the standard error of the mean. The p, in this case, refers to the percentage (proportion) of interest (i.e., the percentage of recidivists say, among those scoring a "5" on the static 99). Thus, you have the percentage of recidivists, multiplied by the percentage of nonrecidivists, all divided by n, the sample size. Then, you take the square root of their product. Hart, Cooke, and Michie utilize a rather more complex formula (called Wilson's formula) for estimating the standard error of a proportion. In fact, there are many different ways to calculate the confidence interval of a proportion, which you can check out at this cool website: EpiTools epidemiological calculators (Sargeant, 2015).

The simpler formula works perfectly well in many situations (Sauro & Lewis, 2005), and is more than adequate to illustrate the fact that increasing sample size "narrows" the size of the standard error of the proportion. This is a key aspect of Imrey and Dawid's paper: whether the statistic is the mean, proportion, or score interval, the size of the confidence interval is largely a function of the sample size; increasing the sample will most always **narrow** the confidence interval.

Now, let's apply the formula to some real data: (Hanson, Thornton, Helmus, & Babchishin, 2016) provide data for their 5-year recidivism rates for the "routine" sample in their new norms (see Table 2). Among the 365 offenders scoring a "5" on the STATIC-99-R, 52 recidivated over 5 years, for a recidivism rate of 14.2%. We can use this information to estimate the underlying population parameter - the percentage of recidivism in a population from which the sample of 365 was drawn.

Table 2. *Sample using STATIC-99-R Data (adapted from Hanson et al., 2015)*

Recidivism data for Score=5	Ns	probability
recidivists	52	p=.14
Non-recidivists	313	(1-p).86
Total	N=365	P*(1-p)=.12

Using our formula, we have: $SE_p = \sqrt{\frac{.12}{365}} = .018$,

And, our $Confidence\ Interval\ of\ the\ proportion = .14\ \pm 1.96 * .018$,

Or $.14 \pm .036$, for a range of .10-.18.

Using our simple formula, we see that the confidence interval of the parameter is (.10, .18). I checked my calculations on the EpiTools website: All methods of calculating the confidence interval (CI) essentially agree, including Wilson's formula. Hanson et al. (2015) use a different approach, logistic regression, to obtain their estimates, which are therefore a bit narrower: .14-.17. Still, not bad for such a simple calculation!

What do population parameters and sample statistics tell us about individuals?

What the foregoing emphasizes is that inferential statistics attempt to describe uncertainty about groups and populations, not about individuals (with the exception of truly individual measures, such as the batting average example, discussed above). In risk assessment using ARAIs, what we know about the individual is not their "individual risk," but the individual's score on a relevant measure (e.g., their Static-99R score). Given that score, they can be compared to a reference group whose risk is known.

Imrey and Dawid suggest that to the extent the individual shares relevant characteristics with the pertinent group (e.g., the same Static-99R score), their predicted risk can be rationally defended to be similar to the group risk:

The individualized guess, that is, prediction, of violence recidivism risk is based upon specified characteristics of the new individual and the relation of these characteristics to recidivism in the larger, more heterogeneous, class of persons used to construct the ARAI. (p. 6)

This approach, of showing there is a good "fit" between the relevant characteristics of the individual with the previously-studied group, is consistent with other discussions of "group to individual" prediction in the literature (e.g., Faigman, Monahan, & Slobogin, 2014). It is important to underline that this prediction is not "unique" to the individual, but instead an attempt to appropriately place them within a population that we know something about. In the same vein, a prediction about the group's proportion of recidivism is not the same as predicting an individual has a specific chance of recidivating (e.g., 33%); individuals either recidivate or they do not (i.e., a dichotomous outcome). Instead, such predictions, at their most basic level, are essentially **group** predictions.

Outside of (or external to) the formal risk assessment, other potential risk factors can exist which may or may not be relevant. The most famous "external" risk factor has to be Paul Meehl's (Meehl, 1957) "broken leg." This refers to a factor which is outside of the empirically derived risk assessment, but which is so highly relevant that it should impact our decision. Most of the time,

however, we do not know the relevance of potential risk factors, nor the extent to which they may incrementally add to, or interact with factors that are already part of the model (Walker, 1995). In fact, there is good reason to believe that most current risk assessment tools (at least in sexual recidivism) are likely close to their theoretical maximum in terms of accuracy (Cook, 2007). Although it is tempting to believe we have greater certainty than suggested by an actuarial or other risk tool when we utilize unstructured "external" factors, this area appears to be one that is highly vulnerable to different types of biases (e.g. the "allegiance bias;" Chevalier, Boccaccini, Murrie, & Varela, 2015), and that such bias is particularly resistant to change (Weinstein & Klein, 1995).

As mentioned above, some complementary perspectives to deal with this problem exist: Hanson et al. (2016, p.22) suggest that actuarial results can be interpreted in light of individual **propensity**: "Within the propensity interpretation, in contrast, the likelihood of actualizing certain outcomes is a property of the individual,[…] not an abstract limit or a subjective belief." Another approach might be to use alternative statistical methods (such as Bayesian analysis). Although Bayesian statistical approaches (e.g., (Mossman, 2015; Scurich & John, 2012) would indeed provide a different conceptualization of individual risk, they would require a major shift in our risk assessment paradigm (i.e., using subjective prior beliefs). It appears unlikely (to me, at least) that such a change will find its way into U.S. courtrooms in the immediate future.

Communicating risk

"Fear of harm ought to be proportional not merely to the gravity of the harm, but also to the probability of the event."

-The monastery at Port-Royal (1662), associated with Blaise Pascal.

It is important for the expert to keep in mind that most risks (especially in high-stakes outcomes) are judged not only by their numeric probability, but by their perceived harm (Bernstein, 1996). Indeed, research suggests (e.g., Krauss & Scurich, 2014) that fairly low rates of potential recidivism are adequate for mock jurors to civilly commit an offender. As Walker (1995) suggests, it is particularly important for the expert not to confuse these two concepts, and to make clear when they are discussing a particular statistic (i.e., proportion of recidivists), and when they are discussing the potential harm of the outcome.

Explaining uncertainty is an essential part of good forensic practice, especially in high-stakes assessments. As experts, we often try to convey uncertainty in our professional role by the use of terms such as "low, moderate, or high risk"(Hilton, Scurich, & Helmus, 2015; Slovic, Monahan, & MacGregor, 2000). Several authors (e.g., Gigerenzer & Edwards, 2003; Walker, 1995) provide helpful discussions about bridging the "gulf" separating group and individual risk prediction and communication; while Walker (1995) agrees that the group to individual prediction depends on how well the individual's characteristics match those of the group, this is rarely enough to make the inferential "leap" with unqualified certainty. Walker suggests that: "The task is learning how

to communicate relative frequency estimates, while at the same time communicating the intended degree of confidence in those estimates, without getting these two tasks hopelessly confused."(p.609).

The communication of risk has been studied extensively (e.g., Gigerenzer & Edwards, 2003; Monahan & Steadman, 1996), and a full discussion of it is beyond the scope of this article. As noted above, the risks that we use in violence prediction are essentially group risks, but that does not imply they are less important, or do not play a central role in the consideration of individual risk. Just as clinicians in the medical field use predictive models to gauge potential outcomes in life or death events including heart attack or cancer survival, the forensic expert is typically asked to gauge the risk of violent outcomes, whether they be of physical violence or sexual violence. Gigerenzer (e.g., Gigerenzer & Edwards, 2003) provides strategies and concrete examples to convey risk in an understandable way, especially to a non-technical audience.

Rather than using probability statements, Gigerenzer suggests employing natural frequency language instead to communicate risk. The PREDICT model ("predict.nhs.uk site," n.d.) uses the following wording for to communicate cancer survival with and without therapy: "41 out of 100 women are alive at 5 years with no adjuvant therapy after surgery. An extra 13 out of 100 women treated are alive because of hormone therapy."

Since this article is about confidence intervals, I will attempt a tentative example here: "Among 100 individuals with scores similar to Mr. X, somewhere between 14 and 17 would be expected to recidivate sexually within the next five years. The best single estimate is that 15 out of 100 would recidivate. That means 85 out of 100 would not." (from Hanson et al., 2016, supplementary recidivism tables, corresponding to five-year risk associated with a Static-99R score of five).

In the arena of judicial reasoning about risk, most cases appear to treat risk in a statistically "casual" manner (that is, suggesting risk is intrinsic to the individual). However, at least one recent court case (*US v. Hall*, 2011) attempted to apply language consistent with group to individual inference:

Dr. X. used actuarial tools [including the Static-99R] and found a group rate of recidivism as high as 36% over 10 years. Such actuarial tests, however, only gauge a risk of recidivism based upon the statistics of the particular groups of sex offenders selected for comparison. According to the evidence presented, knowing the recidivism rate of a particular group does not mean that the individual under consideration poses the same chance of recidivism in the same timeframe; his risk could be higher or lower than that of the group based upon the unique circumstances of his case. (p.464)

This language has been quoted again with approval in a later case (*US v. Wooden*, 2012). If the courts are able to modify their understanding of risk, it is clearly time for forensic experts to follow suit and do the same.

Summary

The goal of this article is a modest one: to provide a "bridge" between the average forensic expert's statistical knowledge about group to individual inference, and the extensive material in the literature surrounding this debate. Readers should seek out this information to fully address the issue for themselves.

The conceptual "gulf" that exists between group and individual risk is often frustrating, particularly in forensic psychology and psychiatry, where individual decisions are essential. The goal of inferential statistics is to estimate attributes of a larger population from a smaller sample drawn from it. Therefore, the majority of the time, our available data and research will never directly answer the question regarding an individual's future risk, because they do not contain the right kind of data about that specific individual's recidivism behavior in highly comparable situations from that individual's past. Sampling distributions, standard error, and the standard error of the proportion are ways of conveying the extent of uncertainty in our process of estimating the population parameter from a sample; decreasing the sample size (down to an n=1) will not tell us about the risk of individuals, but only increase the error around the population estimate.

The individualized prediction of recidivism risk is based upon examining relevant characteristics of the individual, and comparing the similarity of these characteristics to recidivism in the larger, previously-studied group. The prediction is not "unique" to the individual, but instead an attempt to place the individual within a population about which we know something. It should go without saying, that having a knowledge of the group's proportion of recidivism is <u>not</u> the same as an individual's likelihood of recidivism; these are in actuality group predictions. According to Imrey (2016, personal communication), "ARAI 'individual risk' estimates should be represented in court as descriptions of the recidivism experience of groups similar in some characteristics to the offender, not as a latent characteristic of the offender himself."

As forensic experts, we must address the communication difficulties caused by the inference gap between group risk and individualized risk, especially when conveying risk is so often fraught with misunderstanding. Experts interpret risk very differently from judges and attorneys, but since our job is to assist the court, our duty is to be as clear in these concepts as possible.

Acknowledgements

The author would like to thank David Imrey for his helpful comments and suggestions on an earlier version of this manuscript.

References

:: Welcome to predict.nhs.uk site :: (n.d.). Retrieved September 22, 2016, from http://www.predict.nhs.uk/technical.html

Bernstein, P. L. (1996). *Against the gods: The remarkable story of risk*. New York: Wiley. Retrieved from http://www.ams.org/notices/199901/rev-zabell.pdf

Central limit theorem. (2016, August 24). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Central_limit_theorem&oldid=735991814

Chevalier, C. S., Boccaccini, M. T., Murrie, D. C., & Varela, J. G. (2015). Static-99R reporting practices in sexually violent predator cases: Does norm selection reflect adversarial allegiance? *Law and Human Behavior*, *39*(3), 209–218. http://doi.org/10.1037/lhb0000114

Cohen, L. J. (1989). *The philosophy of induction and probability*. Clarendon Press.

Cook, N. R. (2007). Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation*, *115*(7), 928–935. http://doi.org/10.1161/CIRCULATIONAHA.106.672402

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*(4), 259–274. http://doi.org/10.1007/s10979-009-9176-x

Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review*, 417–480.

Gabbay, D. M., Thagard, P., & Woods, J. (2011). *Philosophy of Statistics, Volume 7*. (P. S. Bandyopadhyay & M. R. Forster, Eds.) (1 edition). Amsterdam: North Holland.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ : British Medical Journal*, *327*(7417), 741–744.

Gillies, D. (2000). Varieties of propensity. *The British Journal for the Philosophy of Science*, *51*(4), 807–835.

Hanson, R. K., & Howard, P. D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior*, *34*(4), 275.

Hanson, R. K., Thornton, D., Helmus, L.-M., & Babchishin, K. M. (2016). What Sexual Recidivism Rates Are Associated With Static-99R and Static-2002R Scores? *Sexual*

Abuse: A Journal of Research and Treatment, *28*(3), 218–252. http://doi.org/10.1177/1079063215574710

Hard Coronary Heart Disease | Risk | Framingham Heart Study. (n.d.). Retrieved September 22, 2016, from https://www.framinghamheartstudy.org/risk-functions/coronary-heart-disease/hard-10-year-risk.php

Harris, G. T., Rice, M. E., Quinsey, V. L., & Cormier, C. A. (2015). *Violent offenders: Appraising and managing risk (3rd ed.)* (Vol. xvi). Washington, DC: American Psychological Association.

Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences and the Law*, *31*(1), 81–102. http://doi.org/10.1002/bsl.2049

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments Evaluating themargins of error'of group v. individual predictions of violence. *The British Journal of Psychiatry*, *190*(49), s60–s65.

Hilton, N. Z., Scurich, N., & Helmus, L.-M. (2015). Communicating the Risk of Violent and Offending Behavior: Review and Introduction to this Special Issue. *Behavioral Sciences & the Law*, *33*(1), 1–18. http://doi.org/10.1002/bsl.2160

Imrey, P. B., & Dawid, A. P. (2015). A Commentary on Statistical Assessment of Violence Recidivism Risk. *Statistics and Public Policy*, *2*(1), 1–18. http://doi.org/10.1080/2330443X.2015.1029338

Krauss, D., & Scurich, N. (2014). The impact of case factors on jurors' decisions in a sexual violent predator hearing. *Psychology, Public Policy, and Law*, *20*(2), 135–145. http://doi.org/10.1037/law0000007

McEvoy, J. W., Diamond, G. A., Detrano, R. C., Kaul, S., Blaha, M. J., Blumenthal, R. S., & Jones, S. R. (2014). Risk and the Physics of Clinical Prediction. *American Journal of Cardiology*, *113*(8), 1429–1435. http://doi.org/10.1016/j.amjcard.2014.01.418

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*(4), 268.

Monahan, J., & Steadman, H. J. (1996). Violent storms and violent people: How meteorology can inform risk communication in mental health law. *American Psychologist*, *51*(9), 931–938. http://doi.org/10.1037/0003-066X.51.9.931

Mossman, D. (2015). From group data to useful probabilities: The relevance of actuarial risk assessment in individual instances. *Journal of the American Academy of Psychiatry and the Law Online*, *43*(1), 93–102.

Neller, D. J., & Frederick, R. I. (2013). Classification Accuracy of Actuarial Risk Assessment Instruments. *Behavioral Sciences & the Law*, *31*(1), 141–153. http://doi.org/10.1002/bsl.2047

Sargeant, E. (2015). Epi Tools - EpiTools epidemiological calculators. Retrieved August 15, 2015, from http://epitools.ausvet.com.au.

Sauro, J., & Lewis, J. R. (2005). Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 49, pp. 2100–2103). SAGE Publications. Retrieved from http://pro.sagepub.com/content/49/24/2100.short

Scurich, N., & John, R. S. (2012). A Bayesian approach to the group versus individual prediction controversy in actuarial risk assessment. *Law and Human Behavior*, *36*(3), 237.

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, *24*(3), 271–296. http://doi.org/10.1023/A:1005595519944

Sniderman, A. D., D'Agostino Sr, R. B., & Pencina, M. J. (2015). The role of physicians in the era of predictive analytics. *JAMA*, *314*(1), 25–26.

Stern, R. H. (2012). Individual Risk. *The Journal of Clinical Hypertension*, *14*(4), 261–264. http://doi.org/10.1111/j.1751-7176.2012.00592.x

US v. Hall, 664 F. 3d 456 (Court of Appeals, 4th Circuit October 28, 2011).

US v. Wooden, 693 F. 3d 440 (Court of Appeals, 4th Circuit May 17, 2012).

Walker, V. R. (1995). Direct inference, probability, and a conceptual gulf in risk communication. *Risk Analysis*, *15*(5), 603–607.

Weinstein, N. D., & Klein, W. M. (1995). Resistance of personal risk perceptions to debiasing interventions. *Health Psychology*, *14*(2), 132.