REPRESENTATIVENESS AND STATISTICS IN FIELD PERFORMANCE ASSESSMENT

Gregory B. Baecher[1]

ABSTRACT: Measurement of engineering performance is fundamental to empirical understanding, model development, and to the observational method. It is also expensive. Yet, how representative are field observations of the geotechnical conditions at work, and how informative are they of critical design assumptions? Interesting lessons for geotechnical practice are suggested from considering such issues, and from qualitative statistical reasoning about observing geotechnical performance on limited budgets. The discussion considers three interrelated topics: intuitive misperceptions about samples and sampling variations, the nature of uncertainty and randomness in modeling soil deposits, and simple lessons derived from statistical sampling theory.

INTRODUCTION

Monitoring field performance involves a great deal of sampling, and builds upon lessons learned in a number of other scientific and technical fields. In this regard, field monitoring differs little from other sampling activities, whether in testing pharmaceuticals, inspecting airplane engines, or conducting political polls. We make a limited number of observations and try to draw scientifically defensible conclusions from them. Clearly, the interpretation of field data requires a good deal of insight into geology and engineering mechanics, and also experience with construction practices. Still, intuition fails even the most sophisticated engineer or scientist faced with the vagaries of scattered experimental data, and too little time or too few resources with which to make more observations.

For simplicity, the present discussion limits consideration about the purpose of field monitoring to two objectives: (i) assessing site conditions, and (ii) testing the validity of an analytical model. Clearly, field monitoring has other objectives, for example, in quality control, and in observational methods, but these are deferred to another occasion. The discussion also ignores legal issues, despite their importance to practice.

---

[1] Professor and Chairman, Department of Civil and Environmental Engineering, University of Maryland, College Park, MD 20742. gbaecher@eng.umd.edu

This is a preview. Click here to purchase the full publication.

INTUITION AND THE INTERPRETATION OF DATA

Surprisingly, even trained statisticians are easily led astray when using intuition to interpret sampling observations. Mere scientists and engineers, as a result, have little hope of accurately interpreting sample variations, errors, and biases simply based on inspection. Yet, that is usually the approach taken in practice. Interestingly—but maybe not surprisingly—the errors people make in intuitively interpreting data are remarkably similar from one person to another.

Sample variation

When measurements are made in the laboratory or field they exhibit scatter. These measurements might more generically be called, *observations*, to include things other than instrument readings. A set of observations is typically called a *sample*. If the sample comprises a large number of observations, the data scatter among the latter tends to exhibit regularity. That is, the scatter within a sample and from one sample to another tends to display regular patterns, and over the years statisticians have learned to categorize those patterns, and to use them to draw inferences about the population from which the sample comes, the *parent population*. Patterns of scatter within an individual sample are interpreted against what is known of the probable scatter among samples to make estimates about the parent population.

For convenience, the scatter within a sample or from one sample to another is described by a frequency distribution or histogram, and this in turn can be summarized by its low order statistical moments. The most useful of these are the first two moments, the *mean* or arithmetic average of the observations, m, and the *variance* or mean-square variation about the mean, s^2. The *standard deviation*, s, is the root-mean-square variation (square-root of the variance), and the ratio of standard deviation to mean is the *coefficient of variation*, Ω=s/m. Such mathematical functions of the sample observations are said to be *statistics* of the data, or alternately, *sample statistics*, and form the basis for making inferences about the parent population.

The *law of large numbers*, a fundamental principle of statistical theory, implies that, as the sample becomes larger, the statistical properties of the sample become ever more to resemble the population from which the sample is taken.[2] The operative phrase here is, "as the sample becomes larger." For example, one is often interested in the average value of some parameter or performance measure in the field, across all the elements of the population that may not have been observed in the sample. If the number of observations in a sample is large, it seems reasonable, and the law of large numbers confirms, that one might use the sample average of the set of observed values as indicative of the population average in the field. But, what

[2] The Law of Large Numbers is more specific and limited than this colloquial interpretation (Feller, 1967), but the practical implication is quite broad. See, also, Maistrov (1974) for an historical sketch.

This is a preview. Click here to purchase the full publication.

about the case where the sample is not large? This is almost always true in geotechnical practice.

The law of large numbers says that variations of the moments of a sample about their counterparts in the parent population become ever smaller as sample size increases, but for small samples these variations can be large. Presume we take many samples of size n from the same population, and for each sample we calculate the sample mean, m. The values of m across the many samples themselves exhibit scatter, and could be plotted as a histogram. This distribution of the sample mean, or of any other sample parameter, is called the *sampling distribution*. The sampling distribution is the frequency distribution of some sample statistic over repeated sampling. The theoretical variance of the scatter among the means of each of many samples of size n is $Var(m)=s^2/n$. Var(m) is the second moment of the sampling distribution of m. Correspondingly, the standard deviation of the sample means is, $s_m=s/(n)^{1/2}$.

The coefficient of variation of soil properties measured in the field can be as large as 100%, although values of 30-50% are more common (Kulhawy and Trautmann, 1996; Phoon and Kulhawy, 1996). Thus, if ten (10) tests are made, the variation of their sample average about the (unknown) population (soil deposit) average would have a standard deviation of 10-16%. Since, under very general assumptions, the sampling distribution of the mean is well approximated by a Normal distribution, the range for which one would be comfortable bracketing the population mean is, say, 2 to 2.5 standard deviations, or in this case between ±20-40% of the best estimate.[3] In other words, there is considerable uncertainty in the inference of even average soil properties, when reasonable sample sizes are taken into account. There is, of course, even more uncertainty about inferences of soil properties at specific locations within a soil deposit.

Representativeness

Despite the fact that sampling variations can be large—and in geotechnical practice, they are large—there is an intuitive tendency to treat sample results as representative of—or similar to—the population from which they are taken. Most people believe intuitively that samples should reflect the essential characteristics of the population out of which they arise, and thus the converse, that essential characteristics of the population should mimic those of the sample. People's intuition tells them that a sample should be similar to the population from which it comes, but that is only true in the limit, as sample sizes become large. This leads to errors. Speaking strictly, *representativeness* is a property of sampling plans, not of samples. A sampling plan is representative of the population being sampled if every element of the population has an equal chance of affecting the (weighted) properties of the sample (Cochran,

[3] The Normal limit to the sampling distribution follows from the Central Limit Theorem, which is closely related to the Law of Large Numbers, see, e.g., Feller (1967)

This is a preview. Click here to purchase the full publication.

1977), and from this one speaks of "representative sampling."[4] A sample, in contrast can never be representative: it is a unique collection of particular elements within the population, and each such collection has different properties.

Take, for example, the string of sample outcomes deriving from six tosses of a fair coin, {H,H,H,H,H,H}. Most people intuitively think of this string as less likely to occur than the string, {H,T,T,H,T,H}, even though each has the same probably, $(½)^6$. This is akin to the Gambler's Fallacy that if *heads* has not appeared in some time, it is overdue and should occur with increased probability. Intuition tells us that the sample should represent the population, that is, be similar to the population in salient aspects, and in short runs as well as long. In this case, the sample should have about the same number of heads as tails, and the sequence of heads and tails should be "random," that is, erratic. That this is a misperception is obvious to anyone who thinks about it; yet, our intuition tells us otherwise.

The same thing is true of samples of geotechnical observations. We presume them to be representative of the geotechnical population out of which they arise. The averages within the sample ought to be about the same as the averages in situ. The variability of observations ought to be about the same as the variability in situ. Spatial patterns of variation among the observations ought to mirror spatial patterns in situ. All of these things are true in the limit, but for small samples they are compromised by sampling variability, and may be profoundly untrue. Small samples of the size typical in geotechnical practice seldom display the salient properties of the population; the variability among sample outcomes is simply too great.

Overconfidence

This intuitive belief in representativeness leads people to believe that important characteristics of a population should manifest in every sample, no matter the sample size. Yet, we know from statistical theory that this is not true: small samples exhibit large variation from one to another. This leads people to put too much faith in the results of small numbers of observations and to overestimate the replicability of such results. If tests are repeated, people have unreasonably high expectations that significant results will be replicated. Thus, the ten (10) observations of field performance above are made, and one is surprised that the next set of ten yields a 30% difference in average results. A person's typically response is not to ascribe this difference to expectable statistical variation, but to seek a cause. The engineering literature is filled with well-intentioned attempts to explain away differing sample results, when in fact, such explanations would be more in order had such differences not been observed.

---

[4] In some places the term, representative sampling, is used to mean that the probability of sampling sub-populations is set equal the relative frequency of those sub-populations within the overall population. This meaning is subsumed within the definition here.

This is a preview. Click here to purchase the full publication.

A corollary to this belief in the representativeness of small samples is the over-confidence even scientifically trained people place in their inferences or estimates of unknown quantities. In a famous early study, Alpert and Raiffa (1982) demonstrated that when asked to place 25%:75% or 5%-95% confidence bounds on estimates of unknown quantities, the true values of the quantities being estimated fall outside the assessed bounds considerably more often that the nominal 50% or 10%, respectively. Often more than half the real values fall outside 5%-95% confidence bounds people estimate. This result has been replicated in another early study by Folayan et al (1970) involving engineers' estimates of the properties of San Francisco Bay Mud, and by Hynes and Vanmarcke (1976) involving predictions of embankment height at failure for the MIT I-95 Test Embankment. Data from Folayan et al. are recalculated in Table 1 to show 95% confidence intervals on the subjective assessments of the mean compression ratio, and to show 95% confidence intervals derived from 42 tests at the site. The lack of overlap between the subjects' intervals and that calculated from sample observations suggests strong overconfidence on the part of the subjects.

Subject	2.5% limit	97.5% limit
1	0.29	0.31
2	0.27	0.28
3	0.26	0.29
4	0.26	0.34
5	0.20	0.43
Sample	0.32	0.36

Table 1. 95% confidence intervals on average compression ratio for San Francisco Bay Mud at a particular construction site, subjectively estimated by five engineers. Confidence interval also shown based on n=42 observations at the site (after, Folayan, et al., 1970).

As reliability analysis becomes increasingly important to geotechnical practice, it is sometimes suggested that a field expedient way of assessing the standard deviation of an uncertain quantity is by eliciting the maximum and minimum bounds one could conceive the quantity having, and then assuming that this range spans a certain number of standard deviations of variation, typically, ±3s. The reasoning is that for a Normal variate, ±3 standard deviations spans 99.75% of the variation. But, if people are over confident of their estimates of uncertain quantities—which we know them to be—then people will frequently be surprised in practice to find their maximum and minimum bounds exceeded. Thus, the "six-sigma" rule is unconservative, and possibly quite significantly. This can also be seen in Figure 1, in which the expected range of sample values, $r_n=|x_{max}-x_{min}|$, for a Normal variate is plotted as a function of sample size. Even for samples as large as n=20, the range expected in a sample is less than 4 standard deviations. The reciprocal of this expected range, in fact, makes a useful estimator of standard deviation, and one with known sampling properties (Snedecor and Cochran, 1980).

This is a preview. Click here to purchase the full publication.

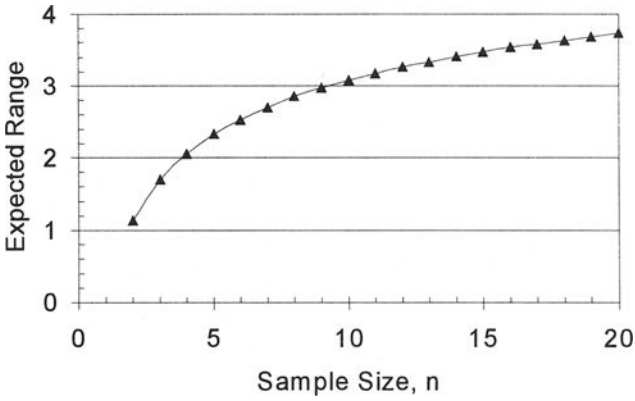

Figure 1. Expected range of Normal sample in standard deviation units

"Law of small numbers"

In a series of celebrated papers in the, 1970's, the late Amos Tversky and Daniel Kahneman, now of Princeton University, introduced the scientific world to the systematic differences between the way people perceive probability and the way statistical theory operates, and to the term *representativeness* as used above (1971, 1974, 1979). That body of work, and the explosion of studies that followed, are sometimes referred to as the "heuristics and biases" school of thought on subjective probability (see, e.g., Morgan and Henrion, 1990).

This body of work emphasizes that the use of representativeness (similarity) to judge probabilities is fraught with difficulty, because it is not affected by factors that should influence judgments of probability. Important among these are the overconfidence described above, a disregard for base rates (*a priori* probabilities), and ignorance of common regression effects. This concept that observers presume samples to be representative of the population seems benign, but leads to serious errors of judgment in practice. Tversky and Kahneman (1971), dubbed this, "The Law of Small Numbers," which states simply, that the Law of Large Numbers applies to small numbers as well.

This overlooking of sample size manifests even when a problem is stated so as to emphasize sample size, and in many different contexts. Consider, for example, a question that arose in a flood hazard damage reduction study. A river basin was analyzed in two different ways to assess levee safety. In the first case, the river was divided into 10miles (6 km) long reaches; in the second, the river was divided into 1 mile (0.6 km) long reaches. Would the average settlements within the levee reaches have greater variability in the first case, the second case, or about the same in each? Of an admittedly unscientific sample of 25 graduate students and engineers, 7 said

This is a preview. Click here to purchase the full publication.

the first (more variation among long reaches), 6 said the second (more variation among short reaches), and 12 said the last (about equal). But clearly, the long reaches have the least variation among their average settlements, because they are larger samples. Smaller samples are more erratic.

Prior probabilities

A second manifestation of representativeness is that people tend to overlook background rates and focus instead on the likelihood of the observed data when drawing conclusions. To review for a moment, Bayes' Theorem says that the probability one ascribes to an event or parameter estimate should be the product of two probabilities: the probability *a priori* to observing new data, and the likelihood (conditional probability) of the new data given the event or parameter value. This is summarized by the familiar expression,

$$\Pr\{\Theta \mid data\} = \Pr\{\Theta\} L\{data \mid \Theta\} \tag{1}$$

in which Θ is an event or parameter (the state of nature), $\Pr\{\Theta\}$ is the probability of Θ prior to observing the data, $\Pr\{\Theta \mid data\}$ is the probability after observing the data, and $L\{data \mid \Theta\}$ is the conditional probability of the data given Θ (i.e., the Likelihood). This relationship led DeFinetti (1937) to say, "data never speak for themselves," they tell us only how to modify what we thought before we saw them to what we should logically think afterward. What the data tell us is summarized in the likelihood function. What we thought before is summarized in the prior probabilities.

Sometimes representativeness leads people to place undue importance on sample data (because they "should be similar to the population"), and in so doing ignore, or at least downplay, prior probabilities (the latter sometimes referred to as *base-rates* in the heuristics and biases literature). As a simple example, in risk analyses for dam safety a geologist might be asked to assess the probability that faults exist undetected in the bottom of a valley. Noting different rock formations on the adjoining valley walls, he or she might assign a high probability to faulting, because of the association of this condition with faulting, in spite of the fact, say, that the base-rate of faulting in the region is low. The two sources of evidence, prior probability and likelihood, should each influence the *a posteriori* probability (Eqn. 1), but intuition leads us to focus on the sample likelihood and, to some extent, ignore the prior probability.

Regression to the mean

Today, we think of regression analysis as fitting lines to data, but when Francis Galton did his pioneering work in the 1870's, and coined the term, his interest was not in best-fit lines but in reversion to the mean (Stigler, 1999). Galton experimented with the sizes of peas, and noted that, on average, size is inherited. Large peas tend to have larger than average offspring, and small peas the reverse. He noted also that, while on average the offspring of large peas are larger than their counter-

This is a preview. Click here to purchase the full publication.

parts, they are also on average smaller than their parents. The offspring revert part of the way back to the population average. The fitting of lines came into the picture because the average distance between the size of the offspring and the population average was a linear function of the distance between the size of the parent and the population average (Figure 2).

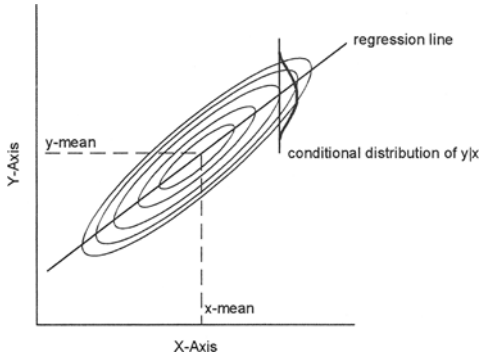

Figure 2. Regression line representing the expected values of y for given value of x. Note, because the regression line is less steep than the axis of the data ellipse, the conditional average of y for a given x is proportionately closer to the y-mean than the value of x is to the x-mean.

This regression effect occurs all the time in everyday life, and is related to the error people make in presuming representativeness. We look at the present or most recent sample or observation, and presume it is representative of the next; but, even eliminating sample size effects for the moment, this may not be the case. Consider that a numerical model with sophisticated constitutive equations is used to predict the performance of some earth structure. This model is applied to a randomly selected test section, and performs well. The predictions it makes of, say, deformations are closely matched by field measurements. Now, the model is applied to another test section. Will it perform as well? No, on average it will not, and one should not be surprised: it's basic regression.

Model predictions are based on theory, simplifications, and statistical parameter estimates. There is necessarily variation in how well a model predicts from one test section to another. Yet, if the model has predictive validity, it will on average be correlated to actual performance, and the accuracies of prediction from one test section to another should be correlated as well. If two predictions are correlated, there exists a regression relationship between them. Invoking Galton's observation, one should expect the second prediction to be less good than the first more than half the time. Of course, the converse is also true. If the first model prediction was not so good, the second will on average be better.

This is a preview. Click here to purchase the full publication.

THE "RANDOM SOIL PROFILE"

In order to circumvent intuitive errors it has become more common to use formal statistical methods in analyzing field monitoring data, and indeed soil testing generally. This is part of a larger trend toward the use of risk and reliability methods in geotechnical engineering, a trend heralded by the emergence of load-resistance factor design (LRFD) in geotechnical codes (Kulhawy and Phoon, 1996), the increasing use of risk analysis in dam safety (Von Tunn, 1996) and flood hazard damage reduction studies (USACE, 1996a, 1996b), and the appearance of prominent lectures on practical applications of reliability.

These new approaches have introduced concepts into geotechnical engineering that are relatively new to practice, and perhaps not fully appreciated by those trying to use them. First, what does it mean for soil properties at a particular site and within a particular soil profile to be "random?" Clearly, unlike the weather, soil properties do not fluctuate erratically with time. In principle, the properties of the soil ought to be knowable everywhere. The only reason they are not known everywhere, and with precision up to our ability to measure, is that limited resources or limited testing technology has precluded them being observed.

Second, what does it mean for predictions of engineering performance to be "uncertain?" Uncertainty comes in many forms. Field measurements are scattered, so the ability to calibrate models to engineering performance is imprecise. Soil test data are biased, so estimates of soil engineering parameters that go into the models are inaccurate. The models used to predict performance are simplifications of reality, so forecasts are only approximations. Do all these different types of uncertainty affect predictions of engineering performance in the same way?

Third, what does it mean for uncertainties to be related to one another, that is, correlated? Some parameters, c and ϕ for example, are not actually separate physical properties but rather curve fitting numbers, and thus dependent on one another. Along a long reach of levee or long excavation, the variation of performance in space may have a systematic although uncertain pattern. Errors in estimating commonly shared parameters may mean that uncertainties in different types of engineering performance are implicitly related, even if mechanistically independent. Do these interdependencies significantly affect predictive uncertainty?

The nature of randomness

> Random *(adjective). Date: 1565. 1. a: lacking a definite plan, purpose, or pattern b: made, done, or chosen at random; 2. a: relating to, having, or being elements or events with definite probability of occurrence. b: being or relating to a set or to an element of a set each of whose elements has equal probability of occurrence. [Merriam-Webster, 1999].*

This is a preview. Click here to purchase the full publication.

We use terms like *randomness, uncertainty, probability,* and *chance* all the time in the course of professional practice, yet without devoting much thought to what they mean, or to the larger philosophical questions their use implies. Most engineers, at least those who deal with the macroscopic world, think of nature as deterministic. For any effect there is a cause, and a cause and its effect are mechanistically linked. What then does it mean for something to be random? If the world is deterministic rather than random—at least at the scale of earth structures—what does it mean to speak of probabilities in relation to the world? Do probabilities describe some fundamental physical process, or do they have to do with limited information?

When we describe something as random, we normally mean that it is inherently unpredictable except probabilistically. Flood frequencies, for example, have been treated as an inherently random aspect of nature for many decades. In flood hazard studies we describe flood discharges only in exceedance probabilities (return periods). Thus, we treat flood discharges as if their magnitudes were generated by a celestial pair of dice. The peak discharge in a specific period of time, such as this year, cannot be predicted. All that can be said is that in a long series of years like this one, some fraction of the years will experience peak discharges larger than some fixed value.

Does this mean that rainfall and runoff are unpredictable processes of nature? No, not necessarily. Given advances in atmospheric science and hydrology, it is becoming ever more common for weather models to be used in predicting rainfall, and thus runoff and flood heights. Such models have also been used to predict probable maximum floods for dam safety studies (Salmon, 1999). When flood discharges are predicted by mechanistic modeling, they cease to be treated as random processes. The uncertainties surrounding predictions of flood flows change from those associated with random events to those associated with model and parameter errors. So, the assumption of randomness is only a convenience of modeling.

Randomness at the macro scale is an assumption, not an inherent part of the physical world. In principle, one ought to be able to predict whether a tossed coin lands heads-up or heads-down, but in practice, it is more convenient to assume that coin tossing is a random process resulting in a consistent frequency of each possible outcome as the experiment is repeated a large number of times. Randomness is not a property of the world; it is an artifact of modeling.

The nature of uncertainty

> Uncertain *(adjective). Date: 14th century. 1: Indefinite, indeterminate 2 : not certain to occur : Problematical 3: not reliable: Untrustworthy 4 a : not known beyond doubt : Dubious b: not having certain knowledge: Doubtful c: not clearly identified or defined 5: not constant: Variable, Fitful [Merriam-Webster, 1999].*

This is a preview. Click here to purchase the full publication.