Measurement theory: Frequently asked questions

Version 3, Sep 14, 1997

Warren S. Sarle
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513, USA
saswss@unx.sas.com

URL: `ftp://ftp.sas.com/pub/neural/measurement.html`

Version 1 originally published in the Disseminations of the International Statistical Applications Institute, volume 1, edition 4, 1995, Wichita: ACG Press, pp. 61-66.

Contents

What is measurement theory?

Measurement theory is a branch of applied mathematics that is useful in measurement and data analysis. The fundamental idea of measurement theory is that measurements are not the same as the attribute being measured. Hence, if you want to draw conclusions about the attribute, you must take into account the nature of the correspondence between the attribute and the measurements.

The mathematical theory of measurement is elaborated in:

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971), *Foundations of measurement, Vol. I: Additive and polynomial representations,* New York: Academic Press.

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1989), *Foundations of measurement, Vol. II: Geometrical, threshold, and probabilistic respresentations,* New York: Academic Press.

Luce, R. D., Krantz, D. H., Suppes, P., and Tversky, A. (1990), *Foundations of measurement, Vol. III: Representation, axiomatization, and invariance,* New York: Academic Press.

Measurement theory was popularized in psychology by S. S. Stevens, who originated the idea of levels of measurement. His relevant articles include Stevens (1946, 1951, 1959, 1968).

For a recent discussion of measurement theory and statistics, see Hand (1996).

What is measurement?

Measurement of some attribute of a set of things is the process of assigning numbers or other symbols to the things in such a way that relationships of the numbers or symbols reflect relationships of the attributes of the things being measured. A particular way of assigning numbers or symbols to measure something is called a *scale* of measurement.

Suppose we have a collection of straight sticks of various sizes and we assign a number to each stick by measuring its length using a ruler. If the number assigned to one stick is greater than the number assigned to another stick, we can conclude that the first stick is longer than the second. Thus a relationship among the numbers (greater than) corresponds to a relationship among the sticks (longer than). If we lay two sticks end-to-end in a straight line and measure their combined length, then the number we assign to the concatenated sticks will equal the sum of the numbers assigned to the individual sticks (within measurement error). Thus another relationship among the numbers (addition) corresponds to a relationship among the sticks (concatenation). These relationships among the sticks must be empirically verified for the measurements to be valid.

Why should I care about measurement theory?

Measurement theory helps us to avoid making meaningless statements. A typical example of such a meaningless statement is the claim by the weatherman on the local TV station that it was twice as warm today as yesterday because it was 40 degrees Fahrenheit today but only 20 degrees yesterday. This statement is meaningless because one measurement (40) is twice the other measurement (20) only in certain arbitrary scales of measurement, such as Fahrenheit. The relationship 'twice-as' applies only to the numbers, not the attribute being measured (temperature).

When we measure something, the resulting numbers are usually, to some degree, arbitrary. We *choose* to use a 1 to 5 rating scale instead of a -2 to 2 scale. We choose to use Fahrenheit instead of Celsius. We choose to use miles per gallon instead of gallons per mile. The conclusions of a statistical analysis should not depend on these arbitrary decisions, because we could have made the decisions differently. We want the statistical analysis to say something about reality, not simply about our whims regarding meters or feet. If a given statement may be either true or false depending on arbitrary, unspecified choices, then that statement is logically meaningless.

Suppose we have a rating scale where several judges rate the goodness of flavor of several foods on a 1 to 5 scale. If we want to draw conclusions about the measurements, i.e. the 1-to-5 ratings, then we need not be concerned about measurement theory. For example, if we want to test the hypothesis that the foods have equal mean ratings, we might do a two-way ANOVA on the ratings.

But if we want to draw conclusions about flavor, then we *must* consider how flavor relates to the ratings, and that is where measurement theory comes in. Ideally, we would want the ratings to be linear functions of the flavors with the same slope for each judge; if so, the ANOVA can be used to make inferences about mean goodness-of-flavors, providing we can justify all the appropriate statistical assumptions. But if the judges have different slopes relating ratings to flavor, or if these functions are not linear, then this ANOVA will *not* allow us to make inferences about mean goodness-of-flavor. Note that this issue is not about statistical interaction; even if there is no evidence of interaction in the ratings, the judges may have different functions relating ratings to flavor.

We need to consider what information we have about the functions relating ratings to flavor for each judge. Perhaps the only thing we are sure of is that the ratings are monotone increasing functions of flavor. In this case, we would want to use a statistical analysis that is valid no matter what the particular monotone increasing functions are. One way to do this is to choose an analysis that yields invariant results no matter what monotone increasing functions the judges happen to use, such as a Friedman test. The study of such invariances is a major concern of measurement theory.

However, no measurement theorist would claim that measurement theory provides a complete solution to such problems. In particular, measurement theory generally does not take random measurement error into account, and if such errors are an important aspect of the measurement process, then additional methods, such as latent variable models, are called for. There is no clear boundary between measurement theory and statistical theory; for example, a Rasch model is both a measurement model and a statistical model.

What are permissible transformations?

Permissible transformations are transformations of a scale of measurement that preserve the relevant relationships of the measurement process. *Permissible* is a technical term; use of this term does not imply that other transformations are prohibited for data analysis any more than use of the term *normal* for probability distributions implies that other distributions are pathological. If Stevens had used the term *mandatory* rather than *permissible*, a lot of confusion might have been avoided.

In the example of measuring sticks, changing the unit of measurement (say, from centimeters to inches) multiplies the measurements by a constant factor. This multiplication does not alter the correspondence of the relationships 'greater than' and 'longer than', nor the correspondence of addition and concatenation. Hence, change of units is a permissible transformation with respect to these relationships.

What are levels of measurement?

There are different levels of measurement that involve different properties (relations and operations) of the numbers or symbols that constitute the measurements. Associated with each level of measurement is a set of permissible transformations. The most commonly discussed levels of measurement are as follows:

Nominal:

- Two things are assigned the same symbol if they have the same value of the attribute.
- Permissible transformations are any one-to-one or many-to-one transformation, although a many-to-one transformation loses information.
- Examples: numbering of football players; numbers assigned to religions in alphabetical order, e.g. atheist=1, Buddhist=2, Christian=3, etc.

Ordinal:

- Things are assigned numbers such that the order of the numbers reflects an order relation defined on the attribute. Two things `x` and `y` with attribute values `a(x)` and `a(y)` are assigned numbers `m(x)` and `m(y)` such that if `m(x) > m(y)`, then `a(x) > a(y)`.
- Permissible transformations are any monotone increasing transformation, although a transformation that is not strictly increasing loses information.
- Examples: Moh's scale for hardness of minerals; grades for academic performance (A, B, C, ...); blood sedimentation rate as a measure of intensity of pathology.

Interval:

- Things are assigned numbers such that differences between the numbers reflect differences of the attribute. If `m(x) - m(y) > m(u) - m(v)`, then `a(x) - a(y) > a(u) - a(v)`.
- Permissible transformations are any affine transformation `t(m) = c * m + d`, where `c` and `d` are constants; another way of saying this is that the origin and unit of measurement are arbitrary.
- Examples: temperature in degrees Fahrenheit or Celsius; calendar date.

Log-interval:

- Things are assigned numbers such that ratios between the numbers reflect ratios of the attribute. If `m(x) / m(y) > m(u) / m(v)`, then `a(x) / a(y) > a(u) / a(v)`.
- Permissible transformations are any power transformation `t(m) = c * m ** d`, where `c` and `d` are constants.
- Examples: density (mass/volume); fuel efficiency in mpg.

Ratio:

- Things are assigned numbers such that differences and ratios between the numbers reflect differences and ratios of the attribute.
- Permissible transformations are any linear (similarity) transformation `t(m) = c * m`, where `c` is a constant; another way of saying this is that the unit of measurement is arbitrary.
- Examples: Length in centimeters; duration in seconds; temperature in degrees Kelvin.

Absolute:

- Things are assigned numbers such that all properties of the numbers reflect analogous properties of the attribute.
- The only permissible transformation is the identity transformation.
- Examples: number of children in a family, probability.

These measurement levels form a partial order based on the sets of permissible transformations:

```
Weaker <-----------------------------------> Stronger

                     - Interval -
                    /            \
Nominal -- Ordinal <              > Ratio -- Absolute
                    \            /
                     Log-interval
```

In real life, a scale of measurement may not correspond precisely to any of these levels of measurement. For example, there can be a mixture of nominal and ordinal information in a single scale, such as in questionnaires that have several non-response categories. It is common to have scales that lie somewhere between the ordinal and interval levels in that the measurements can be assumed to be a smooth monotone function of the attribute. For many subjective rating scales (such as the 'strongly agree,' 'agree,' ... 'strongly disagree' variety) it cannot be shown that the intervals between successive ratings are exactly equal, but with reasonable care and diagnostics it may be safe to say that no interval represents a difference more than two or three times greater than another interval.

The above list of measurement levels is not exhaustive. It is not unusual to encounter other scales of measurement that do not have such widely recognized names. For example, directional or circular data may be measured on what might be called a periodic-interval scale, which has an arbitrary origin and unit as well as a period related to the unit. Time of day, for example, conventionally has an origin of midnight, a unit of hours, and a period of 24 hours.

Unfortunately, there are also many situations where the measurement process is too ill-defined for measurement theory to apply. In such cases, it may still be fruitful to consider what arbitrary choices were made in the course of measurement, what effect these choices may have had on the measurements, and whether some plausible class of permissible transformations can be determined.

What about binary (0/1) variables?

For a binary variable, the classes of one-to-one transformations, monotone increasing/decreasing transformations, and affine transformations are identical--you can't do anything with a one-to-one transformation that you can't do with an affine tranformation. Hence binary variables are at least at the interval level. If the variable connotes presence/absence or if there is some other distinguishing feature of one category, a binary variable may be at the ratio or absolute level.

Nominal variables are often analyzed in linear models by coding binary dummy variables. This procedure is justified since binary variables are at the interval level or higher.

Is measurement level a fixed, immutable property of the data?

Measurement level depends on the correspondence between the measurements and the attribute. Given a set of data, one cannot say what the measurement level is without knowing what attribute is being measured. It is possible that a certain data set might be treated as measuring different attributes at different times for different purposes.

Consider a rat in a Skinner box who pushes a lever to get food pellets. The number of pellets dispensed in the course of an experiment is obviously an absolute-level measurement of the number of pellets dispensed. If number of pellets is considered as a measure of some other attribute, the measurement level may differ. As a measure of amount of food dispensed, the number of pellets is at the ratio level under the assumption that the pellets are of equal size; if the pellets are not of equal size, a more elaborate measurement model is required, perhaps one involving random measurement error if the pellets are dispensed in random order. As a measure of duration during the experiment, the number of pellets is at an ordinal level. As a measure of response effort, the number of pellets might be approximately ratio level, but we would need to consider whether the rat's responses were executed in a consistent way, whether the rat may miss the lever, and so forth. As a measure of amount of reward, the number of pellets could only be justified by some very strong assumptions about the nature of rewards; the measurement level would depend on the precise nature of those assumptions. The main virtue of measurement theory is that it encourages people to consider such issues.

Velleman and Wilkinson (1993) have pointed out that decisions about the scale of measurement should not be made in haste. Not only may information about the measurement process be inadequate, but during the course of a statistical analysis, we may revise our theories about what attributes we are trying to measure. Especially in predictive modeling, we may discover that some of the predictor variables contain previously unsuspected information. However, measurement, like experimental design, is something that *should* be considered carefully before collecting data.

Once a set of measurements have been made on a particular scale, it may be possible to transform the measurements to yield a new set of measurements at a different level. It is always possible to transform from a stronger level to a weaker level. For example, a temperature measurement in degrees Kelvin is at the ratio level. If we convert the measurements to degrees Celsius, the level is interval. If we rank the measurements, the level becomes ordinal. In some cases it is possible to convert from a weaker scale to a stronger scale. For example, correspondence analysis can convert nominal measurements to an interval scale under appropriate assumptions, and multidimensional scaling or conjoint analysis can convert ordinal measurements to an interval scale if the model is correct.

Isn't an ordinal scale just an interval scale with error?

You can view an ordinal scale as an interval scale with error if you really want to, but the errors are not independent, additive, or identically distributed as required for many statistical methods. The errors would involve complicated dependencies to maintain monotonicity with the interval scale. In the example above with number of pellets as a measure of duration, the errors would be cumulative, not additive, and the error variance would increase over time. Hence for most statistical purposes, it useless to consider an ordinal scale as an interval scale with measurement error.

What does measurement level have to do with discrete vs. continuous?

Measurement level has nothing to do with discrete vs. continuous variables.

The distinction between discrete and continuous random variables is commonly used in statistical theory, but that distinction is rarely of importance in practice. A continuous random variable has a continuous cumulative distribution function. A discrete random variable has a stepwise-constant cumulative distribution function. A discrete random variable can take only a finite number of distinct values in any finite interval. There exist random variables that are neither continuous nor discrete; for example, if Z is a standard normal random variable and Y=max(0,Z), then Y is neither continuous nor discrete, but has characteristics of both.

While measurements are always discrete due to finite precision, attributes can be conceptually either discrete or continuous regardless of measurement level. Temperature is usually regarded as a continuous attribute, so temperature measurement to the nearest degree Kelvin is a ratio-level measurement of a continuous attribute. However, quantum mechanics holds that the universe is fundamentally discrete, so temperature may actually be a discrete attribute. In ordinal scales for continuous attributes, ties are impossible (or have probability zero). In ordinal scales for discrete attributes, ties are possible. Nominal scales usually apply to discrete attributes. Nominal scales for continuous attributes can be modeled but are rarely used.

Don't the theorems in a statistics textbook prove the validity of statistical methods without reference to measurement theory?

Mathematical statistics is concerned with the connection between inference and data. Measurement theory is concerned with the connection between data and reality. Both statistical theory and measurement theory are necessary to make inferences about reality.

Does measurement level detemine what statistics are valid?

Measurement theory cannot determine some single statistical method or model as appropriate for data at a specific level of measurement. But measurement theory does in fact show that some some statistical methods are inappropriate for certain levels of measurement if we want to make meaningful inferences about the attribute being measured.

If we want to make statistical inferences regarding an attribute based on a scale of measurement, the statistical method must yield invariant or equivariant results under the permissible transformations for that scale of measurement. If this invariance or equivariance does not hold, then the statistical inferences apply only to the measurements, not to the attribute that was measured.

If we record the temperature in degrees Fahrenheit in Cary, NC, at various times, we can compute statistics such as the mean, standard deviation, and coefficient of variation. Since Fahrenheit is an interval scale, only statistics that are invariant or equivariant under change of origin or unit of measurement are meaningful. The mean is meaningful because it is equivariant under change of origin or unit. The standard deviation is meaningful because it is invariant under change of origin and equivariant under change of unit. But the coefficient of variation is meaningless because it lacks such invariance or equivariance. The mean and standard deviation can easily be converted back and forth from Fahrenheit to Celsius, but we cannot compute the coefficient of variation in degrees Celsius if we know only the coefficient of variation in degrees Fahrenheit.

Paul Thompson provides an example where interval and ratio levels are confused:

> ... I recently published a paper in Psychiatric Research. We discuss the BPRS, a very common rating scale in psychiatry. Oddly enough, in the BPRS in the US, '1' means NO PATHOLOGY. However, a frequent statistic computed is percent improved:
>
> ```
> PI = (BPRS(Base)-BPRS(6week) ) / BPRS(Base) * 100
> ```
>
> If you use the '1 implies no pathology' model, you are not measuring according to a ratio scale, which requires a true 0. We show that this has very bad characteristics, which include a flat impossibility for a certain % improvement at certain points in the scale.
>
> This is pretty trivial, but should have an effect. As one reviewer said, 'This is so obvious that I am surprised that no one has ever thought of it before.' Nonetheless, the scale was being misused.

It is clear that if we are estimating a parameter that lacks invariance or equivariance under permissible transformations, we are estimating a chimera. The situation for hypothesis testing is more subtle. It is nonsense to test a null hypothesis the truth of which is not invariant under permissible transformations. For example, it would be meaningless to test the null hypothesis that the mean temperature in Cary in July is twice the mean temperature in December using a Fahrenheit or Celsius scale--we would need a ratio scale for that hypothesis to be meaningful.

But it is possible for the null hypothesis to be meaningful even if the error rates for a given test are not invariant. Suppose that we had an ordinal scale of temperature, and the null hypothesis was that the distribution of temperatures in July is identical to the distribution in December. The truth of this hypothesis is invariant under strictly increasing monotone transformations and is therefore meaningful under an ordinal scale. But if we do a t-test of this hypothesis, the error rates will not be invariant under monotone transformations. Hard-core measurement theorists would therefore consider a t-test inappropriate. But given a null hypothesis, there are usually many different tests that can be performed with accurate or conservative significance levels but with different levels of power against different alternatives. The fact that different tests have different error rates does not make any of them correct or incorrect. Hence a soft-core measurement theorist might argue that invariance of error rates is not a prerequisite for a meaningful hypothesis test--only invariance of the null hypothesis is required.

Nevertheless, the hard-core policy rules out certain tests that, while not incorrect in a strict sense, are indisputably poor tests in terms of having absurdly low power. Consider the null hypothesis that two random variables are independent of each other. This hypothesis is invariant under one-to-one transformations of either variable. Suppose we have two nominal variables, say, religion and preferred statistical software product, to which we assign arbitrary numbers. After verifying that at least one of the two variables is approximately normally distributed, we could test the null hypothesis using a Pearson product-moment correlation, and this would be a valid test. However, the power of this test would be so low as to be useless unless we were lucky enough to assign numbers to categories in such a way as to reveal the dependence as a linear relationship. Measurement theory would suggest using a test that is invariant under one-to-one transformations, such as a chi-squared test of independence in a contingency table. Another possibility would be to use a Pearson product-moment correlation after assigning numbers to categories in such a way as to maximize the correlation (although the usual sampling distribution of the correlation coefficient would not apply). In general, we can test for independence by maximizing some measure of dependence over all permissible transformations.

However, it must be emphasized that there is no need to *restrict* the transformations in a statistical analysis to those that are permissible. That is not what *permissible transformation* means. The point is that statistical methods should be used that give invariant results under the class of permissible transformations, because those transformations do not alter the meaning of the measurements. *Permissible* was undoubtedly a poor choice of words, but Stevens was quite clear about what he meant. For example (Stevens 1959):

> In general, the more unrestricted the permissible transformations, the more restricted the statistics. Thus, nearly all statistics are applicable to measurements made on ratio scales, but only a very limited group of statistics may be applied to measurements made on nominal scales.

The connection between measurement level and statistical analysis has been hotly disputed in the psychometric and statistical literature by people who fail to distinguish between inferences regarding the attribute and inferences regarding the measurements. If one is interested only in making inferences about the measurements without regard to their meaning, then measurement level is, of course, irrelevant to choice of statistical method. The classic example is Lord's (1953) article "On the Statistical Treatment of Football Numbers." Lord argued that statistical methods could be applied regardless of level of measurement, and concocted a silly example involving the jersey numbers assigned to football players, which Lord claimed were nominal-level measurements of the football players. Lord contrived a situation in which freshmen claimed they were getting lower numbers than the sophomores, so the purpose of the analysis was to make inferences about the numbers, not about some attribute measured by the numbers. It was therefore quite reasonable to treat the numbers as if they were on an absolute scale. However, this argument completely misses the point by eliminating the measured attribute from the scenario.

The confusion between measurements and attributes was perpetuated by Velleman and Wilkinson (1993), who set up a series of straw men and knocked some of them down, while consistently misunderstanding the meaning of *meaning* and of *permissible transformation*. For example, they claimed that the number of cylinders in an automobile engine can be treated, depending on the circumstances, as nominal, ordinal, interval, or ratio, and hence the concept of measurement level "simplifies the matter so far as to be false." In fact, the number of cylinders is at the absolute level of measurement. Thus, measurement theory would dictate that any statistical analysis of number of cylinders must be invariant under an identity transformation. Obviously, *any* analysis is invariant under an identity transformation, so all of the analyses that Velleman and Wilkinson claimed might be appropriate *are* acceptable according to measurement theory. What is false is not measurement theory but Velleman and Wilkinson's backwards interpretation of it.

It is important to understand that the level of measurement of a variable does not mandate how that variable must appear in a statistical model. However, the measurement level does suggest reasonable ways to use a variable by default. Consider the analysis of fuel efficiency in automobiles. If we are interested in the average distance that can be driven with a given amount of gas, we should analyze miles per gallon. If we are interested in the average amount of gas required to drive a given distance, we should analyze gallons per mile. Both miles per gallon and gallons per mile are measurements of fuel efficiency, but they may yield quite different results in a statistical analysis, and there may be no clear reason to use one rather than the other. So how can we make inferences regarding fuel efficiency that do not depend on the choice between these two scales of measurement? We can do that by recognizing that both miles per gallon and gallons per mile are measurements of the same attribute on a log-interval scale, and hence that the logarithm of either can be treated as a measurement on an interval scale. Thus, if we were doing a regression, it would be reasonable to begin the analysis using log(mpg). If evidence of nonlinearity were detected, then other transformations could still be considered.

Rank tests are broadly useful for ordinal data because ranking often produces the required invariance of test statistics. But ranking is not some sort of ordinal-to-interval conversion. An hypothesis that is meaningless for ordinal data does not become meaningful when the data are ranked. For example, in a two-way factorial design, the hypothesis of additivity (no interaction) of the effects requires an interval or stronger scale. Ranking does not allow tests of interaction for ordinal data because no well-defined hypothesis is being tested--the truth of the hypothesis of additivity of ranks can change depending on how many cases are in each cell of the design.

The cookbook approach to measurement theory and rank tests has yielded some peculiar ideas. Measurement theory certainly does not demand that ordinal data be ranked, since there are other ways of achieving the necessary invariance (e.g., Agresti, 1984; Gifi, 1990). Neither does measurement theory forbid the use of rank tests, as some people have argued under the misguided notion that ranks, being ordinal, cannot be summed; sums of ranks can be used meaningfully when they have the necessary invariance properties.

But measurement level has been shown empirically to be irrelevant to statistical results, hasn't it?

What has been shown is that various statistical methods are more or less robust to distortions that could arise from smooth monotone transformations; in other words, there are cases where it makes little difference whether we treat a measurement as ordinal or interval. But there can hardly be any doubt that it often makes a huge difference whether we treat a measurement as nominal or ordinal, and confusion between interval and ratio scales is a common source of nonsense.

Suppose we are doing a two-sample t-test; we are sure that the assumptions of ordinal measurement are satisfied, but we are not sure whether an equal-interval assumption is justified. A smooth monotone tranformation of the entire data set will generally have little effect on the p value of the t-test. A robust variant of a t-test will likely be affected even less (and, of course, a rank version of a t-test will be affected not at all). It should come as no surprise then that a decision between an ordinal or an interval level of measurement is of no great importance in such a situation, but anyone with lingering doubts on the matter may consult the simulations in Baker, Hardyck, and Petrinovich (1966) for a demonstration of the obvious.

On the other hand, suppose we were comparing the variability instead of the location of the two samples. The F test for equality of variances is not robust, and smooth monotone transformations of the data could have a large effect on the p value. Even a more robust test could be highly sensitive to smooth monotone transformations if the samples differed in location.

Measurement level is of greatest importance in situations where the meaning of the null hypothesis depends on measurement assumptions. Suppose the data are 1-to-5 ratings obtained from two groups of people, say males and females, regarding how often the subjects have sex: frequently, sometimes, rarely, etc. Suppose that these two groups interpret the term 'frequently' differently as applied to sex; perhaps males consider 'frequently' to mean twice a day, while females consider it to mean once a week. Females may report having sex more 'frequently' than men on the 1-to-5 scale, even if men in fact have sex more frequently as measured by sexual acts per unit of time. Hence measurement considerations are crucial to the interpretation of the results.

What are some more examples of how measurement level relates to statistical methodology?

As mentioned earlier, it is meaningless to claim that it was twice as warm today as yesterday because it was 40 degrees Fahrenheit today but only 20 degrees yesterday. Fahrenheit is not a ratio scale, and there is no meaningful sense in which 40 degrees is twice as warm as 20 degrees. It would be just as meaningless to compute the geometric mean or coefficient of variation of a set of temperatures in degrees Fahrenheit, since these statistics are not invariant or equivariant under change of origin. There are many other statistics that can be meaningfully applied only to data at a sufficiently strong level of measurement.

Consider some measures of location: the mode requires a nominal or stronger scale, the median requires an ordinal or stronger scale, the arithmetic mean requires an interval or stronger scale, and the geometric mean or harmonic mean require a ratio or stronger scale.

Consider some measures of variation: entropy requires a nominal or stronger scale, the standard deviation require an interval or stronger scale, and the coefficient of variation requires a ratio or stronger scale.

Simple linear regression with an intercept requires that both variables be on an interval or stronger scale. Regression through the origin requires that both variables be on a ratio or stronger scale.

A generalized linear model using a normal distribution requires the dependent variable to be on an interval or stronger scale. A gamma distribution requires a ratio or stronger scale. A Poisson distribution requires an absolute scale.

The general principle is that an appropriate statistical analysis must yield invariant or equivariant results for all permissible transformations. Obviously, we cannot actually conduct an infinite number of analyses of a real data set corresponding to an infinite class of transformations. However, it is often straightforward to verify or falsify the invariance mathematically. The application of this idea to summary statistics such as means and coefficients of variation is fairly widely understood.

Confusion arises when we come to linear or nonlinear models and consider transformations of variables. Recall that Stevens did *not* say that transformations that are not 'permissible' are prohibited. What Stevens said was that we should consider *all* 'permissible' transformations and verify that our conclusions are invariant.

Consider, for example, the problem of estimating the parameters of a nonlinear model by maximum likelihood (ML), and comparing various models by likelihood ratio (LR) tests . We would want the LR tests to be invariant under the permissible transformations of the variables. One way to do this is to parameterize the model so that any permissible transformation can be inverted by a corresponding change in the parameter estimates. In other words, we can make the ML and LR tests invariant by making the inverse-permissible transformations mandatory (this is the same set of transformations as the permissible transformations except for a degeneracy here and there which I won't worry about).

To illustrate, suppose we are modeling a variable Y as a function `f()` of variables `N, O, I, L, R`, and `A` at the nominal, ordinal, etc. measurement levels, respectively. Then we can ensure the desired invariance by setting up the model as:

```
   Y = f( arb(N), mon(O), a+bI, cL^d, eR, A, ...)
```

where `arb()` is any (estimated) function, `mon()` is any (estimated) monotone function, and `a, b, c, d`, and `e` are parameters. Then any permissible transformations of `N, O, I, L, R`, and `A` can be absorbed by the estimation of the `arb()` and `mon()` functions and the parameters. The function `f()` can involve *any other* transformations such as sqrts or logs or whatever. `f()` can be as complicated as you like--the presence of the permissible transformations as part of the model to be estimated guarantees the desired invariance.

If we were designing software for fitting linear or nonlinear models, we might want to provide these 'permissible' or 'mandatory' transformations in a convenient way. This, in fact, was the motivation for numerous programs developed by psychometricians that anticipated many of the features of ACE and generalized additive models.

Are there other theories of measurement?

Michell (1986, 1990; also discussed by Hand, 1996) has distinguished among "representational" measurement theory as described in this FAQ, "operational" theory, and "classical" theory. The representational theory assumes that there exists a "reality" that is being measured, and that scientific theories are about this reality. The operational theory avoids the assumption of an underlying reality, requiring only that measurement consist of precisely specified operations; scientific theories concern only relationships among measurements. The classical theory, like the representational theory, assumes an objective reality, but, unlike the representational theory, holds that only quantitative attributes are measurable, and measurement involves the discovery of the magnitudes of these attributes. In the classical theory, like the operational theory, meaningfulness comes from empirical support for scientific theories describing the interrelationships of various measurements.

It is interesting to consider the dramatic international rise in IQ scores (the "Flynn effect": Flynn, 1987; Neisser, 1997) in light of these three theories of measurement. Has intelligence increased, or just the test scores? The representational theory makes clear the importance of this distinction. The operational theory makes the question impossible to ask, since there is no such thing as "intelligence" as distinct from scores on intelligence tests. The classical theory allows the question, "What do IQ tests really measure?" (Flynn, 1987), but it is difficult to see how intelligence can be regarded as having a magnitude susceptible to classical measurement by Michell's (1986) definition.

The distinctions among Michell's theories of measurement are not always clear. Consider latent variable models, such as a Rasch model. Michell (1986, p. 404) says that "Such an approach to psychological measurement owes more to the classical theory of measurement than to either the operational or the representational theories." Hand (1996, p. 455) says, "It seems to me, however, that the so-called *latent* variables are operationally defined by their relationships to the observed variables." But Hand later (p. 457) describes a Rasch model as an example of classical measurement. My opinion is that a Rasch model is a form of representational measurement involving probabilitic relationships--an extension, but a natural extension, of Stevens's idea of measurement.

The operational theory has the virtue of discouraging sloppiness. As Hand (1996) points out, operational measurement may be good enough for predictive (as opposed to explanatory) models. But operationalism has some severe philosophical disadvantages. Robert Klee (1997) says about operationalism (pp. 53-54):

> The most influential doctrine about [correspondence rules] ever to circulate among practicing scientists themselves (and not just philosophers of science) was *operationalism.* Operationalism was first introduced to a wide scientific audience by the physicist Percy Bridgman in 1927, just as logical positivism was starting up in central Europe. Operationalism did not last long in the physical sciences; but, for reasons that continue to puzzle philosophers of science, it survives to this day with considerable influence in the social and behavioral sciences (especially psychology), where the methodological war cry to "operationalize your variables!" persists among practitioners in certain quarters despite the problems with operationalism that we are about to investigate.

Michell (1990, p. 28) says that operationalism is logically false, hence the operational theory of measurement must be rejected. He also claims (p. 49) that the representational theory, while not logically incoherent, is nevertheless empirically false. However, this conclusion seems to be based more on philosophical convictions than empirical results, primarily on the claim that numbers are empirical entities, not abstractions. Most of Michell's (1990) book is devoted to promoting the classical theory of measurement, in which numbers are empirical, but which in other respects seems very similar to the representational theory. However, in the discussion of Hand's (1996, pp. 481-482) paper, Michell emphasizes that the three theories of measurement are mutually contradictory.

While I find most of Michell's (1990) conclusions unconvincing, I think his emphasis on philosophy is enlightening with respect to the ongoing arguments about measurement. In the debate about the implications of measurement theory for statistical practice, it often seems that the two sides are arguing past each other, each side considering their own position to be self-evident. Klee (1997) points out that a similar situation exists with regard to arguments between realist and anti-realist philosophers of science. Operationalism is dead, but operational measurement could still be supported by some varieties of anti-realism, especially the pragmatic school. Perhaps the arguments about measurement theory go nowhere because of unstated philosophical assumptions.

What's the bottom line?

Measurement theory shows that strong assumptions are required for certain statistics to provide meaningful information about reality. Measurement theory encourages people to think about the meaning of their data. It encourages critical assessment of the assumptions behind the analysis. It encourages responsible real-world data analysis.

References

Agresti, A. (1984), *Analysis of Ordinal Categorical Data.* NY: Wiley.

Baker, B. O., Hardyck, C, and Petrinovich, L. F. (1966), "Weak measurement vs. strong statistics:

An empirical critique of S.S. Stevens' proscriptions on statistics," Educational and Psychological Measurement, 26, 291-309.

Bridgman, P. (1927), *The Logic of Modern Physics,* NY: Macmillan.

Flynn, J.R. (1987), "Massive IQ gains in 14 nations: What IQ tests really measure," Psychological Bulletin, 101, 171-191.

Gifi, A. (1990), *Nonlinear Multivariate Analysis,* Chichester: Wiley.

Hand, D.J. (1996), "Statistics and the theory of measurement," with discussion, J. of the Royal Statistical Society, Series A, 159, 445-492.

Klee, R. (1997), *Introduction to the Philosophy of Science: Cutting Nature at Its Seams,* NY: Oxford University Press.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971), *Foundations of measurement, Vol. I: Additive and polynomial representations,* New York: Academic Press.

Lord, F.M. (1953), "On the Statistical Treatment of Football Numbers," American Psychologist, 8, 750-751,

Luce, R. D., Krantz, D. H., Suppes, P., and Tversky, A. (1990), *Foundations of measurement, Vol. III: Representation, axiomatization, and invariance,* New York: Academic Press.

Michell, J. (1986), "Measurement scales and statistics: a clash of paradigms," Psychological Bulletin, 100, 398-407.

Michell, J. (1990), *An Introduction to the Logic of Psychological Measurement,* Hillsdale: Erlbaum.

Neisser, U. (1997), "Rising scores on intelligence tests," American Scientist, 85, 440-447.

Suppes, P., Krantz, D. H., Luce, R. D., and Tversky, A. (1989), *Foundations of measurement, Vol. II: Geometrical, threshold, and probabilistic respresentations,* New York: Academic Press.

Stevens, S. S. (1946), "On the theory of scales of measurement," Science, 103, 677-680.

Stevens, S. S. (1951), "Mathematics, measurement, and psychophysics," in S. S. Stevens (ed.), *Handbook of experimental psychology,* pp 1-49), New York: Wiley.

Stevens, S. S. (1959), "Measurement," In C. W. Churchman, ed., *Measurement: Definitions and Theories,* pp. 18-36, New York: Wiley. Reprinted in G. M. Maranell, ed., (1974) *Scaling: A Sourcebook for Behavioral Scientists,* pp. 22-41, Chicago: Aldine.

Stevens, S. S. (1968), "Measurement, statistics, and the schemapiric view," Science, 161, 849-856.

Velleman, P.F., and Wilkinson, L. (1993), "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading," The American Statistician, 47, 65-72.

LBA for WSS, 18 Mar 1996