# V

## Validity of Scales

Elisabeth D. Svensson
Professor Emerita
Örebro University, Örebro, Sweden

The concepts of quality of measurements made by rating scales and multi-scale questionnaires are *validity* and *reliability*. Corresponding concepts for quantitative data (interval and ratio data) are accuracy and precision. A rating scale is *valid* if it measures what it is intended to measure in the specific study. The *validity* of self-estimated subjective phenomena is relative and cannot be assessed absolutely. The validity of a scale is study specific, and must be considered each time the scale or the ▶questionnaire is chosen for a new study. Therefore there are various concepts of validity, each addressing a specific type of quality assessment. The main concepts are *criterion, construct*, and *content* validity, but a large number of sub concepts are used. The meaning of these concepts is not univocal and depends on applications and research paradigms. *Criterion validity* refers to the conformity of a scale to a true state or a gold standard, and depending on the purpose of the study sub concepts like *clinical, predictive* and *concurrent validity* will be used.

*Construct validity* refers to the consistency between scales having the same theoretical definition in the absence of a true state or a gold standard. Sub concepts like *convergent, descriptive, discriminant, divergent, factorial, translation validity and parallel reliability* have been used in studies. *Biologic validity* refers to the closeness of scale assessments to the hypothesized expectation when comparing with other measures in a specific population. Discriminative rating scales are used to distinguish between individuals or groups, when no external criterion is available, then *discriminant validity* is to be assessed. *Parallel reliability* refers to the interchangeability of scales.

The concept *content validity* refers to the completeness of the scale or multi-scale questionnaire in the coverage of important areas. Sub concepts like *face, ecological, decision, consensual, sampling validity, comprehensiveness* and *feasibility* have been used.

Assessments on *rating scales* generate ordinal data having rank-invariant properties only, which means that the responses indicate a rank order and not a mathematical value. The results of statistical treatments of data must not being changed when relabeling the ordered responses. Appropriate statistical methods for evaluation of criterion and construct validity often refer to the order consistency or to the relationship between the scales of comparison.

The scatter plot of 48 paired assessments of perceived back pain on a visual analogue scale (VAS) and on a verbal descriptive scale (VDS-5) having five ordered categorical responses is shown the Fig. 1.

As evident from the plot there is a large overlapping between the assessments. The probability of discordance in paired observations (X,Y),

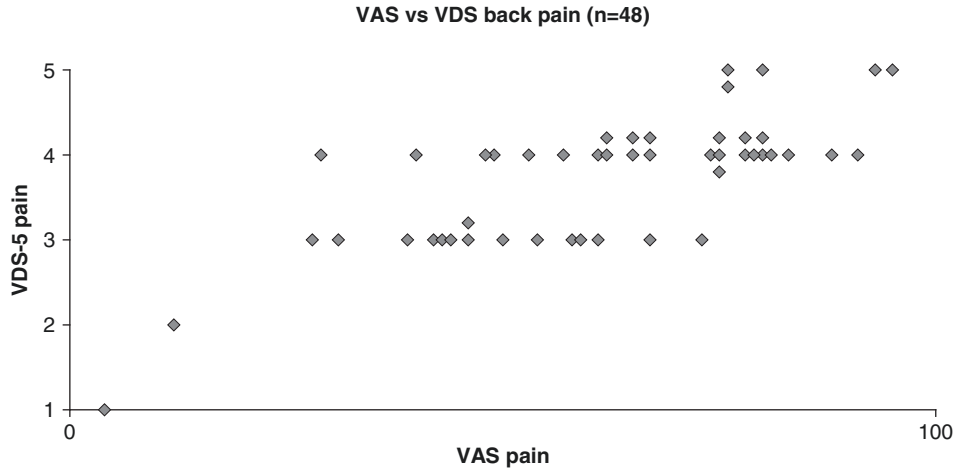
$$P\left[(X_\ell < X_k) \cap (Y_\ell > Y_k)\right] + P\left[(X_\ell > X_k) \cap (Y_\ell < Y_k)\right],$$

is estimated by the empirical measure of disorder D. In this case D equals 0.07, which means that 7% of all possible combinations of different pairs are disordered. The expected pattern of complete order consistency, the rank-transformable pattern of agreement (RTPA), is constructed by pairing off the two sets of distributions of data against each other. The measure of disorder expresses the observed dispersion of pairs from this order consistent distribution of inter-changeability between the scales. The cut-off response values for inter-scale calibration are also provided, and it is obvious that there is no linear correspondence between VAS and discrete scale assessments (see Fig. 2).

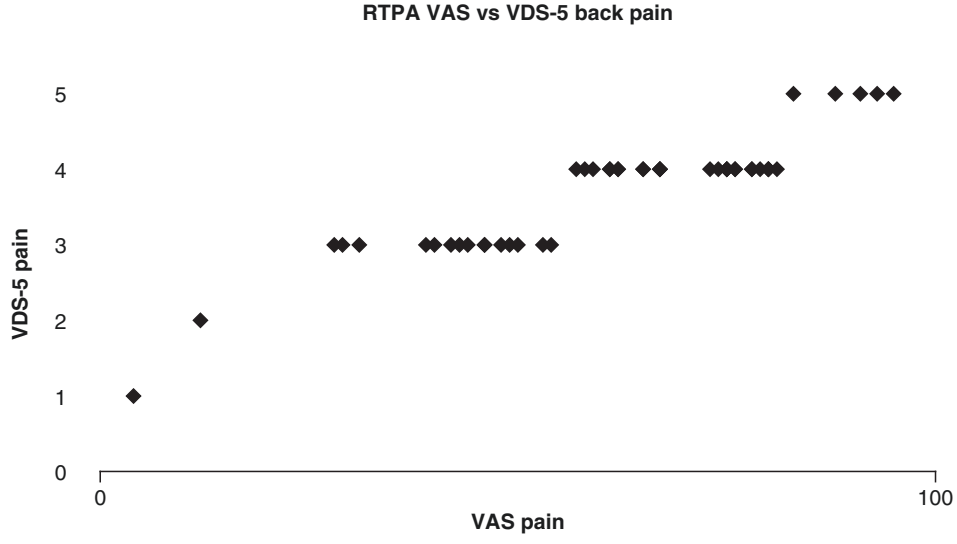
There are other measures that could be applied to evaluation of various kinds of validity of scales. Dependent on the purpose the Spearman rank-order correlations coefficient, The Goodman-Kruskal's gamma, the Kendall's tau-b (see ▶Kendall's Tau), the Somers delta or the Stuart's tau-c could be suitable. Spearman rank order correlation coefficient is a commonly used non-parametric measure of association. However, a strong association does not necessarily mean a high level of order consistency, and does not indicate that two scales are interchangeable.

The Pearson correlation coefficient, the t-test and the ▶Analysis of Variance are also common in validity studies. A serious drawback is that these methods assume normally

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,




**Validity of Scales. Fig. 1** The distribution of paired assessments of back pain on a visual analogue pain scale and a five point verbal descriptive pain scale




**Validity of Scales. Fig. 2** The rank-transformable pattern of agreement, RTPA, uniquely defined by the two sets of frequency distributions of data in Fig. 1

distributed quantitative data, and such requirements are not met by data from rating scales. When applying statistical methods on data that do not have the assumed properties then the results run the risk of being invalid and unreliable.

## About the Author

For biography *see* the entry ►Ranks.

## Cross References

►Analysis of Variance
►Correlation Coefficient
►Kendall's Tau
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Rating Scales
►Scales of Measurement and Choice of Statistical Methods
►Student's *t*-Tests

## References and Further Reading

Svensson E (2000a) Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometrical J 42:417–434

Svensson E (2000b) Concordance between ratings using different scales for the same variable. Stat Med 19(24):3483–3496

Svensson E, Schillberg B, Kling AM, Nyström B (2009) The balanced inventory for spinal disorders. The validity of a disease specific questionnaire for evaluation of outcomes in patients with various spinal disorders. Spine 34(18):1976–1983

# Variables

Rabija Somun- Kapetanović
Professor
University of Sarajevo, Sarajevo, Bosnia and Herzegovina

A variable is a characteristic that can take several values of a set of possible data upon which a measure or a quality can be applied. Thus, a variable varies in value among subjects in a sample or population. Each subject of the observed set has a particular value for a variable.

Examples of variables are gender (with values being female and male), nationality (American, French, German, . . .), level of education (Ph.D., Master, Bachelor, Baccalaureate, . . .), number of children in a family (0, 1, 2, . . .), and annual income in Euros.

Variables can be classified in many ways and terminology varies between different fields. For example, we may classify variables as (a) *qualitative* or *quantitative*, (b) *independent* or *dependent*, (c) *univariate* (one dimensional) or *multivariate* (multidimensional), (d) *latent* (hidden) or *observed*, (e) *endogenous* or *exogenous*, (f) *explanatory, intermediate*, or *response*, and (g) *monitoring* or *moderating*. Classification is further complicated because mixtures of different types occur quite commonly. Depending on the nature of measurement there are also different scales for measuring the variable: *nominal, ordinal, interval*, and *ratio* measurements. The scale of measurement determines the amount of information contained in a set of data and shows the most appropriate statistical methods for analyzing that data. We will focus only on the distinction between qualitative and quantitative variables.

## Qualitative Variables

Qualitative variables contain values that express a quality in a descriptive way, such as sex, nationality, or level of education. Qualitative variables, also called categorical variables, are divided into nominal and ordinal ones.

- *Nominal variables* imply the fact that the labels are unordered. Indeed, there is no criterion that allows determining a label (a value) to be greater than or smaller than other labels. Thus the gender and nationality are nominal variables. Accordingly, the marital status, name, and country of residence are also nominal variables, which are measured on a nominal scale.
- *Ordinal variables* represent labels that can be ordered according to some logical criterion. Hence, the level of education is an ordinal variable as are opinions concerning a subject (excellent, good, poor. . .). The set of labels that satisfies a hierarchical criterion and is measured on an ordinal scale is an ordinal variable.

Mathematical operations are not allowed in qualitative variables, but for ordinal variables, counting and comparison are permitted. Qualitative nominal and ordinal variables can be numerically encoded. Indeed, for instance, it can be supposed that the variable "gender" takes the value 1 for female and 2 for male. Also, if the variable considered is an opinion, the value 1 can be used to represent excellent, 2 for good, and 3 for poor. However, these numbers have no meaning as such and cannot be the object of any mathematical operations.

## Quantitative Variables

Quantitative variables are expressed through measurable values, that is, in terms of numbers. They can be measured on an interval or ratio scale and can be classified as either discrete or continuous.

- *Discrete variables* take only a countable and usually finite number of real values that are the result of a counting process. These variables typically take integer values. For instance, discrete variables are the number of children in a family, the number of students attending a class, and the number of employees in a company.
- *Continuous variables* take an infinite number of real values arising from a measuring process. In practice the number of values that continuous variables can take depends on the precision of the measuring instruments. For instance, the height or the weight is expressed in decimal points when they are measured.

In practice, it is sometimes difficult to distinguish discrete and continuous variables because of the way they are actually measured. Quantitative variables can be used to perform more admissible mathematical operations. The use of quantitative variables is widespread because it contributes to obtaining important results as more statistical methods for analyzing can be applied.

## Cross References

►Data Analysis
►Dummy Variables
►Exchangeability
►Instrumental Variables

►Random Variable
►Rating Scales
►Scales of Measurement
►Scales of Measurement and Choice of Statistical Methods

## References and Further Reading

Anderson DR, Sweeney DJ, Williams TA (2008) Essentials of statistics for business and economics, 5th edn. Cincinnati, South Western Educational Publishing

Berenson ML, Levine DM, Krehbiel TC (2007) Basic business statistics: Concepts and applications, 11th edn. New Yersey, Prentice Hall

Dehon C, Droesbeke JJ, Vermandele C (2008) Eléments de statistique, 5th edn. Editions de l'Universié de Bruxelles, Ellipses, Bruxelles, Paris

Wonnacott TH, Wonnacott RJ (1990) Introductory statistics for business and economics, 4th edn. Wiley, New York

# Variance

Abdulbari Bener[1], Miodrag Lovric[2]
[1]Professor
Weill Cornell Medical College, Doha, Qatar
[2]Professor
University of Kragujevac, Kragujevac, Serbia

The term "variance" was coined by Ronald Fisher in 1918 in his famous paper on population genetics, The Correlation Between Relatives on the Supposition of Mendelian Inheritance, published by Royal Society of Edinburgh: "It is … desirable in analyzing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance …" (p. 399). Interestingly, according to O. Kempthorne, this paper was initially rejected by the Royal Society of London, "probably the reason was that it constituted such a great advance on the thought in the area that the reviewers were unable to make a reasonable assessment."

The variance of a random variable (or a data set) is a measure of variable (data) dispersion or spread around the mean (expected value).

**Definition** Let $X$ be a random variable with second moment $E(X^2)$ and let $\mu = E(X)$ be its mean. The variance of $X$ is defined by (see, e.g., Feller 1968, p. 228)

$$Var(X) = E\left[(X-\mu)^2\right] = E(X^2) - \mu^2. \tag{1}$$

The variance of a random variable is also frequently denoted by $V(X)$, $\sigma_X^2$ or simply $\sigma^2$, when the context is clear. The positive square root of variance is called the standard deviation.

From (1), the variance of $X$ can be interpreted as the "mean of the *squares* of deviations from the mean" (Kendall 1945, p. 39). Since the deviations are squared, it is clear that variance cannot be negative. Variance is a measure of dispersion "since if the values of a random variable $X$ tend to be far from their mean, the variance of $X$ will be larger than the variance of a comparable random variable $Y$ whose values tend to be near their mean" (Mood et al. 1974, p. 67). It is obvious that a constant has variance 0, since there is no spread. Because the deviations are squared, the variance is expressed in the original units squared (inches$^2$, euro$^2$) which are difficult to interpret.

To compute the variance of a random variable, it is required to know the probability distribution of $X$. If $X$ is a discrete random variable, then

$$Var(X) = \sum_i (x_i - \mu)^2 P(X = x_i) = \sum_i x_i^2 P(X = x_i) - \mu^2. \tag{2}$$

When $X$ is a continuous random variable with probability density function $f(x)$, then

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2. \tag{3}$$

*Example* 1 If $X$ has a Uniform distribution on $[a, b]$, with pdf $1/(b-a)$, then

$$E(X) = \frac{1}{b-a}\int_a^b x dx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

and

$$E(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

Hence the variance is equal to

$$Var(X) = E(X^2) - \mu^2 = \frac{(b-a)^2}{12}.$$

The following table provides expressions for variance for some standard univariate discrete and continuous probability distributions.

The Cauchy distribution possesses neither mean nor variance.

Next, we list some important properties of variance.

1. The variance of a constant is 0; in other words, if all observations in the data set are identical, the variance takes its minimum possible value, which is zero.
2. If $b$ is a constant then

| Distribution | Notation | Variance |
|---|---|---|
| Bernoulli | $Be(p)$ | $pq$ |
| Binomial | $Bin(n,p)$ | $npq$ |
| Geometric | $Ge(p)$ | $q/p^2$ |
| Poisson | $Po(\lambda)$ | $\lambda$ |
| Uniform | $U(a,b)$ | $(b-a)^2/12$ |
| Exponential | $Exp(\lambda)$ | $1/\lambda^2$ |
| Normal | $N(\mu,\sigma)$ | $\sigma^2$ |
| Standard Normal | $N(0,1)$ | 1 |
| Student | $t(\nu)$ | $\nu(\nu-2)$ for $\nu>2$ |
| F | $F(\nu_1,\nu_2)$ | $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2>4$ |
| Chi-square | $Chi(\nu)$ | $2\nu$ |

$$Var(X+b) = Var\,X,$$

which means that adding a constant to a random variable does not change the variance.

3. If $a$ and $b$ are constants, then

$$Var(aX+b) = a^2\,Var\,X$$

4. If two variables $X$ and $Y$ are independent, then

$$Var(X+Y) = Var\,X + Var\,Y$$
$$Var(X-Y) = Var\,X + Var\,Y$$

5. The previous property can be generalized, i.e., the variance of the sum of independent random variables is equal to the sum of variances of these random variables

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

This result is called Bienaymé equality (see Loève 1977, p. 12, or Roussas p. 171).

6. If two random variables $X$ and $Y$ are independent and $a$ and $b$ are constants, then

$$Var(aX+bY) = a^2\,Var\,X + b^2\,Var\,Y.$$

In practice, the variance of a population, $\sigma^2$, is usually not known, and therefore it can only be estimated using the information contained in a sample of observations drawn from that population. If $x_1, x_2, \ldots, x_n$ is a random sample of size $n$ selected from a population with mean $\mu$, then the sample variance is usually denoted by $s^2$ and is defined by

$$S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}, \tag{4}$$

where $\bar{x}$ is the sample mean. The sample variance depicts the dispersion of sample observations around the sample mean. The squared deviations in (4) are divided by $n-1$, not by $n$, in order to obtain the unbiased estimator of the population variance, $E(s^2) = \sigma^2$. The factor $1/(n-1)$ increases sample variance enough to make it unbiased. This factor is known as Bessel's correction (after Friedrich Bessel). Although the sample variance defined as in (4) is an unbiased estimator of population variance, the same does not relate to its square root, standard deviation; the sample standard deviation is a *biased* estimate of the population standard deviation.

*Example 2* The first column of the following table contains first five measurements of the speed of light in suitable units (000 km/s) from the classical experiments performed by Michelson in 1879 (data obtained from the Ernest N. Dorsey's 1944 paper "The Velocity of Light").

| $x_i$ | $x_i-\bar{x}$ | $(x_i-\bar{x})^2$ | $x_i^2$ |
|---|---|---|---|
| 299.85 | −0.048 | 0.002304 | 89,910.0225 |
| 299.74 | −0.158 | 0.024964 | 89,844.0676 |
| 299.90 | 0.002 | 0.000004 | 89,940.0100 |
| 300.07 | 0.172 | 0.029584 | 90,042.0049 |
| 299.93 | 0.032 | 0.001024 | 89,958.0049 |
| Σ 1499.49 | 0.000 | 0.057880 | 449,694.1099 |

Since the sample mean is equal to $\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{1499.49}{5} = 299.898$ using the formula given in (4) results in the variance value

$$S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{0.057880}{4} = 0.01447.$$

In the past, instead of the "definitional" formula (4), the following (so-called shorthand) formula was commonly used, but it has become obsolete with the wide access of

statistical software, spreadsheets, and Internet java applets:

$$S^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{449,694.1099 - \frac{1,499.49^2}{5}}{4}$$

$$= 0.01447.$$

## About the Author

Abdulbari Bener, Ph.D., has joined the Department of Public Health at the Weill Cornell Medical College as Research Professor of Public Health. Professor Bener is Director of the Medical Statistics and Epidemiology Department at Hamad Medical Corporation/Qatar. He is also an advisor to the World Health Organization and Adjunct Professor and Coordinator for the postgraduate and master public health programs (MPH) of the School of Epidemiology and Health Sciences, University of Manchester. He is Fellow of Royal Statistical Society (FRSS) and Fellow of Faculty of Public Health (FFPH). Dr Bener holds a Ph.D. degree in Medical Statistics (Biometry) and Genetics from the University College of London, and a B.Sc. degree from Ankara University, Faculty of Education, Department of Management, Planning and Investigation. He completed research fellowships in the Departments of Genetics and Biometry and Statistics and Computer Sciences at the University College of London. He has held academic positions in public health, epidemiology, and statistics at universities in Turkey, Saudi Arabia, Kuwait, the United Arab Emirates, Qatar, and England. Professor Bener has been author or coauthor of more than 430 published journal articles; Editor, Associate Editor, Advisor Editor, and Asst. Editor for several Journals; and Referee for over 23 journals. He has contributed to more than 15 book chapters and supervised thesis of 40 postgraduate students (M.Sc., MPH, M.Phil. and Ph.D.).

## Cross References

►Expected Value
►Mean Median and Mode
►Mean, Median, Mode: An Introduction
►Semi-Variance in Finance
►Standard Deviation
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

## References and Further Reading

Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance. Philos Trans Roy Soc Edinb 52: 399–433

Dorsey EN (1944) The velocity of light. T Am Philos Soc 34(Part 1): 1–110, Table 22

Feller W (1968) An introduction to the probability theory and its applications, 3rd edn. Wiley, New York

Kempthorne O (1968) Book reviews. Am J Hum Genet 20(4): 402–403

Kendall M (1945) The advanced theory of statistics. Charles Griffin, London

Loève M (1977) Probability theory I, 4th edn. Springer, New York

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, 3rd edn. McGraw-Hill, London

Roussas G (1997) A course in mathematical statistics, 2nd edn. Academic, Hardcover

# Variation for Categorical Variables

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

By definition, a categorical variable has a measurement scale that consists of a set of categories, either nominal (i.e., categories without any natural ordering) or ordinal (i.e., categories that are ordered). For a categorical variable with $n$ categories and the probability distribution $P_n = (p_1, \ldots, p_n)$ where $p_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$, some measurement of variation (dispersion) is sometimes of interest. Any such measure will necessarily depend on whether the variable (or set of categories or data) is nominal or ordinal.

## Nominal Case

In the nominal case, variation is generally considered to increase strictly as the probabilities (or proportions) $p_i (i = 1, \ldots, n)$ become increasingly equal, with the variation being maximum for the uniform distribution $P_n^1 = (1/n, \ldots, 1/n)$ and minimum for the degenerate distribution $P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0)$ and for any given $n$. In terms of *majorization theory* (Marshall and Olkin 1979, Ch. 1), this requires that a nominal variation measure be strictly Schur-concave. Another typically imposed requirement is that the measure should be normed to the [0,1] - interval for ease of interpretation.

The best known measures meeting those two requirements are the *index of qualitative variation* ($IQV$), the normed entropy ($H^*$), and the normed form of the *variation ratio* ($VR^*$) defined as follows (e.g., Weisberg 1992):

$$IQV = \left(\frac{n}{n-1}\right)\left(1 - \sum_{i=1}^{n} p_i^2\right), \qquad (1)$$

$$H^* = \frac{-\sum_{i=1}^{n} p_i \log p_i}{\log n}, \tag{2}$$

$$VR^* = \left(\frac{n}{n-1}\right)(1 - \max\{p_1, \ldots, p_n\}). \tag{3}$$

Note that the logarithmic terms in (2) can be to any base since such terms appear both in the numerator and denominator. Those three measures range in value from 0 (when $P_n = P_n^0$) to 1 (when $P_n = P_n^1$) where

$$P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0), \quad P_n^1 = (1/n, \ldots, 1/n), \tag{4}$$

and for any given $n$. The measures in (1) and (2) can be seen to be strictly Schur-concave, while $VR^*$ in (3) is Schur-concave but not strictly so (see Marshall and Olkin 1979, Ch. 3). Also, while $IQV$ and $H^*$ are continuous functions of all the probability components $p_1, \ldots, p_n$, $VR^*$ is a function of only the modal probability.

Although $IQV$ and $H^*$ in (1)–(2) have a number of nice properties, they both lack an important one: they both overstate the true extent of variation. To illustrate this fact, consider $P_2 = (0.75, 0.25)$ for which each element is the arithmetic mean of the corresponding elements of $P_2^0 = (1, 0)$ and $P_2^1 = (0.5, 0.5)$ so that one would reasonably expect that the variation for this $P_2$ should be 0.5, i.e., the mean of the variations for $P_2^0$ and $P_2^1$ (i.e., 0 and 1, respectively). However, one finds the $IQV(0.75, 0.25) = 0.75$ and $H^*(0.75, 0.25) = 0.81$. In order for a variation measure to take on reasonable numerical values, and thereby avoid invalid and misleading results and conclusions, Kvälseth (1995) proposed the following *coefficient of nominal variation* ($CNV$) as a simple transformation of $IQV$:

$$CNV = 1 - \sqrt{1 - IQV}. \tag{5}$$

Besides having the same types of properties as $IQV$, this $CNV$ takes on values that appear to be entirely reasonable throughout the $[0, 1]$ – interval. For instance, $CNV(0.75, 0.25) = 0.50$ as is only reasonable.

Note also that the $CNV$ in (5) can be expressed in terms of metric distances as follows. In terms of the Euclidean distance $d_2(X, Y) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2}$ between the two points $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, $CNV$ can be expressed as

$$CNV = 1 - \frac{d_2\left(P_n, P_n^1\right)}{d_2\left(P_n^0, P_n^1\right)}, \tag{6}$$

for any distribution $P_n$, with $P_n^0$ and $P_n^1$ defined in (4). That is, $CNV$ is the relative extent to which the Euclidean distance $d_2\left(P_n, P_n^1\right)$ is less than its maximum possible value. Or, $CNV$ is the relative (metric) proximity of $P_n$ to $P_n^1$. Thus, the expression in (6) provides $CNV$ with a reasonable interpretation and a solid basis.

In terms of the standard deviation $s$ of $p_1, \ldots, p_n$ (using the usual divisor $n - 1$), it is readily seen that $CNV$ is given by

$$CNV = 1 - s\sqrt{n}. \tag{7}$$

Similarly, in terms of the pair-wise differences between the $p_i$'s,

$$CNV = 1 - \left(\frac{1}{n-1}\sum\sum_{1 \le i < j \le n} |p_i - p_j|^2\right)^{1/2}. \tag{8}$$

A parameterized family of such difference-based variation measures may also be formulated (Kvälseth 1998), but no other family member appears to be superior to $CNV$.

## Ordinal Case

In the ordinal case, and when the order information is accounted for, it is considered that variation is zero for the degenerate distribution $P_n^0$ and maximal for the polarized distribution $P_n^{(1)}$ defined as

$$P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0), \quad P_n^{(1)} = (0.5, 0, \ldots, 0, 0.5), \tag{9}$$

(see, e.g., Leik 1966; Weisberg 1992). When the $n$ categories are ordered, it makes sense to use cumulative probabilities $F_i = \sum_{j=1}^{i} p_j$ for $i = 1, \ldots, n$ with $F_n = 1$. Thus, for any given $P_n = (p_1, \ldots, p_n)$, and for the particular distributions in (9), the following cumulative distributions can be defined:

$$F_{(n)} = (F_1, \ldots, F_{n-1}, 1), \quad F_{(n)}^0 = (0, \ldots, 0, 1, 1, \ldots, 1),$$
$$F_{(n)}^{(1)} = (.5, \ldots, .5, 1). \tag{10}$$

A measure of variation for ordinal categorical data may then be based on cumulative probabilities.

The first such proposed measure appears to be Leik's (1966) ordinal variation measure ($LOV$), which can be expressed as

$$LOV = 1 - \frac{\sum_{i=1}^{n-1} |2F_i - 1|}{n - 1}, \tag{11}$$

which ranges in value from 0 to 1, equals 0 for $F_{(n)}^{(0)}$ and 1 for $F_{(n)}^{(1)}$ in (10). An alternative measure is the *coefficient of ordinal variation* ($COV$) by Kvälseth (1995a,b) defined,

and somewhat analogous to *CNV* in (5), as

$$COV = 1 - \sqrt{1 - \Delta^*}, \quad \Delta^* = \frac{2}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}|i-j|p_i p_j$$

$$= \frac{4}{n-1}\sum_{i=1}^{n-1} F_i(1-F_i) \tag{12}$$

where $COV \in [0,1]$, $COV\left(F^0_{(n)}\right) = 0$, and $COV\left(F^{(1)}_{(n)}\right) = 1$. The *COV* can also be expressed as

$$COV = 1 - \left(\frac{\sum_{i=1}^{n-1}|2F_i - 1|^2}{n-1}\right)^{1/2}. \tag{13}$$

It would appear from (11) and (13) that *LOV* and *COV* are both members of the same family. In fact, expressed in terms of an $\alpha$ – order arithmetic mean, both measures belong to the family of ordinal variation measures

$$OV_\alpha = 1 - \left(\frac{\sum_{i=1}^{n-1}|2F_i - 1|^\alpha}{n-1}\right)^{1/\alpha}, \quad -\infty < \alpha < \infty \tag{14}$$

where $LOV = OV_1$, and $COV = OV_2$. Furthermore, in terms of the Minkowski metric distance of order $\alpha \geq 1$ (i.e., $d_\alpha(X,Y) = \left(\sum_{i=1}^{n}|x_i - y_i|^\alpha\right)^{1/\alpha}$),

$$OV_\alpha = 1 - \frac{d_\alpha\left(F_{(n)}, F^{(1)}_{(n)}\right)}{d_\alpha\left(F^0_{(n)}, F^{(1)}_{(n)}\right)}, \quad \alpha \geq 1 \tag{15}$$

with $F_{(n)}, F^{(0)}_{(n)}$, and $F^{(1)}_{(n)}$ defined in (10). Clearly, $d_\alpha\left(F_{(n)}F^{(1)}_{(n)}\right) \leq d_\alpha\left(F^0_{(n)}, F^{(1)}_{(n)}\right)$ since $|F_i - 0.5| \leq 0.5$ for all $i$. Thus, $OV_\alpha \in [0,1]$, $OV_\alpha\left(F^0_{(n)}\right) = 0$, and $OV_\alpha\left(F^{(1)}_{(n)}\right) = 1$. The expressions in (14) – (15), especially (15), provide interpretations and bases for *LOV* and *COV*, with *LOV* and *COV* being based, respectively, on city-block (Hamming) distances ($\alpha = 1$) and Euclidean distances ($\alpha = 2$) (see also Blair and Lacy 1996).

## Statistical Inferences

For a generic variation measure $V$, consider now $(a)$ that $V(P_n)$ is the sample value based on the distribution $P_n = (p_1, \ldots, p_n)$ of sample probabilities $n_i/N$ for $i = 1, \ldots, n$ with sample size $N = \sum_{i=1}^{n} n_i$ and $(b)$ that $V(\Pi_n)$ is the population value based on the corresponding population distribution $\Pi_n = (\pi_1, \ldots, \pi_n)$. It may then be of interest to construct a confidence interval or test an hypothesis about $V(\Pi_n)$. This can be done using the *delta method* (Agresti 2002, Ch. 14). Accordingly, under multinomial sampling with $N$ reasonably large, $V(P_n)$ is approximately normally distributed with mean $V(\Pi_n)$ and estimated variance

$$\hat{\sigma}^2_V = \frac{1}{N}\left[\sum_{i=1}^{n} p_i \hat{\phi}^2_{V_i} - \left(\sum_{i=1}^{n} p_i \hat{\phi}_{V_i}\right)^2\right], \tag{16}$$

where

$$\hat{\phi}_{V_i} = \left.\frac{\partial V(\Pi_n)}{\partial \pi_i}\right|_{\pi_i = p_i}, \quad i = 1, \ldots, n \tag{17}$$

i.e., $\hat{\phi}_{V_i}$ is the partial derivative of $V(\Pi_n)$ with respect to $\pi_i$, which is then replaced with $p_i$, for $i = 1, \ldots, n$.

In the case of *CNV* in (5), it follows from (17) (with $V = CNV$) that

$$\hat{\phi}_{CNV_i} = \frac{-n}{(n-1)(1-CNV)} p_i, \quad i = 1, \ldots, n,$$

so that; from (16),

$$\hat{\sigma}^2_{CNV} = \left(\frac{1}{N}\right)\left(\frac{n}{(n-1)(1-CNV)}\right)^2\left[\sum_{i=1}^{n} p_i^3 - \left(\sum_{i=1}^{n} p_i^2\right)^2\right]. \tag{18}$$

For the case of *COV* in (13), and with $V = COV$, it is found from (17) that

$$\hat{\phi}_{COV_i} = \begin{cases} \dfrac{2}{(n-1)(1-COV)}\left[n - i - 2\sum_{j=1}^{n-1} F_j\right], & i = 1, \ldots, n-1 \\ 0, & i = n \end{cases} \tag{19}$$

which can then be used to compute $\hat{\sigma}^2_{COV}$ from (16).

As a numerical example, consider the respective multinomial frequencies $n_i = 20, 15, 5, 60$ so that, with $N = 100$, $P_4(0.20, 0.15, 0.05, 0.60)$. From (1) and (5), $IQV = 0.77$ and $CNV = 0.52$. From (18), with $CNV = 0.5170$, $\hat{\sigma}^2_{CNV} = 0.0036$. Therefore, an approximate 95% confidence interval for the population measure $CNV(\Pi_4)$ becomes $0.5170 \pm 1.96\sqrt{0.0036}$ or $(0.40, 0.63)$. If the four categories are ordinal so that $F_i = 0.20, 0.35, 0.40, 1$ for $i = 1, \ldots, 4$, it follows from (13) and (19) that $COV = 0.5959$ and $\hat{\phi}_{COV_i} = 1.8148, 0.8249, 0.3300, 0$ for $i = 1, \ldots, 4$ so that, from (16), with $V = COV$, $\hat{\sigma}^2_{COV} = 0.0051$. Therefore, an approximate 95% confidence interval for $COV(\Pi_4)$ becomes $0.5959 \pm 1.96\sqrt{0.0051}$, or $(0.46, 0.74)$.

## About the Author

For biography *see* the entry ►Entropy.

## Cross References

►Association Measures for Nominal Categorical Variables
►Categorical Data Analysis
►Scales of Measurement
►Variables

## References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken, NJ
Blair J, Lacy MG (1996) Measures of variation for ordinal data as functions of the cumulative distribution. Percept Mot Skills 82:411–418
Kvälseth TO (1995a) Coefficients of variation for nominal and ordinal categorical data. Percept Mot Skills 80:843–847
Kvälseth TO (1995b) Comment on the coefficient of ordinal variation. Percept Mot Skills 81:621–622
Kvälseth TO (1998) On difference – based summary measures. Percept Mot Skills 87:1379–1384
Leik RK (1966) A measure of ordinal consensus. Pacific Sociol Rev 9:85–90
Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic Press, San Deigo, CA
Weisberg HF (1992) Central tendency and variability. (Sage University Paper Series No. 07-083). Sage Publications, Newbury Park, CA

# Vector Autoregressive Models

Helmut Lütkepohl
Professor of Econometrics
European University Institute, Firenze, Italy

Vector autoregressive (VAR) processes are popular in economics and other sciences because they are flexible and simple models for multivariate time series data. In econometrics they became standard tools when Sims (1980) questioned the way classical simultaneous equations models were specified and identified and advocated VAR models as alternatives. A textbook treatment of these models with details on the issues mentioned in the following introductory exposition is available in Lütkepohl (2005).

## The Model Setup

The basic form of a VAR process is

$$y_t = Dd_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ (the prime denotes the transpose) is a vector of $K$ observed time series variables, $d_t$ is a vector of deterministic terms such as a constant, a linear trend and/or seasonal ►dummy variables, $D$ is the associated parameter matrix, the $A_i$'s are $(K \times K)$ parameter matrices attached to the lagged values of $y_t$, $p$ is the lag order or VAR order and $u_t$ is an error process which is assumed to be white noise with zero mean, that is, $E(u_t) = 0$, the covariance matrix, $E\left(u_t u_t'\right) = \Sigma_u$, is time invariant and the $u_t$'s are serially uncorrelated or independent.

VAR models are useful tools for forecasting. If the $u_t$'s are independent white noise, the minimum mean squared error (MSE) $h$-step forecast of $y_{t+h}$ at time $t$ is the conditional expectation given $y_s$, $s \le t$,

$$\begin{aligned} y_{t+h|t} &= E(y_{t+h}|y_t, y_{t-1}, \ldots) \\ &= Dd_{t+h} + A_1 y_{t+h-1|t} + \cdots + A_p y_{t+h-p|t}, \end{aligned}$$

where $y_{t+j|t} = y_{t+j}$ for $j \le 0$. Using this formula, the forecasts can be computed recursively for $h = 1, 2, \ldots$. The forecasts are unbiased, that is, the forecast error $y_{t+h} - y_{t+h|t}$ has mean zero and the forecast error covariance is equal to the MSE matrix. The 1-step ahead forecast errors are the $u_t$'s.

VAR models can also be used for analyzing the relation between the variables involved. For example, Granger (1969) defined a concept of causality which specifies that a variable $y_{1t}$ is causal for a variable $y_{2t}$ if the information in $y_{1t}$ is helpful for improving the forecasts of $y_{2t}$. If the two variables are jointly generated by a VAR process, it turns out that $y_{1t}$ is not Granger-causal for $y_{2t}$ if a simple set of zero restrictions for the coefficients of the VAR process are satisfied. Hence, Granger-causality is easy to check in VAR processes.

Impulse responses offer another possibility for analyzing the relation between the variables of a VAR process by tracing the responses of the variables to impulses hitting the system. If the VAR process is stable and stationary, it has a moving average representation of the form

$$y_t = D^* d_t + \sum_{j=0}^{\infty} \Phi_j u_{t-j},$$

where the $\Phi_j$'s are $(K \times K)$ coefficient matrices which can be computed from the VAR coefficient matrices $A_i$ with $\Phi_0 = I_K$, the $(K \times K)$ identity matrix. This representation can be used for tracing the effect of a specific forecast error through the system. For example, if $u_t = (1, 0, \ldots, 0)'$, the coefficients of the first columns of the $\Phi_j$ matrices represent the marginal reactions of the $y_t$'s. Unfortunately, these so-called *forecast error impulse responses* are often not of interest for economists because they may not reflect properly what actually happens in a system of variables. Given that the components of $u_t$ are typically instantaneously correlated, such shocks or impulses are not likely to appear in isolation. Impulses or shocks of interest for economists are usually instantaneously uncorrelated. They are obtained from the forecast errors, the $u_t$'s, by some transformation,

for example, $\varepsilon_t = Bu_t$ may be a vector of shocks of interest if the $(K \times K)$ matrix $B$ is such that $\varepsilon_t \sim (0, \Sigma_\varepsilon)$ has a diagonal covariance matrix $\Sigma_\varepsilon$. The corresponding moving average representation in terms of the $\varepsilon_t$'s becomes

$$y_t = D^* d_t + \sum_{j=0}^{\infty} \Theta_j \varepsilon_{t-j},$$

where $\Theta_j = \Phi_j B^{-1}$.

There are many $B$ matrices with the property that $Bu_t$ is a random vector with diagonal covariance matrix. Hence, there are many shocks $\varepsilon_t$ of potential interest. Finding those which are interesting from an economic point of view is the subject of *structural VAR analysis*.

## Estimation and Model Specification

In practice the process which has generated the time series under investigation is usually unknown. In that case, if VAR models are regarded as suitable, the lag order has to be specified and the parameters have to be estimated. For a given VAR order $p$, estimation can be conveniently done by equationwise ordinary ▶least squares (OLS). For a sample of size $T$, $y_1, \ldots, y_T$, and assuming that in addition presample values $y_{-p+1}, \ldots, y_0$ are also available, the OLS estimator of the parameters $B = [D, A_1, \ldots, A_p]$ can be written as

$$\hat{B} = \left( \sum_{t=1}^{T} y_t Z_{t-1}' \right) \left( \sum_{t=1}^{T} Z_{t-1} Z_{t-1}' \right)^{-1},$$

where $Z_{t-1}' = \left(d_t', y_{t-1}', \ldots, y_{t-p}'\right)$. Under standard assumptions the estimator is consistent and asymptotically normally distributed. In fact, if the residuals and, hence, the $y_t$'s are normally distributed, that is, $u_t \sim$ i.i.d.$\mathcal{N}(0, \Sigma_u)$, the OLS estimator is equal to the maximum likelihood (ML) estimator with the usual asymptotic optimality properties. If the dimension $K$ of the process is large, then the number of parameters is also large and estimation precision may be low if a sample of typical size in macroeconomic studies is available for estimation. In that case it may be useful to exclude redundant lags of some of the variables from some of the equations and fit so-called subset VAR models. In general, if zero or other restrictions are imposed on the parameter matrices, other estimation methods may be more efficient.

VAR order selection is usually done by sequential tests or model selection criteria (see ▶Model Selection). ▶*Akaike's information criterion* (AIC) is, for instance, a popular model selection criterion (Akaike, 1973). It has the form

$$\mathrm{AIC}(m) = \log\det(\hat{\Sigma}_m) + 2mK^2/T,$$

where $\hat{\Sigma}_m = T^{-1} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$ is the residual covariance matrix of a VAR($m$) model estimated by OLS. The criterion consists of the determinant of the residual covariance matrix which tends to decline with increasing VAR order whereas the penalty term $2mK^2/T$, which involves the number of parameters, grows with $m$. The VAR order is chosen which optimally balances both terms. In other words, models of orders $m = 0, \ldots, p_{\max}$ are estimated and the order $p$ is chosen such that it minimizes the value of AIC.

Once a model is estimated it should be checked that it represents the data features adequately. For this purpose a rich toolkit is available. For example, descriptive tools such as plotting the residuals and residual autocorrelations may help to detect model deficiencies. In addition, more formal methods such as tests for residual autocorrelation, conditional heteroskedasticity, nonnormality and structural stability or tests for parameter redundancy may be applied.

## Extensions

If some of the time series variables to be modeled with a VAR have stochastic trends, that is, they behave similarly to a ▶random walk, then another model setup may be more useful for analyzing especially the trending properties of the variables. Stochastic trends in some of the variables are generated by models with unit roots in the VAR operator, that is, $\det(I_K - A_1 z - \cdots - A_p z^p) = 0$ for $z = 1$. Variables with such trends are nonstationary and not stable. They are often called integrated. They can be made stationary by differencing. Moreover, they are called cointegrated if stationary linear combinations exist or, in other words, if some variables are driven by the same stochastic trend. Cointegration relations are often of particular interest in economic studies. In that case, reparameterizing the standard VAR model such that the cointegration relations appear directly may be useful. The so-called *vector error correction model* (VECM) of the form

$$\Delta y_t = Dd_t + \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + u_t$$

is a simple example of such a reparametrization, where $\Delta$ denotes the differencing operator defined such that $\Delta y_t = y_t - y_{t-1}$, $\Pi = -(I_K - A_1 - \cdots - A_p)$ and $\Gamma_i = -(A_{i+1} + \cdots + A_p)$ for $i = 1, \ldots, p-1$. This parametrization is obtained by subtracting $y_{t-1}$ from both sides of the standard VAR representation and rearranging terms. Its advantage is that $\Pi$ can be decomposed such that the cointegration relations are directly present in the model. More precisely, if all variables are stationary after differencing once, and there are $K - r$ common trends, then the matrix $\Pi$ has rank $r$ and can be decomposed as $\Pi = \alpha\beta'$, where $\alpha$ and $\beta$ are

$(K \times r)$ matrices of rank $r$ and $\beta$ contains the cointegration relations. A detailed statistical analysis of this model is presented in Johansen (1995) (see also Part II of Lütkepohl (2005)).

There are also other extensions of the basic VAR model which are often useful and have been discussed extensively in the associated literature. For instance, in the standard model all observed variables are treated as endogenous, that is, they are jointly generated. This setup often leads to heavily parameterized models, imprecise estimates and poor forecasts. Depending on the context, it may be possible to classify some of the variables as exogenous and consider partial models which condition on some of the variables. The latter variables remain unmodeled.

One may also question the focus on finite order VAR models and allow for an infinite order. This can be done by either augmenting a finite order VAR by a finite order MA term or by accounting explicitly for the fact that the finite order VAR approximates some more general model. Details on these and other extensions are provided, e.g., by Hannan and Deistler (1988) and Lütkepohl (2005).

## About the Author

Professor Lütkepohl was Dean of the School of Economics and Business Administration, Humboldt University, Berlin (1998–2000); Head of the Economics Department, European University Institute, Florence (2006–2008).

## Cross References

►Akaike's Information Criterion
►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Asymptotic Normality
►Econometrics: A Failed Science?
►Forecasting: An Overview
►Likelihood
►Random Walk
►Residuals
►Seasonal Integration and Cointegration in Economic Time Series
►Time Series

## References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds) 2nd International Symposium on Information Theory, Académiai Kiadó, Budapest, pp 267–281

Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438

Hannan EJ, Deistler M (1988) The statistical theory of linear systems. Wiley, New York

Johansen S (1995) Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford

Lütkepohl H (2005) New introduction to multiple time series analysis. Springer-Verlag, Berlin

Sims CA (1980) Macroeconomics and reality. Econometrica 48:1–48