K

Kalman Filtering

MOHINDER S. GREWAL
Professor
California State University, Fullerton, CA, USA

Theoretically, a Kalman filter is an estimator for what is called the linear quadratic Gaussian (*LQG*) problem, which is the problem of estimating the instantaneous "state" of a linear dynamic system perturbed by Gaussian white noise, by using measurements linearly related to the state, but corrupted by Gaussian white noise. The resulting estimator is statistically optimal with respect to any quadratic function of estimation error. R. E. Kalman introduced the "filter" in 1960 (Kalman 1960).

Practically, the Kalman filter is certainly one of the greater discoveries in the history of statistical estimation theory, and one of the greatest discoveries in the twentieth century. It has enabled humankind to do many things that could not have been done without it, and it has become as indispensable as silicon in the makeup of many electronic systems. The Kalman filter's most immediate applications have been for the control of complex dynamic systems, such as continuous manufacturing processes, aircraft, ships, spacecraft, and satellites.

In order to control a dynamic system, one must first know what the system is doing. For these applications, it is not always possible or desirable to measure every variable that one wants to control. The Kalman filter provides a means for inferring the missing information from indirect (and noisy) measurements. In such situations, the Kalman filter is used to estimate the complete state vector from partial state measurements and is called an observer. The Kalman filter is also used to predict the outcome of dynamic systems that people are not likely to control, such as the flow of rivers during flood conditions, the trajectories of celestial bodies, or the prices of traded commodities.

Kalman filtering is an algorithm made from mathematical models. The Kalman filter makes it easier to solve a problem, but it does not solve the problem all by itself. As with any algorithm, it is important to understand its use and function before it can be applied effectively.

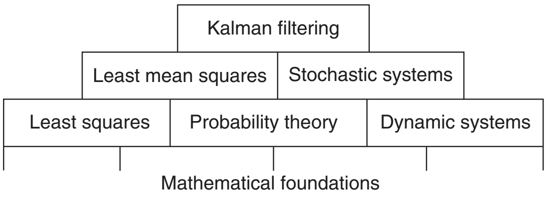
The Kalman filter is a recursive algorithm. It has been called "ideally suited to digital computer implementation," in part because it uses a finite representation of the estimation problem-by a finite number of variables (Gelb et al. 1974). It does, however, assume that these variables are real numbers with infinite precision. Some of the problems encountered in its use arise from the distinction between finite dimension and finite information, and the distinction between finite and manageable problem sizes. These are all issues on the practical side of Kalman filtering that must be considered along with the theory.

It is a complete statistical characterization of an estimation problem. The Kalman filter is much more than an estimator, because it propagates the entire probability distribution of the variables it is tasked to estimate. This is a complete characterization of the current state of knowledge of the dynamic system, including the influence of all past measurements. These probability distributions are also useful for statistical analysis and predictive design of sensor systems.

In a limited context, the Kalman filter is a learning process. It uses a model of the estimation problem that distinguishes between phenomena (what we are able to observe), noumena (what is really going on), and the state of knowledge about the noumena that we can deduce from the phenomena. That state of knowledge is represented by probability distributions. To the extent that those probability distributions represent knowledge of the real world, and the cumulative processing of knowledge is learning, this is a learning process. It is a fairly simple one, but quite effective in many applications. Figure 1 depicts the essential subjects forming the foundations for Kalman filtering theory. Although this shows Kalman filtering as the apex of a pyramid, it is but part of the foundations of another discipline-modern control theory-and a proper subset of statistical decision theory (Grewal and Andrews 2008).

Applications of Kalman filtering encompass many fields. As a tool, the algorithm is used almost exclusively for estimation and performance analysis of estimators and as observers for control of a dynamical system. Except for a few fundamental physical constants, there is hardly anything in the universe that is truly constant. The orbital parameters of the asteroid Ceres are not constant, and

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



Kalman Filtering. Fig. 1 Foundational concepts in Kalman filtering

even the "fixed" stars and continents are moving. Nearly all physical systems are dynamic to some degree. If we want very precise estimates of their characteristics over time, then we must take their dynamics into consideration.

We do not always know the dynamics very precisely. Given this state of partial ignorance, the best we can do is express ignorance more precisely-using probabilities. The Kalman filter allows us to estimate the state of such systems with certain types of random behavior by using such statistical information. A few examples of common estimation problems are shown in Table 1. The third column lists some sensor types that we might use to estimate the state of the corresponding dynamic systems. The objective of design analysis is to determine how best to use these sensor types for a given set of design criteria. These criteria are typically related to estimation accuracy and system cost.

Because the Kalman filter uses a complete description of the probability distribution of its estimation errors to determine the optimal filtering gains, this probability distribution may be used to assess its performance as a function of the design parameters of an estimation system, such as the types of sensors to be used, the locations and orientations of the various sensor types with respect to the system to be estimated, the allowable noise characteristics of the sensors, the prefiltering methods for smoothing sensor noise, the data sampling rates for the various sensor types, and the level of model simplification to reduce implementation requirements.

This analytical capability of the Kalman filter enables system designers to assign "error budgets" to subsystems of an estimation system and to trade off the budget allocations to optimize cost or other measures of performance while achieving a required level of estimation accuracy. Furthermore, it acts like an observer by which all the states not measured by the sensors can be constructed for use in the control system applications.

Linear Estimation

Linear estimation addresses the problem of estimating the state of a linear stochastic system by using measurements or sensor outputs that are linear functions of the state. We suppose that the stochastic systems can be represented by the types of plant and measurement models (for continuous and discrete time) shown as equations in Table 2, with dimensions of the vector and matrix quantities. The measurement and plant noise v_k and w_k, respectively, are assumed to be zero-mean ►Gaussian processes, and the initial value x_o is a Gaussian random variable with known mean x_0 and known covariance matrix P_0. Although the noise sequences w_k and v_k are assumed to be uncorrelated, this restriction can be removed, modifying the estimator equations accordingly.

A summary of equations for the discrete-time Kalman estimator are shown in Table 3, where Q_k, R_k are process and measurement noise covariances, Φ_k is the state transition matrix, H_k is the measurement sensitivity matrix, $\overline{K}_k$ is the Kalman gain. $P_k(-), P_k(+)$ are covariances before and after measurement updates.

Implementation Methods

The Kalman filter's theoretical performance has been characterized by the covariance matrix of estimation uncertainty, which is computed as the solution of a matrix Riccati differential and difference equation. A relationship between optimal deterministic control and optimal estimation problems has been described via the separation principle.

Soon after the Kalman filter was first implemented on computers, it was discovered that the observed mean-square estimation errors were often much larger than the values predicted by the covariance matrix, even with simulated data. The variances of the filter estimation errors were observed to diverge from their theoretical values, and the solutions obtained for the Riccati equations were observed to have negative variances. Riccati equations should have positive or zero variances.

Current work on the Kalman filter primarily focuses on development of robust and numerically stable implementation methods. Numerical stability refers to robustness against roundoff errors. Numerically stable implementation methods are called square root filtering because they use factors of the covariance matrix of estimation uncertainty or its inverse, called the information matrix.

Numerical solution of the Riccati equation tends to be more robust against roundoff errors if Cholesky factors of a symmetrical nonnegative definite matrix P is a matrix C such that $CC^T = P$. Cholesky decomposition algorithms solve for C that is either upper triangular or lower triangular. Another method is modified Cholesky decomposition. Here, algorithms solve for diagonal factors and either a lower triangular factor L or an upper triangular

Kalman Filtering. Table 1 Examples of estimation problems

Application	Dynamic system	Sensor types
Process control	Chemical plant	Pressure, temperature, flow rate, gas analyzer
Flood prediction	River system	Water level, rain gauge, weather radar
Tracking	Spacecraft	Radar, imaging system
Navigation	Ships	Sextant
	Aircraft, missiles	Log
	Smart bombs	Gyroscope
	Automobiles	Accelerometer
	Golf carts	Global Positioning System (GPS) receiver
	Satellites	GPS receiver
	Space shuttle	GPS receiver, Inertial Navig. Systems (INS)

Kalman Filtering. Table 2 Linear plant and measurement models

<table>
<tr><th>Model</th><th colspan="2">Continuous time</th><th colspan="2">Discrete time</th></tr>
<tr><td>Plant</td><td colspan="2">$\dot{x}(t) = F(t)x(t) + w(t)$</td><td colspan="2">$x_k = \Phi_{k-1}x_{k-1} + w_{k-1}$</td></tr>
<tr><td>Measurement</td><td colspan="2">$z(t) = H(t)x(t) + v(t)$</td><td colspan="2">$z_k = H_k x_k + v_k$</td></tr>
<tr><td>Plant noise</td><td colspan="2">$E\langle w(t)\rangle = 0$
$E\langle w(t)w^T(s)\rangle = \delta(t-s)Q(t)$</td><td colspan="2">$E\langle w_k\rangle = 0$
$E\left\langle w_k w_i^T\right\rangle = \Delta(k-i)Q_k$</td></tr>
<tr><td>Observation noise</td><td colspan="2">$E\langle v(t)\rangle = 0$
$E\langle v(t)v^T(s)\rangle = \delta(t-s)R(t)$</td><td colspan="2">$E\langle v_k\rangle = 0$
$E\left\langle v_k v_i^T\right\rangle = \Delta(k-i)R_k$</td></tr>
<tr><th>(Linear model)</th><th>Symbol</th><th>Dimensions</th><th>Symbol</th><th>Dimensions</th></tr>
<tr><td>Dimensions of vectors and matrices</td><td>x, w</td><td>$n \times 1$</td><td>Φ, Q</td><td>$n \times n$</td></tr>
<tr><td></td><td>z, v</td><td>$\ell \times 1$</td><td>H</td><td>$\ell \times n$</td></tr>
<tr><td></td><td>R</td><td>$\ell \times \ell$</td><td>Δ, δ</td><td>Scalar</td></tr>
</table>

factor U such that $P = UD_uU^T = LD_LL^T$ where D_L and D_u are diagonal factors with nonnegative diagonal elements. Another implementation method uses square root information filters that use a symmetric product factorization of the information matrix P^{-1}. Another implementation with improved numerical properties is the “sigmaRho filter.” Individual terms of the covariance matrix can be interpreted as $P_{ij} = \sigma_i\sigma_j\rho_{ij}$ where P_{ij} is the ijth of the covariance matrix, σ_i is the standard deviation of the estimate of the ith state component, and ρ_{ij} is the correlation coefficient between ith and jth state component (Grewal and Kain 2010).

Alternative Kalman filter implementations use these factors of the covariance matrix (or its inverse) in three types of filter operations: (1) temporal updates, (2) observation updates, and (3) combined updates (temporal and observation). The basic algorithm methods used in these alternative Kalman filter implementations fall into four general categories. The first three of these categories are concerned with decomposing matrices into triangular factors and maintaining the triangular form of the factors through all the Kalman filtering operation. The fourth category includes standard matrix operations (multiplication, inversion, etc.) that have been specialized for triangular

Kalman Filtering. Table 3 Discrete-time Kalman filter equations

System dynamic model	$x_k = \Phi_{k-1} x_{k-1} + w_{k-1}, \quad w_k \sim N(0, Q_k)$
Measurement model	$z_k = H_k x_k = v_k, \quad v_k \sim N(0, R_k)$
Initial conditions	$E\langle x_0 \rangle = \widehat{x}_0, \quad E\left\langle \widetilde{x}_0 \widetilde{x}_0^T \right\rangle = P_0$
Independence assumption	$E\left\langle w_k v_j^T \right\rangle = 0$ for all k and j
State estimate extrapolation	$\widehat{x}_k(-) = \Phi_{k-1} \widehat{x}_{k-1}(+)$
Error covariance extrapolation	$P_k(-) = \Phi_{k-1} P_{k-1}(+) \Phi_{k-1}^T + Q_{k-1}$
State estimate observational update	$\widehat{x}(+) = \widehat{x}_k(-) + \overline{K}_k [z_k - H_k \widehat{x}_k(-)]$
Error covariance update	$P_k(+) = [I - \overline{K}_k H_k] P_k(-)$
Kalman gain matrix	$\overline{K}_k = P_k(-) H_k^T \left[H_k P_k(-) H_k^T + R_k\right]^{-1}$

matrices. These implementation methods have succeeded where the conventional Kalman filter implementations have failed (Grewal and Andrews 2008).

Even though uses are being explored in virtually every discipline, research is particularly intense on successful implementation of Kalman filtering to global positioning systems (GPS), inertial navigation systems (INS), and guidance and navigation. GPS is a satellite-based system that has demonstrated unprecedented levels of positioning accuracy, leading to its extensive use in both military and civil arenas. The central problem for GPS receivers is the precise estimation of position, velocity, and time, based on noisy observations of satellite signals. This provides an ideal setting for the use of Kalman filtering. GPS technology is used in automobile, aircraft, missiles, ships, agriculture, and surveying. Currently, the Federal Aviation Agency (FAA) is sponsoring research on the development of wide-area augmentation system (WAAS) for precision landing and navigation of commercial aircraft (Grewal et al. 2007).

Kalman filters are used in bioengineering, traffic systems, photogrammetry, and myriad process controls. The Kalman filter is observer, parameter identifier in modeling, predictor, filter, and smoother in a wide variety of applications. It has become integral to twenty-first century technology (Grewal and Kain 2010; Grewal et al. 2007).

About the Author

Dr. Mohinder S. Grewal, P.E., coauthored *Kalman Filtering: Theory & Practice Using MATLAB* (with A.P. Andrews, 3rd edition, Wiley & Sons 2008) and *Global Positioning Systems, Inertial Navigation, & Integration* (with L.R. Weill and A.P. Andrews, 2nd edition, Wiley & Sons 2007). Dr. Grewal has consulted with Raytheon Systems, Geodetics, Boeing Company, Lockheed-Martin, and Northrop on application of Kalman filtering. He has published over 60 papers in IEEE and ION refereed journals and proceedings, including the Institute of Navigation's Redbook. Currently, Dr. Grewal is Professor of Electrical Engineering at California State University, Fullerton, in Fullerton, California, where he received the 2009 Outstanding Professor award. He is an architect of the GEO Uplink Subsystem (GUS) for the Wide Area Augmentation System (WAAS), including the GUS clock steering algorithms, and holds two patents in this area. His current research interest is in the area of application of GPS, INS integration to navigation. Dr. Grewal is a member of the Institute of Navigation, Senior Member of IEEE, and a Fellow of the Institute for the Advancement of Engineering.

Cross References

►Conditional Expectation and Probability
►Intervention Analysis in Time Series
►Statistical Inference for Stochastic Processes
►Structural Time Series Models

References and Further Reading

Gelb A et al (1974) Applied optimal estimation. MIT Press, Cambridge

Grewal MS, Andrews AP (2008) Kalman filtering theory and practice using MATLAB, 3rd edn. Wiley, New York

Grewal MS, Kain J (September 2010) Kalman filter implementation with improved numerical properties, Transactions on automatic control, vol 55(9)

Grewal MS, Weill LR, Andrews AP (2007) Global positioning systems, inertial navigation, & integration, 2nd edn. Wiley, New York

Kalman RE (1960) A new approach to linear filtering and prediction problems. ASME J Basic Eng 82:34–45

Kaplan-Meier Estimator

Irène Gijbels
Professor, Head of Statistics Section
Katholieke Universiteit Leuven, Leuven, Belgium

The Kaplan-Meier estimator estimates the distribution function of a lifetime T based on a sample of randomly right censored observations. In survival analysis the lifetime T is a nonnegative random variable describing the time until a certain event of interest happens. In medical applications examples of such events are the time till death of a patient suffering from a specific disease, the time till recovery of a disease after the start of the treatment, or the time till remission after the curing of a patient. A typical difficulty in survival analysis is that the observations might be incomplete. For example, when studying the time till death of a patient with a specific disease, the patient might die from another cause. As a consequence the lifetime of this patient is not observed, and is only known to be larger than the time till the patient was "censored" by the other cause of death. Such a type of censoring mechanism is called right random censorship. Other areas of applications in which one encounters this type of data are reliability in industry and analysis of duration data in economics, to just name a few.

Let $T_1, T_2, \cdots, T_n$ denote n independent and identically distributed random variables, all having the same distribution as the lifetime T. Denote by $F(t) = P\{T \leq t\}$ the cumulative distribution function of T. Due to the right random censoring, the lifetime T might be censored by a censoring time C, having cumulative distribution function G. Associated at each lifetime T_i there is a censoring time C_i. Under a right random censorship model the observations consist of the pairs

$$(Z_i, \delta_i) \quad \text{where} \quad Z_i = \min(T_i, C_i) \quad \text{and}$$
$$\delta_i = I\{T_i \leq C_i\} \qquad i = 1, \cdots, n.$$

The indicator random variable $\delta = I\{T \leq C\}$ takes value 1 when the lifetime T is observed, and is 0 when the censoring time is observed instead. A crucial assumption in this model is that the lifetime T_i (also often called survival time) is independent of the censoring time C_i for all individuals. An observation (Z_i, δ_i) is called uncensored when $\delta_i = 1$ and hence the survival time T_i for individual i has been observed. When $\delta_i = 0$ the observed time is the censoring time C_i and one only has the incomplete observation that $T_i > C_i$.

Kaplan and Meier (1958) studied how to estimate the survival function $S(t) = 1 - F(t) = P\{T > t\}$, based on observations $(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$ from n patients. The estimation method does not make any assumptions about a specific form of the cumulative distribution functions F and G, and is therefore a nonparametric estimate. When there are no tied observations the estimate is defined as

$$\widehat{S}(t) = \begin{cases} \prod_{j:Z_{(j)} \leq t} \left(\dfrac{n-j}{n-j+1} \right)^{\delta_{(j)}} & \text{if} \quad t < Z_{(n)} \\ 0 & \text{if} \quad t \geq Z_{(n)}, \end{cases}$$

where $Z_{(1)} \leq Z_{(2)} \cdots \leq Z_{(n)}$ denote the ordered Z_i's, and $\delta_{(i)}$ is the indicator variable associated with $Z_{(i)}$. In case of a tie between a censored and an uncensored observation, the convention is that the uncensored observation happened just before the censored observation. An equivalent expression, for $t < Z_{(n)}$, is

$$\widehat{S}(t) = \prod_{j:Z_{(j)} \leq t} \left(1 - \frac{\delta_{(j)}}{n-j+1} \right).$$

In case of n complete observations, $\delta_{(i)} = 1$ for all individuals, and the Kaplan-Meier estimate reduces to $S_n(t) = 1 - \#\{j : Z_j \leq t\}/n$, i.e., one minus the empirical cumulative distribution function. The latter estimate is a decreasing step function with downward jumps of size $1/n$ at each observation $Z_j = T_j$.

Suppose now that there are tied observations, and that there are only r distinct observations. Denote by $Z_{(1)} \leq Z_{(2)} \leq Z_{(r)}$ the r ordered different observations, and by d_j the number of times that $Z_{(j)}$ has been observed. Then, for $t < Z_{(r)}$, the Kaplan-Meier estimate is defined as

$$\prod_{j:Z_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right)^{\delta_{(j)}},$$

where n_j denotes the number of individuals in the sample that are at risk at time point $Z_{(j)}$, i.e., the set of individuals that are still "alive" just before the time point $Z_{(j)}$. The Kaplan-Meier estimate is also called the product-limit estimate.

In studies of life tables (see ▸Life Table), the actuarial estimate for the survival function was already around much earlier. One of the first references for the product-limit estimate, obtained as a limiting case of the actuarial estimate, is Bähmer (1912).

The Kaplan-Meier estimate of the survival function $S = 1 - F$ is a decreasing step function, which jumps at the uncensored observations but remains constant when passing a censored observation. In contrast to the empirical estimate S_n based on a complete sample of size n, the sizes of the jumps are random.

The construction of the Kaplan-Meier estimate $\widehat{S}(t)$ also has a very simple interpretation due to Efron (1967). The mass $1/n$ that is normally attached to each observation in the empirical estimate for S, is now for a censored observation redistributed equally over all observations that are larger than the considered one.

Kaplan and Meier (1958) give the maximum likelihood derivation of the product-limit estimate and discuss mean and variance properties of it. An estimate for the variance of $\widehat{S}(t)$ was already established by Greenwood (1926). Greenwood's formula estimates the variance of $\widehat{S}(t)$ by

$$\widehat{\mathrm{Var}}\left(\widehat{S}(t)\right) = \left(\widehat{S}(t)\right)^2 \sum_{j:Z_{(j)}\leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

This estimate can be used to construct confidence intervals.

The theoretical properties of the Kaplan-Meier estimate have been studied extensively. For example, weak convergence of the process $\sqrt{n}\left(\widehat{S}(t) - S(t)\right)$ to a Gaussian process was established by Breslow and Crowley (1974), and uniform strong consistency of the Kaplan-Meier estimate was proven by Winter et al. (1978).

The Kaplan-Meier estimate is implemented in most statistical software packages, and is a standard statistical tool in survival analysis.

About the Author

Professor Irène Gijbels is Head of the Statistics Section, Department of Mathematics, of the Katholieke Universiteit Leuven, and is the Chair of the Research Commission of the Leuven Statistics Research Center. She is Fellow of the Institute of Mathematical Statistics and Fellow of the American Statistical Association. She has (co-)authored over 70 articles in internationally reviewed scientific journals.

Cross References

- ►Astrostatistics
- ►Life Table
- ►Mean Residual Life
- ►Modeling Survival Data
- ►Survival Data

References and Further Reading

Böhmer PE (1912) Theorie der unabhängigen Wahrscheinlichkeiten. Rapports Mémoires et Procès-verbaux de Septième Congrès International d'Actuaries. Amsterdam, vol 2. pp 327–343

Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann Stat 2:437–453

Efron B (1967) The two sample problem with censored data. In: Proceedings of the 5th Berkeley Symposium, vol 4. pp 831–853

Greenwood M (1926) The natural duration of cancer. In: Reports on public health and medical subjects, vol 33. His Majesty's Stationery Office, London

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Winter BB, Földes A, Rejtö L (1978) Glivenko-Cantelli theorems for the product limit estimate. Probl Control Inform 7:213–225

Kappa Coefficient of Agreement

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

Introduction

When two (or more) observers are independently classifying items or observations into the same set of k mutually exclusive and exhaustive categories, it may be of interest to use a measure that summarizes the extent to which the observers agree in their classifications. The Kappa coefficient first proposed by Cohen (1960) is one such measure.

In order to define this measure, let p_{ij} be the proportion of items assigned to category i by Observer 1 and to category j by Observer 2. Furthermore, let p_{i+} be the proportion of items assigned to category i by Observer 1 and p_{+j} the proportion of items assigned to category j by Observer 2. If these proportions or sample probabilities are represented in terms of a two-way contingency table with k rows and k columns, then p_{ij} becomes the probability in the cell corresponding to row i and column j. With row i being the same as column i for $i = 1, \ldots, k$, the diagonal of this table with probabilities p_{ii} $(i = 1, \ldots, k)$ represents the agreement probabilities, whereas the off-diagonal entries represent the disagreement probabilities.

The observed probability of agreement $P_{ao} = \sum_{i=1}^{k} p_{ii}$ could, of course, be used as an agreement measure. However, since there may be some agreement between the two observers based purely on chance, it seems reasonable that a measure of interobserver agreement should also account for the agreement expected by chance. By defining chance-expected agreement probability as $P_{ac} = \sum_{i=1}^{k} p_{i+}p_{+i}$ and based on independent classifications between the two observers, Cohen (1960) introduced the Kappa coefficient as

$$K = \frac{P_{ao} - P_{ac}}{1 - P_{ac}} \qquad (1)$$

where $K \leq 1$, with $K = 1$ in the case of perfect agreement, $K = 0$ when the observed agreement probability equals that due to chance, and $K < 0$ if the observed agreement probability is less than the chance-expected one.

Kappa can alternatively be expressed in terms of the observed probability of disagreement P_{do} and the chance-expected probability of disagreement P_{dc} as

$$K = 1 - \frac{P_{do}}{P_{dc}}; \quad P_{do} = \sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} p_{ij}, \quad P_{dc} = \sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} p_{i+}p_{+j} \quad (2)$$

The form of K in (1) is the most frequently used one. However, it should be pointed out that the normalization used in (1), i.e., using the dominator $1 - P_{ac}$ such that $K \leq 1$, is not unique. In fact, there are infinitely many such normalizations. Thus, for any given marginal probability distributions $\{p_{i+}\}$ and $\{p_{+j}\}$, one could, for example, instead of the denominator in (1), use $\sum_{i=1}^{k} \left(\frac{1}{2}p_{i+}^{\alpha} + \frac{1}{2}p_{+i}^{\alpha}\right)^{1/\alpha} - P_{ac}$ for any real value of α. For $\alpha \to -\infty$, this alternative denominator would become $\sum_{i=1}^{k} \min\{p_{i+}, p_{+i}\} - P_{ac}$. No such non-uniqueness issue would arise by using the form of K in (2). This K also has the simple interpretation of being the proportional difference between the chance and observed disagreement probabilities, i.e., the relative extent to which P_{do} is less than P_{dc}.

Weighted Kappa

In the case when the $k > 2$ categories are ordinal, or also possibly in some cases involving nominal categories, some disagreements may be considered more serious than others. Consequently, the weighted Kappa (K_w) was introduced (Cohen 1968). In terms of the set of nonnegative agreement weights $v_{ij} \in [0,1]$ and disagreement weights $w_{ij} \in [0,1]$ for all i and j, K_w can be defined as

$$K_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}} \quad (3)$$

$$= 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}} \quad (4)$$

where $w_{ij} = 1 - v_{ij}, w_{ij} = w_{ji}$ for all i and j, and $w_{ij} = 0$ for all $i = j$. Of course, when w_{ij} is the same for all $i \neq j$, K_w reduces to K in (1) – (2). From (4), which seems to be the most intuitive and preferred form of K_w, it is clear that $K_w \leq 1$, with $K_w = 1$ if, and only if, $p_{ij} = 0$ for all $i \neq j$ (i.e., if all disagreement cells have zero probability), $K_w = 0$ under independence (i.e., $p_{ij} = p_{i+}p_{+j}$ for all i and j), and K_w may also take on negative values. Unless there are particular justifications to the contrary, the most logical choice of weights would seem to be $w_{ij} = |i - j|/(k-1)$ or $w_{ij} = (i-j)^2/(k-1)^2$ for all i and j.

Specific Category Kappa

Besides measuring the overall agreement between two observers, it may be of interest to determine their extent of agreement on specific categories. As first proposed by Spitzer et al. (1967) (see also (Fleiss et al. 2003)), such measurement required the original $k \times k$ table to be collapsed into 2×2 tables, one for each specific category. Thus, to measure the agreement on a specific category s, the original $k \times k$ table would need to be collapsed into a 2×2 table with one category being the original s category and the other category being "all others". The agreement measurement K_s was then obtained by computing the value of K in (1) based on the collapsed 2×2 table.

As an alternative way of obtaining the agreement K_s on the specific category s, without the need to collapse the original $k \times k$ table, Kvålseth (1989) proposed the specific – category Kappa as

$$K_s = \frac{p_{ss} - p_{s+}p_{+s}}{\bar{p}_s - p_{s+}p_{+s}}, \quad \bar{p}_s = \frac{p_{s+} + p_{+s}}{2} \quad (5)$$

The K_s can alternatively be expressed as

$$K_s = 1 - \frac{\sum\sum_{D_s} p_{ij}}{\sum\sum_{D_s} p_{i+}p_{+j}} \quad (6)$$

where $\sum\sum_{D_s}$ denotes the summation over all disagreement cells for category s, i.e.,

$$D_s = \{(s,j) \text{ for all } j \neq s \text{ and } (i,s) \text{ for all } i \neq s\} \quad (7)$$

From (6), K_s is the proportional difference between the chance - expected disagreement and the observed disagreement for the specific category s. Note that K in (1) and (2) are weighted arithmetic means of K_s in (5) and (6), respectively, for $s = 1, \ldots, k$, with the weights being based on the denominators in (6) – (7).

When K_s is expressed as in (6), an extension to the case when disagreements should be unequally weighted is rather obvious. Thus, for disagreement weights $w_{ij} \in [0,1]$ for all i and j, with $w_{ij} = 0$ for all $i = j$, the following weighted specific – category Kappa has been proposed (Kvålseth 2003):

$$K_{ws} = 1 - \frac{\sum\sum_{D_s} w_{ij}p_{ij}}{\sum\sum_{D_s} w_{ij}p_{i+}p_{+j}} \quad (8)$$

where D_s is the set of disagreement cells in (7). When w_{ij} is the same for all $(i,j) \in D_s$, (8) reduces to (6). Note also that

K_w in (4) is a weighted arithmetic mean of the K_{ws} for $s = 1, \ldots, k$, with the weights based on the denominator in (8).

The possible values of K_s and K_{ws} range from 1 (when the disagreement probabilities for category s are all zero), through 0 (under the independence $p_{sj} = p_{s+}p_{+j}$ for all j and $p_{is} = p_{i+}p_{+s}$ for all i), and to negative values when observed disagreement exceeds chance - expected disagreement for category s.

Conditional and Asymmetric Kappa

Light (1971) considered the agreement between two observers for only those items (observations) that Observer 1 assigned to category i, with the conditional Kappa defined as

$$K_{2|1}^{(i)} = \frac{p_{ii}/p_{i+} - p_{+i}}{1 - p_{+i}} \tag{9}$$

This measure can also be expressed as

$$K_{2|1}^{(i)} = 1 - \frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{k} p_{ij}}{\sum\limits_{\substack{j=1\\ j\neq i}}^{k} p_{i+}p_{+j}} \tag{10}$$

which immediately suggests the following weighted form (Kvålseth 1985):

$$K_{2|1,w}^{(i)} = 1 - \frac{\sum\limits_{\substack{j=1\\ j\neq i}}^{k} w_{ij}p_{ij}}{\sum\limits_{\substack{j=1\\ j\neq i}}^{k} w_{ij}p_{i+}p_{+j}} \tag{11}$$

Whereas K in (1) – (2) and K_w in (3) – (4) treat the two observers equivalently, i.e., these measures are effectively symmetric, asymmetric versions of Kappa may be defined in terms of the weighted means of the measures in (9) – (11) as

$$\overline{K}_{2|1} = \sum_{i=1}^{k} p_{i+}K_{2|1}^{(i)}, \; \overline{K}_{2|1,w} = \sum_{i=1}^{k} p_{i+}K_{2|1,w}^{(i)} \tag{12}$$

Such measures as in (12) may be appropriate if Observer 1 is to be designated as the "standard" against which classifications by Observer 2 are to be compared (Kvålseth 1991).

Statistical Inferences

Consider now that the above Kappa coefficients are estimates (and estimators) based on sample probabilities (proportions) $p_{ij} = n_{ij}/N$ for $i = 1, \ldots, k$ and $j = 1, \ldots, k$ and sample size $N = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}$, with $\{\pi_{ij}\}$ being the corresponding population probabilities. It may then be of interest to make statistical inferences, especially confidence - interval construction, about the corresponding $\{\pi_{ij}\}$ – based population coefficients or measures. Such approximate inferences can be made based on the *delta method* (Bishop et al. 1975).

Consequently, under multinomial sampling and when N is reasonably large, the various Kappa coefficients introduced above are approximately normally distributed with means equal to the corresponding population coefficients and with variances that can be determined as follows. Since those Kappa coefficients can all be expressed in terms of K_w in (4) by special choices among the set of weights $\{w_{ij}\}$, it is sufficient to determine the variance of (the estimator) K_w. For instance, in the case of K_s in (6) and K_{ws} in (8), one can simply set $w_{ij} = 0$ in (4) for all cells that do not belong to D_s in (7). Thus, the estimated variance of K_w has been given in (Kvålseth 2003) as

$$\mathrm{Var}(K_w) = \left(NF_w^2\right)^{-1}\left\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}E_{ij}^2 - \left[K_w - (1-F_w)(1-K_w)\right]^2\right\} \tag{13a}$$

where

$$E_{ij} = 1 - w_{ij} - (2 - \overline{w}_{i+} - \overline{w}_{+j})(1 - K_w) \tag{13b}$$

$$\overline{w}_{i+} = \sum_{j=1}^{k} w_{ij}p_{+j}, \; \overline{w}_{+j} = \sum_{i=1}^{k} w_{ij}p_{i+}, \; F_w = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}. \tag{13c}$$

Note that F_w is the dominator of K_w in (4).

Example As an example of this inference procedure, consider the (fictitious) probabilities (proportions) in Table 1.

In the case of category 1, e.g., it follows from (5) or (6) and Table 1 that the interobserver agreement $K_1 = 0.72$.

Kappa Coefficient of Agreement. Table 1 Results from two observers' classifications with three categories

	Observer 2			
Observer 1	1	2	3	Total
1	0.40	0.07	0.01	0.48
2	0.04	0.20	0.06	0.30
3	0.02	0.05	0.15	0.22
Total	0.46	0.32	0.22	1.00

If, however, the categories in Table 1 are ordinal and the weights $w_{ij} = |i - j|/(k - 1)$ are used, it is found from (7) – (8) and Table 1, with D_1 consisting of the cells (1,2), (1,3), (2,1), and (3,1), that $K_{w1} = 0.76$. Similarly, $K_2 = 0.49, K_3 = 0.59$, $K_{w2} = 0.49$, and $K_{w3} = 0.69$, whereas, from (1) – (4), $K = 0.61$ and $K_w = 0.67$.

In order to construct a confidence interval for the population equivalent of K_{w1}, (13) can be used by (a) setting $w_{ij} = 0$ for those cells that do not belong to D_1 in (7), i.e., the cells (2,2), (2,3), (3,2) and (3,3) and (b) replacing K_w and F_w with K_{ws} and F_{ws} (the denominator of K_{ws}). Thus, with $K_{w1} = 1 - 0.0850/0.3526 = 0.7589$ (and $F_{w1} = 0.3526$), it is found from (13b) – (13c) that $E_{11} = 0.6986$, $E_{12} = 0.1673, \ldots, E_{33} = 0.7444$ so that, from (13a), if the data in Table 1 are based on sample size $N = 100$, it is found that $\text{Var}(K_{w1}) = 0.0042$. Consequently, an approximate 95% confidence interval for the population equivalent of K_{w1} becomes $0.76 \pm 1.96\sqrt{0.0042}$, or $(0.63, 0.89)$. By comparison, setting $w_{12} = w_{13} = w_{21} = w_{31} = 1$ and all other $w_{ij} = 0$, it is found in (Kvålseth 2003) that a 95% confidence interval for the population equivalent of the unweighted K_1 is (0.58, 0.86).

Concluding Comment

While the overall Kappa and its weighted form in (1) – (4) are the most popular measures of interobserver agreement, they are not without some criticism or controversy. In particular, they depend strongly on the marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$ so that, when those distributions are highly uneven (non-uniform), values of Kappa tend to be unreasonably small. Also, since the p_{ii} $(i = 1, \ldots, k)$ are included in the marginal distributions, the agreement probabilities enter into both the overall probability of agreement as observed and as expected by chance.

About the Author

For biography *see* the entry ►Entropy.

Cross References

►Measures of Agreement
►Psychiatry, Statistics in

References and Further Reading

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis, Ch. 14. MIT Press, Cambridge

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Cohen J (1968) Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220

Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, Ch. 18, 3rd edn. Wiley, Hoboken

Kvålseth TO (1985) Weighted conditional Kappa. B Psychonomic Soc 23:503–505

Kvålseth TO (1989) Note on Cohen's Kappa. Psychol Rep 65:223–226

Kvålseth TO (1991) A coefficient of agreement for nominal scales: an asymmetric version of Kappa. Educ Psychol Meas 51:95–101

Kvålseth TO (2003) Weighted specific – category Kappa measure of interobserver agreement. Psychol Rep 93:1283–1290

Light RJ (1971) Measure of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 76:365–377

Spitzer RL, Cohen J, Fleiss JL, Endicott J (1967) Quantification of agreement in psychiatric diagnosis. Arch Gen Psychiat 17:83–87

Kendall's Tau

Llukan Puka
Professor
University of Tirana, Tirana, Albania

Kendall's *Tau* is a nonparametric measure of the degree of correlation. It was introduced by Maurice Kendall in 1938 (Kendall 1938).

Kendall's *Tau* measures the strength of the relationship between two ordinal level variables. Together with Spearman's rank correlation coefficient, they are two widely accepted measures of rank correlations and more popular rank correlation statistics.

It is required that the two variables, X and Y, are paired observations. Then, provided both variables are at least ordinal, it would be possible to calculate the correlation between them. In general, application of the product-moment correlation coefficient is limited by the requirement that the trend must be linear. A less restrictive measure of correlation is based on the probabilistic notion that the correlation between variables X and Y is strong if on average, there is a high probability that an increase in X will be accompanied by an increase in Y (or decrease in Y). Then the only limitation imposed on the trend line is that it should be either continually increasing or continually decreasing.

One of the properties of coefficients that adopt this notion of correlation, like *Kendall's Tau* coefficient, is that the definition of the correlation depends only on the ranks of the data values and not on the numerical values. To this end, they can be applied either to data from scaled variables that has been converted to ranks, or to ordered categorical variables.

Formula for Calculation of Kendall's Tau Coefficient, (Hollander and Wolfe 1998)

For any sample of n paired observations of a bivariate variables (X, Y), there are $m = \frac{n(n-1)}{2}$ possible comparisons of points (X_i, Y_i) and (X_j, Y_j). A pair of observation data set (X_i, Y_i), (X_j, Y_j) is called concordant if $X_j - X_i$ and $Y_j - Y_i$ has the same sign. Otherwise, if they have opposite signs, the pair is called discordant. If $X_i = X_j$, or $Y_i = Y_j$ or both, the comparison is called a "tie." Ties are not counted as concordant or discordant.

If C is the number of pairs that are concordant and D is the number of pairs that are discordant, then the value *Tau* of Kendall's *Tau* is

$$Tau = \frac{C - D}{m}$$

The quantity $S = C - D$ is known as Kendall S. A predominance of concordant pairs resulting in a large positive value of S indicates a strong positive relationship between X and Y; a predominance of discordant pairs resulting in a large negative value of S indicates a strong negative relationship between X and Y.

The denominator m is a normalizing coefficient such that the *Kendall's Tau* coefficient can assume values between −1 and +1: $-1 \leq Tau \leq 1$.

The interpretation of *Kendall's Tau* value is similar as for the other correlation coefficients: when the value is +1, then the two rankings are the same (the concordance between two variables is perfect); when the value is −1, the discordance is perfect (the ranking of one of variables is reverse to the other); and finally any other value between −1 and +1 is interpreted as a sign of the level of relationship, a positive relationship (*Kendall's Tau* > 0, both variables increase together), or a negative relationship (*Kendall's Tau* < 0, the rank of one variable increases, the other one decreases); the value 0 is an indication for non relationship.

If there are a large number of ties, then the dominator has to be corrected by $\sqrt{(m - n_x)(m - n_y)}$ where n_x is the number of ties involving X and n_y is the number of ties involving Y.

For inferential purposes, *Kendall's Tau* coefficient is used to test the hypothesis that X and Y are independent, $Tau = 0$, against one of the alternatives: $Tau \neq 0$, $Tau > 0$, $Tau < 0$. Critical values are tabulated, Daniel (1990), Abdi (2007). The problem of ties is considered also by Sillitto (1947), Burr (1960).

In large samples, the statistic

$$\frac{3 \times Kendall's\ Tau \times \sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

has approximately a normal distribution with mean 0 and standard deviation 1, and therefore can be used as a test statistic for testing the null hypothesis of zero correlation. It can be used also to calculate the confidence intervals (Noether 1967).

Kendall's *Tau* and Spearman's *Rho*

Kendall's *Tau* is equivalent to Spearman's *Rho*, with regard to the underlying assumptions. But they differ in their underlying logic and also computational formulas are quite different. The relationship between the two measures is given by

$$-1 \leq \{(3 \times \text{Kendall's } Tau) - (2 \times \text{Spearman's } Rho)\} \leq +1.$$

Their values are very similar in most cases, and when discrepancies occur, it is probably safer to interpret the lower value. More importantly, Kendall's *Tau* and Spearman's *Rho* imply different interpretations. Spearman's *Rho* is considered as the regular Pearson's correlation coefficient in terms of the proportion of variability accounted for, whereas Kendall's *Tau* represents a probability, i.e., the difference between the probabilities that the observed data are in the same order versus the probability that the observed data are not in the same order.

The distribution of Kendall's *Tau* has better statistical properties. In most of the situations, the interpretations of Kendall's *Tau* and Spearman's rank correlation coefficient are very similar and thus invariably lead to the same inferences. In fact neither statistics has any advantage in being easier to apply (since both are freely available in statistical packages) or easier to interpret. However Kendall's statistics structure is much simpler than that of the Spearman coefficient and has the advantage that it can be extended to explore the influence of a third variable on the relationship.

There are two different variations of Kendall's *Tau* that make adjustment for ties: *Tau b* and *Tau c*. These measures differ only as to how tied ranks are handled.

Kendall's *Tau-b*

Kendall's *Tau-b* is a nonparametric measure of correlation for ordinal or ranked variables that take ties into account. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values ranges from −1 to 1. The calculation formula for *Kendall's Tau-b* is given by the following:

$$Tau - b = \frac{C - D}{\sqrt{(C + D + X_0)(C + D + Y_0)}}$$

where X_0 is the number of pairs tied only on the X variable, Y_0 is the number of pairs tied only on the Y variable. When there are no ties, the values of Kendall's *Tau* and Kendall's *Tau b* are identical.

The *Kendall's Tau-b* has properties similar to the properties of the *Spearman Rho*. Because it does estimate a population parameter, many statisticians prefer the Kendall's *Tau-b* to the Spearman rank correlation coefficient.

Kendall's *Tau-c*

Kendall's *Tau-c*, is a variant of *Tau-b* used for situations of unequal-sized sets of ordered categories. It equals the excess of concordant over discordant pairs, multiplied by a term representing an adjustment for the size of the table. It is also called *Kendall–Stuart Tau-c* (or Stuart's Tau-c) and is calculated by formula

$$Tau - c = \frac{2m \times (C - D)}{n^2(m-1)}$$

where m is the smaller of the number of rows and columns, and n is the sample size.

Kendall's *Tau-b* and Kendall's *Tau-c* are superior to other measures of ordinal correlation when a test of significance is required.

About the Author

Dr. Llukan Puka is a Professor, Department of Mathematics, Faculty of Natural Science, University of Tirana, Albania. He was the Head of Statistics and Probability Section at the Faculty, Head of Mathematics Department too. During 1997–2007, he was the Dean of the Faculty. Professor Puka is an Elected Member of the ISI, associated to the IASC. He has authored and coauthored more than 30 papers and 15 textbooks, university level and high school level, mainly concerning probability, statistics, and random processes. He was the Head of National Council of Statistics (2001–2005). Professor Puka is the President of the Albanian Actuarial Association.

Cross References

- ►Copulas: Distribution Functions and Simulation
- ►Frailty Model
- ►Measures of Dependence
- ►Nonparametric Statistical Inference
- ►Sequential Ranks
- ►Statistics on Ranked Lists
- ►Tests of Independence
- ►Validity of Scales

References and Further Reading

Abdi H (2007) The Kendall rank correlation coefficient. In: Salkin NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks

Burr EJ (1960) The distribution of Kendall's score S for a pair of tied rankings. Biometrika 47:151–171

Daniel WW (1990) Applied nonparametric statistics, 2nd edn. PWS-KENT Publishing Company, Boston

Hollander M, Wolfe DA (1998) Nonparametric statistical methods, 2nd edn. Wiley, New York

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Noether GE (1967) Elements of nonparametric statistics. Wiley, New York

Sillitto GP (1947) The distribution of Kendall's coefficient of rank correlation in rankings containing ties. Biometrika 34:36–40

Khmaladze Transformation

Hira L. Koul, Eustace Swordson
Professor and Chair, President of the Indian Statistical Association
Michigan State University, East Lansing, MI, USA

Background

Consider the problem of testing the null hypothesis that a set of random variables $X_i, i = 1, \ldots, n$, is a random sample from a specified continuous distribution function (d.f.) F. Under the null hypothesis, the empirical d.f.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\}$$

must "agree" with F. One way to measure this agreement is to use omnibus test statistics from the empirical process (see ►Empirical Processes)

$$v_n(x) = \sqrt{n}(F_n(x) - F(x)).$$

The time transformed uniform empirical process

$$u_n(t) = v_n(x), \quad t = F(x)$$

is an empirical process based on random variables $U_i = F(X_i), i = 1, \ldots, n$, that are uniformly distributed on $[0,1]$ under the null hypothesis. Hence, although the construction of u_n depends on F, the null distribution of this process does not depend on F any more (Kolmogorov (1933), Doob (1949)). From this sprang a principle, universally accepted in goodness of fit testing theory, that one should choose tests of the above hypothesis based on statistics $A(v_n, F)$ which can be represented as statistics $B(u_n)$ just from u_n. Any such statistic, like, for example, weighted Cramér-von Mise statistics $\int v_n^2(x)\alpha(F(x))dF(x)$, or Kolmogorov-Smirnov statistics $\max_x |v_n(x)|/\alpha(F(x))$, will have a null distribution free from F, and hence this distribution can be calculated once and used for many different F – still a very desirable property in present times, in spite of great advantages in computational power. It is called the distribution free property of the test statistic.

However, as first clarified by Gikhman (1954) and Kac et al. (1955), this property is lost even asymptotically as soon as one is fitting a family of parametric d.f.'s. More precisely, suppose one is given a parametric family of d.f.'s F_θ, θ a k-dimensional Euclidean parameter, and one wishes to test the hypothesis that $X_i, i = 1, \ldots, n$, is a random sample from some F_θ. Denoting $\hat{\theta}_n$ a $n^{1/2}$-consistent estimator of θ, the relevant process here is the parametric empirical process

$$\hat{v}_n(x) = \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x)).$$

To describe the effect of estimation of θ on $\hat{v}_n$, let $\dot{F}_\theta(x) = \partial F_\theta(x)/\partial\theta$ and y^T denote the transpose of a k-vector y. Under simple regularity conditions,

$$\begin{aligned}\hat{v}_n(x) &= \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x))\\ &= v_n(x) - \dot{F}_\theta(x)^T\sqrt{n}(\hat{\theta}_n - \theta) + o_P(1).\end{aligned}$$

If additionally, for a k-vector of square integrable functions ψ,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \int \psi dv_n + o_P(1),$$

then $\hat{v}_n$ converges weakly to a mean zero Gaussian process $\hat{v}$, different from the weak limit of v_n, with a covariance function that depends on the unknown parameter θ via F_θ and ψ in a complicated fashion (Durbin (1973), Khmaladze (1979)). Critical values of any test based on this process are difficult to find even for large samples. Thus the goodness of fit testing theory was in danger of being fragmented into large number of particular cases and becoming computationally heavy and complex.

Khmaladze Transformation

To overcome this shortcoming, Khmaladze devised a transformation of $\hat{v}_n$ whose asymptotic null distribution under the parametric null hypothesis is distribution free while at the same time this transformed process stays in one-to-one correspondence with the process $\hat{v}_n$ without the loss of any "statistical information."

To describe this transformation, let f_θ denote density of F_θ and $\psi_\theta = \partial \log f_\theta/\partial\theta$ and let v denote the limit in distribution of empirical process v_n. Equip the process $\hat{v}$ with filtration $\mathcal{H} = \{\mathcal{H}_x, -\infty < x < \infty\}$, where each σ-field $\mathcal{H}_x = \sigma\{v(y), y \le x, \int \psi_\theta dv\}$ is generated not only by the "past" of v but also $\int \psi_\theta dv$, which contains a "little bit of a future" as well. This filtration is not an intrinsic part of the testing problem as it is usually formulated in statistics. Nevertheless, Khmaladze (1981) suggested to use it, because then it is natural to speak about martingale part $\{w, \mathcal{H}\}$ of the resulting semi-martingale $\{\hat{v}, \mathcal{H}\}$. Let $h_\theta^T(x) = (1, \psi_\theta(x))$ be "extended" score function and let $\Gamma_{x,\theta}$ be covariance matrix of $\int_x h_\theta dv$. Then this martingale part has the form

$$w(x) = v(x) - \int^x h_\theta(y)\Gamma_{y,\theta}^{-1}\int_y h_\theta dv\, dF_\theta(y). \quad (1)$$

The change of time $t = F_\theta(x)$ will transform it to a standard Brownian motion (see ►Brownian Motion and Diffusions) on $[0,1]$ – a convenient limiting process, with the distribution independent from F_θ. The substitution of $\hat{v}_n$ in (1) produces a version of empirical process w_n, which, basically, is the Khmaladze transform (KhT hence forth). It was shown to possess the following asymptotic properties: it will not change, regardless of which function ψ, or which estimator $\hat{\theta}_n$, was used in $\hat{v}_n$; it stays in one-to-one correspondence with $\hat{v}_n$, if $\hat{\theta}_n$ is the maximum likelihood estimator; and also the centering of empirical distribution function F_n in empirical process is unnecessary. Hence, the final form of KhT for parametric hypothesis is

$$w_{n,\theta}(x) = \sqrt{n}\left[F_n(x) - \int^x h_\theta(y)\Gamma_{y,\theta}^{-1}\int_y h_\theta dF_n\, dF_\theta(y)\right].$$

If the hypothesis is true, after time transformation $t = F_\theta(x)$, the processes $w_{n,\theta}$ and $w_{n,\hat{\theta}_n}$ converge weakly to standard Brownian motion. Consequently a class of tests based on time transformed $w_{n,\hat{\theta}_n}$ are asymptotically distribution free.

A slightly different point of view on $w_{n,\theta}$ is that its increment

$$dw_{n,\theta}(x) = \sqrt{n}\left[dF_n(x) - h_\theta(x)\Gamma_{x,\theta}^{-1}\int_x h_\theta dF_n\, dF_\theta(x)\right]$$

is (normalized) difference between $dF_n(x)$ and its linear regression on $F_n(x)$ and $\int_x h_\theta dF_n$.

If θ is known, i.e., if the hypothesis is simple, then $w_{n,\theta}$ reduces to what is called in the theory of empirical processes the basic martingale (see, e.g., Shorack and Wellner (1986)).

It is well known that the analog of Kolmogorov test is not distribution free when fitting a multivariate d.f. Khmaladze (1988, 1993) developed an analog of KhT in this case also, using the notion of so called scanning martingales.

Tsigroshvili (1998), and in some cases Khmaladze and Koul (2009), show that the KhT is well defined even if the matrix $\Gamma_{x,\theta}$ is not of full rank.

Some power properties of tests based on the $w_{n,\hat{\theta}_n}$ were investigated in a number of publications, including Janssen & Ünlü (2008), Koul and Sakhanenko (2005) and Nikitin (1995).

The specific form of $w_{n,\theta}$ and the practicality of its use for some particular parametric families was studied, e.g., in Koul and Sakhanenko (2005) and Haywood and Khmaladze (2007).

KhT for Counting Processes

If $N(t), t \geq 0$, is a point process (see ▸Point Processes) then Aalen (1978) used an appropriate filtration and the corresponding random intensity function $\lambda(t)$ to create the martingale

$$M(t) = N(t) - \int_0^t \lambda(s)ds.$$

This in turn gave rise to broad and successful theory, especially in survival analysis with randomly censored observations, as explained in the monograph by Andersen et al. (1993). However, if the $\lambda = \lambda_\theta$ depends on unspecified parameter, which needs to be estimated using N itself, then the process $\hat{M}(t)$ is not a martingale any more and suffers from the same problems as the process $\hat{v}_n$.

Again, by including the estimator $\hat{\theta}$ in the filtration used, the KhT for $\hat{M}(t)$ was constructed in Maglaperidze et al. (1998), Nikabadze and Stute (1997), and later in O'Quigley (2003), Sun et al. (2001) and Scheike and Martinussen (2004).

KhT in Regression

The transformation was taken into new direction of the quantile regression problems in Koenker and Xiao (2002), where some additional problems were resolved. The practicality of the approach was demonstrated by the software, created by Roger Koenker and his colleagues. Recent extension to the case of autoregression is presented in discussion paper Koenker and Xiao (2006).

In the classical mean regression set up with covariate X and response Y, $Y = \mu(X) + \epsilon$, where error ϵ is independent of X, $E\epsilon = 0$, and $\mu(x) = E(Y|X = x)$. Let (X_i, Y_i), $i = 1, \cdots, n$, be a random sample from this model.

Here the two testing problems are of interest. One is the goodness-of-fit of an error d.f. and the second is the problem of lack-of-fit of a parametric regression function $m_\theta(x)$. In parametric regression model, tests for the first problem are based on the residual empirical process $\hat{v}_n(x)$ of the residuals $\hat{\epsilon}_i = Y_i - m_{\hat{\theta}_n}(X_i)$, $i = 1, \cdots, n$, where $\hat{\theta}_n$ is a $n^{1/2}$-consistent estimator of θ. Khmaladze and Koul (2004) develops the KhT of $\hat{v}_n$. Similar results were obtained for nonparametric regression models in Khmaladze and Koul (2009). It is shown, somewhat unexpectedly, that in nonparametric regression models, KhT not only leads to an asymptotically distribution free process, but also tests based on it have larger power than the tests based on $\hat{v}_n$ with non-parametric residuals $Y_i - \hat{m}_n(X_i)$.

Tests of lack-of-fit are typically based on the partial sum processes of the form

$$\sum_{i=1}^{n} g(\hat{\epsilon}_i) I\{X_i \leq x\},$$

for some known function g. However, again their limiting distribution depend on the form of the regression function, on the estimator $\hat{\theta}_n$ used and on the particular value of the parameter. Starting with Stute et al. (1998) this tradition was changed and KhT was introduced for these partial sum processes, which again, led to the process converging to standard Brownian motion. Khmaladze and Koul (2004) studied the analog of KhT for partial sum process when design variable is multi-dimensional.

Extension to some time series models are discussed in Koul and Stute (1999), Bai (2003) and Koul (2006). Koul and Song (2008, 2009, 2010), Dette and Hetzler (2008, 2009) illustrate use of KhT in some other problems in the context of interval censored data, Berkson measurement error regression models and fitting a parametric model to the conditional variance function.

About the Author

Dr. Hira Lal Koul is Professor and Chair, Department of Statistics and Probability, Michigan State University. He was President of the International Indian Statistical Association (2005–2006). He was awarded a Humboldt Research Award for Senior Scientists (1995). He is a Fellow of the American Statistical Association, Institute of Mathematical Statistics, and Elected member of the International Statistical Institute. He is Co-Editor in Chief of *Statistics and Probability Letters*, Associate Editor for *Applicable Analysis and Discrete Mathematics*, and Co-Editor of *The J. Mathematical Sciences* he has (co-)authored about 110 papers, and several monographs and books. He is joint editor of the text *Frontiers in Statistics* (with Jianqing Fan, dedicated to Prof. Peter J Bickel in honor of his 65th birthday, Imperial College Press, London, UK, 2006).

Cross References

▸Empirical Processes
▸Martingales
▸Point Processes

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic porcesses. Ann Math Stat 23:193–212

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Bai J (2003) Testing parametric conditional distributions of dynamic models. Rev Econ Stat 85:531–549

Dette H, Hetzler B (2008) A martingale-transform goodness-of-fit test for the form of the conditional variance. http://arXiv.org/abs/0809.4914?context=stat

Dette H, Hetzler B (2009) Khmaladze transformation of integrated variance processes with applications to goodness-of-fit testing. Math Meth Stat 18:97–116

Doob JL (1949) Heuristic approach to the Kolmogorov-Smirnov theorems. Ann Math Stat 20:393–403

Durbin J (1973) Weak convergence of the sample distribution function when parameters are estimated. Ann Statist 1:279–290

Gikhman II (1954) On the theory of ω^2 test. Math Zb Kiev State Univ 5:51–59

Haywood J, Khmaladze EV (2007) On distribution-free goodness-of-fit testing of exponentiality. J Econometrics 143:5–18

Janssen A, Ünlü H (2008) Regions of alternatives with high and low power for goodness-of-fit tests. J Stat Plan Infer 138:2526–2543

Kac M, Kiefer J, Wolfowitz J (1955) On tests of normality and other tests of goodness of fit based on distance methods. Ann Math Stat 26:189–211

Khmaladze EV (1979) The use of $\omega2$ tests for testing parametric hypotheses. Theor Probab Appl 24(2):283–301

Khmaladze EV (1981) Martingale approach in the theory of goodness-of-fit tests. Theor Probab Appl 26:240–257

Khmaladze EV (1988) An innovation approach to goodness-of-fit tests in R^m. Ann Stat 16:1503–1516

Khmaladze EV (1993) Goodness of fit problem and scanning innovation martingales. Ann Stat 21:798–829

Khmaladze EV, Koul HL (2004) Martingale transforms goodness-of-fit tests in regression models. Ann Stat 32:995–1034

Khmaladze EV, Koul HL (2009) Goodness of fit problem for errors in non-parametric regression: distribution free approach. Ann Stat 37:3165–3185

Koenker R, Xiao Zh (2002) Inference on the quantile regression process. Econometrica 70:1583–1612

Koenker R, Xiao Zh (2006) Quantile autoregression. J Am Stat Assoc 101:980–990

Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari 4: 83–91

Koul HL, Stute W (1999) Nonparametric model checks for time series. Ann Stat 27:204–236

Koul HL, Sakhanenko L (2005) Goodness-of-fit testing in regression: A finite sample comparison of bootstrap methodology and Khmaladze transformation. Stat Probabil Lett 74: 290–302

Koul HL (2006) Model diagnostics via martingale transforms: a brief review. In: Frontiers in statistics, Imperial College Press, London, pp 183–206

Koul HL, Yi T (2006) Goodness-of-fit testing in interval censoring case 1. Stat Probabil Lett 76(7):709–718

Koul HL, Song W (2008) Regression model checking with Berkson measurement errors. J Stat Plan Infer 138(6):1615–1628

Koul HL, Song W (2009) Model checking in partial linear regression models with berkson measurement errors. Satistica Sinica

Koul HL, Song W (2010) Conditional variance model checking. J Stat Plan Infer 140(4):1056–1072

Maglaperidze NO, Tsigroshvili ZP, van Pul M (1998) Goodness-of-fit tests for parametric hypotheses on the distribution of point processes. Math Meth Stat 7:60–77

Nikabadze A, Stute W (1997) Model checks under random censorship. Stat Probabil Lett 32:249–259

Nikitin Ya (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge, pp xvi+274

O'Quigley J (2003) Khmaladze-type graphical evaluation of the proportional hazards assumption. Biometrika 90(3): 577–584

Scheike TH, Martinussen T (2004) On estimation and tests of time-varying effects in the proportional hazards model. Scand J Stat 31:51–62

Shorack GR, Wellner JA (1986) Empirical processes with application to statistics. Wiley, New York

Stute W, Thies S, Zhu Li-Xing (1998) Model checks for regression: an innovation process approach. Ann Stat 26:1916–1934

Sun Y, Tiwari RC, Zalkikar JN (2001) Goodness of fit tests for multivariate counting process models with applications. Scand J Stat 28:241–256

Tsigroshvili Z (1998) Some notes on goodness-of-fit tests and innovation martingales (English. English, Georgian summary). Proc A Razmadze Math Inst 117:89–102

Kolmogorov-Smirnov Test

Raul H. C. Lopes
Research Fellow
Brunel University, Uxbridge, UK
Professor of Computer Science at UFES, Vitoria
Brazil

Applications of Statistics are frequently concerned with the question of whether two sets of data come from the same distribution function, or, alternatively, of whether a probabilistic model is adequate for a data set. As an example, someone might be interested in evaluating the quality of a computer random numbers generator, by testing if the sample is uniformly distributed. A test like that is generally called a goodness-of-fit test. Examples of it are the χ^2 test and the Kolmogorov-Smirnov test.

Generally given a sample $X = x_0, x_1, \ldots, x_{n-1}$ and a probability distrbution function $P(x)$ the target would be to test the Null Hypothesis H_0 that P is the sample's distribution function. When testing with two data sets the Null Hypothesis H_0 states that they both have the same distribution functions.

The choice of a statistical test must take into account at least two factors: (1) whether the data is continuous or discrete, (2) and if the comparison to be performed uses two data sets or a one set against a fitting probability model. Testing that a set of computer generated pseudo-random real numbers follows a uniformly distributed model is an example of testing a continuous data set against a probabilistic model, while comparing the amount of Vitamin C in two different brands of orange juice would fit a comparison of two continuous data sets.

The χ^2 test was designed to test discrete data sets against a probability model. However, it could be applied in the test of the computer random numbers generator by discretising the sample. Given the set $X = x_0, x_1, \ldots, x_{n-1}$ of generated numbers, a set of k intervals (bins)

$$(-\infty, z_1), (z_1, z_2), \ldots, (z_{k-1}, \infty)$$

could be used to define a discrete function

$$X_j = i, \text{ when } x_j \in (z_{i-1}, z_i).$$

Kolmogorov (1933) and Smirnov (1948) proved a result, also Schmid (1958), that is the basis for a much more efficient goodness-of-fit test when continuous data is involved. The test starts with the definition of a function $F_{X,n}(x)$ that gives the fraction of points $x_i, i \in (0, \ldots, n-1)$, in a sample X that are below x as follows (E.W. Dijkstra's uniform notation for quantifiers is used, with $\# i : P(x_i)$ denoting the number of elements in the set satisfying the property $P(x_i)$, for all possible i.):

$$F_{X,n}(x) = \frac{\# i : x_i \leq x}{n}$$

Assuming that another sample $Y = y_0, y_1, \ldots, y_{m-1}$ is given, then its function can be defined:

$$F_{Y,m}(y) = \frac{\# i : y_i \leq y}{m}$$

And any statistic could be used to measure the difference between X and Y, by measuring the difference between $F_{X,n}(x)$ and $F_{Y,m}(x)$. Even the area between the curves defined by these functions could be used. The Kolmogorov-Smirnov distance, is defined as the maximum absolute value of the difference between $F_{X,n}(x)$ and $F_{Y,m}(x)$ for all possible values of x:

$$D = max\, x : -\infty < x < \infty : F_{X,n}(x) - F_{Y,m}(x)$$

In a test trying to fit one sample with a probabilistic model defined by the function $P(x)$, the distance, also called Kolmogorov-Smirnov statistic, would be defined as

$$D = max\, x : -\infty < x < \infty : F_{X,n}(x) - P(x)$$

The distribution of the Kolmogorov-Smirnov statistic in the case of a Null Hypothesis test can be computed, giving a significance level for the observed value. For that purpose, let $D^\star$ be the following function of the observed value:

$$D^\star(d) = \left[\sqrt{n_e} + 0.12 + 0.11/\sqrt{n_e}\right] d$$

In the definition of $D^\star$, the quantity n_e is defined as follows:

- n_e is the number of points in the sample, when doing a one-sample test.
- $n_e = \frac{n*m}{n+m}$, in the case of a two-sample test, with n and m being the sizes of the samples.

The significance level can then be computed using the function Q below (Stephens 1970):

$$Q(d) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

Given a d, computed by the Kolmogorov-Smirnov distance, the significance level of d, which comes to be the probability that the null hypotheses (that the two distributions are the same) is invalid, is given by

$$\text{Probability}(D > d) = Q(D^\star(d))$$

The Kolmogorov-Smirnov test offers several advantages over the χ^2 test:

- It can be applied to continuous data.
- The distribution of its statistic is invariant under re-parametrisation and it can be easily implemented by computers.
- It can be extended to multivariate data.

Several statistics packages implement the Kolmogorov-Smirnov test. The package **R** (Crawley 2007), freely available (Software and documentation from http://www.r-project.org) for most operating systems, offers a Kolmogorov-Smirnov test in the function *ks.test*.

Adapting goodness-of-fit tests to multivariate data is considered a challenge. In particular, tests based on binning suffer from what has been described as the "curse of multi-dimensionality": the multi-dimensional space is essentially empty and binning tests tend to be ineffective even with large data sets.

Peacock in (Peacock 1983) introduced an extension of the Kolmogorov-Smirnov test to multivariate data. The idea consists in taking into account the distribution function of the two samples in all possible orderings, $2^d - 1$ orderings when d dimensional data is being considered. Given n points, in a two-dimensional space, Peacock proposed to compute the distribution functions in the $4n^2$ quadrants of the plane defined by all pairs (x_i, y_i), x_i and y_i being coordinates of all points of two given samples. This gives an algorithm of $\Omega(n^3)$ complexity. Fasano e Franceschini introduced in (Fasano and Franceschini 1987) an approximation of the Peacock's test that computes the statistic over all quadrants centred in each point of the given samples. Their test can be computed in time $\Omega(n^2)$. Lopes et alii introduced an algorithm (Available, under GPL license, from http://www.inf.ufes.br/ raul/cern.2dks.tar.bz2) based on range-counting trees that computes this last statistic in $O(n \lg n)$, which is a lower-bound for the test (Lopes et al. 2008).

Cross References

►Chi-Square Tests
►Cramér-Von Mises Statistics for Discrete Distributions
►Normality Tests
►Normality Tests: Power Comparison
►Parametric and Nonparametric Reliability Analysis
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Crawley MJ (2007) The R book. Wiley, Chichester

Fasano G, Franceschini A (1987) A multi-dimensional version of the Kolmogorov-Smirnov test. Monthly Notices of the Royal Astronomy Society 225:155–170

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell' Instituto Italiano degli Attuari 4:83–91

Lopes RHC, Hobson PR, Reid I (2008) Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test. Conference Series, Journal of Physics, p 119

Peacock JA (1983) Two-dimensional goodness-of-fit in Astronomy. Monthly Notices of the Royal Astronomy Society 202: 615–627

Schmid P (1958) On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. Ann Math Stat 29(4):1011–1027

Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann Math Stat 19(2):279–281

Stephens MA (1970) Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. J R Stat Soc 32:115–122

Kullback-Leibler Divergence

James M. Joyce
Chair and Professor of Philosophy and of Statistics
University of Michigan, Ann Arbor, MI, USA

Kullback-Leibler divergence (Kullback and Leibler 1951) is an information-based measure of disparity among probability distributions. Given distributions P and Q defined over X, with Q absolutely continuous with respect to P, the *Kullback-Leibler divergence* of Q from P is the P-expectation of $-\log_2\{P/Q\}$. So, $D_{KL}(P,Q) = -\int_X \log_2(Q(x)/P(x))dP$. This quantity can be seen as the difference between the *cross-entropy for Q on P*, $H(P,Q) = -\int_X \log_2(Q(x))dP$, and the *self-entropy* (Shannon 1948) of P, $H(P) = H(P,P) = -\int_X \log_2(P(x))dP$. Since $H(P,Q)$ is the P-expectation of the number of bits of information, beyond those encoded in Q, that are needed to identify points in X, $D_{KL}(P,Q) = H(P) - H(P,Q)$ is the expected difference, from the perspective of P, between the information encoded in P and the information encoded in Q.

D_{KL} has a number of features that make it plausible as a measure of probabilistic divergence. Here are some of its key properties:

Premetric. $D_{KL}(P,Q) \geq 0$, with identity if and only if $P = Q$ a.e. with respect to P.

Convexity. $D_{KL}(P,Q)$ is convex in both P and Q.

Chain Rule. Given joint distributions $P(x,y)$ and $Q(x,y)$, define the KL-divergence conditional on x as $D_{KL}(P(y|x),Q(y|x)) = \int_X D_{KL}(P(y|x),Q(y|x))dP_x$ where P_x is P's x-marginal. Then,

$$D_{KL}(P(x,y),Q(x,y)) = D_{KL}(P_x,Q_x) + D_{KL}(P(y|x),Q(y|x)).$$

Independence. When X and Y are independent in both P and Q the Chain Rule assumes the simple form $D_{KL}(P(x,y),Q(x,y)) = D_{KL}(P_x,Q_x) + D_{KL}(P_y,Q_y)$, which reflects the well-known idea that independent information is additive.

It should be emphasized that KL-divergence is not a genuine metric: it is not symmetric and fails the triangle inequality. Thus, talk of Kullback-Leibler "distance" is misleading. While one can create a symmetric divergence measure by setting $D^*_{KL}(P,Q) = \frac{1}{2}D_{KL}(P,Q) + \frac{1}{2} D_{KL}(Q,P)$, this still fails the triangle inequality.

There is a close relationship between KL-divergence and a number of other statistical concepts. Consider, for example, *mutual information*. Given a joint distribution

$P(x,y)$ on $X \times Y$ with marginals P_X and P_Y, the mutual information of X and Y with respect to P is defined as $I_P(X,Y) = -\int_{X\times Y} \log_2(P(x,y)/[P_X(x)\cdot P_Y(y)])dP$. If we let $P_\perp(x,y) = P_X(x)\cdot P_Y(y)$ be the factorization of P, then $I_P(X,Y) = D(P,P_\perp)$. Thus, according to KL-divergence, mutual information measures the dissimilarity of a joint distribution from its factorization.

There is also a connection between KL-divergence and maximum likelihood estimation. Let $l_x(\theta) = p(x|\theta)$ be a likelihood function with parameter $\theta \in \Theta$, and imagine that enough data has been collected to make a certain empirical distribution $f(x)$ seem reasonable. In MLE one often hopes to find an estimate for θ that maximizes *expected log-likelihood* relative to one's data, i.e., we seek $\hat{\theta} = \text{argmax}_\theta E_f[\log_2(p(x|\theta)]$. To find this quantity it suffices to minimize the KL-divergence between $f(x)$ and $p(x|\hat{\theta})$ since

$$\begin{aligned}
&\text{argmin}_\theta\, D_{KL}(f, p(\cdot|\hat{\theta}))\\
&\qquad = \text{argmin}_\theta - \int_X f(x)\cdot\log_2(p(x|\hat{\theta})/f(x))dx\\
&\qquad = \text{argmin}_\theta[H(f,f) - H(f,p(\cdot|\hat{\theta}))]\\
&\qquad = \text{argmax}_\theta H(f,p(\cdot|\hat{\theta}))\\
&\qquad = \text{argmax}_\theta E_f[\log_2(p(x|\theta))].
\end{aligned}$$

In short, MLE minimizes Kullback-Leibler divergence from the empirical distribution.

Kullback-Leibler also plays a role in ►model selection. Indeed, Akaike (1973) uses D_{KL} as the basis for his "information criterion" (AIC). Here, we imagine an unknown true distribution $P(x)$ over a sample space X, and a set Π_θ of models each element of which specifies a parameterized set of distributions $\pi(x|\theta)$ over X. The models in Π_θ are meant to approximate P, and the aim is to find the best approximation in light of data drawn from P. For each π and θ, $D_{KL}(P,\pi(x|\theta))$ measures the information lost when $\pi(x|\theta)$ is used to approximate P. If θ were known, one could minimize information loss by choosing π to minimize $D_{KL}(P,\pi(x|\theta))$. But, since θ is unknown one must estimate. For each body of data y and each π, let $\hat{\theta}_y$ be the MLE estimate for θ given y, and consider $D_{KL}(P,\pi(x|\hat{\theta}_y))$ as a random variable of y. Akaike maintained that one should choose the model that minimizes the expected value of this quantity, so that one chooses π to minimize $E_y[D_{KL}(P,\pi(x|\hat{\theta}_y))] = E_y[H(P,P) - H(P,\pi(\cdot|\hat{\theta}_y))]$. This is equivalent to maximizing $E_yE_x[\log_2(\pi(x|\hat{\theta}_y))]$. Akaike proved that $2k - \log_2(l_x(\hat{\theta}))$ is an unbiased estimate of this quantity for large samples, where $\hat{\theta}$ is the MLE estimate of θ and k is the number of estimated parameters. In this way, some have claimed, the policy of minimizing KL-divergence leads one to value simplicity in models since the "$2k$" term functions as a kind of penalty for complexity. (see Sober 2002).

KL-divergence also figures prominently in Bayesian approaches experimental design, where it is treated as a utility function. The objective in such work is to design experiments that maximize KL-divergence between the prior and posterior. The results of such experiments are interpreted as having a high degree of informational content. Lindley (1956) and De Groot (1962) are essential references here.

Bayesians have also appealed to KL-divergence to provide a rationale for Bayesian conditioning and related belief update rules, e.g., the probability kinematics of Jeffrey (1965). For example, Diaconis and Zabell (1982) show that the posterior probabilities prescribed by Bayesian conditioning or by probability kinematics minimize KL-divergence from the perspective of the prior. Thus, in the sense of information divergence captured by D_{KL}, these forms of updating introduce the least amount of new information consistent with the data received.

About the Author

For biography *see* the entry the ►St. Petersburg paradox.

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Chernoff Bound
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Estimation
►Information Theory and Statistics
►Measurement of Uncertainty
►Radon–Nikodým Theorem
►Statistical Evidence

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Proceedings of the international symposium on information theory. Budapest, Akademiai Kiado

De Groot M (1962) Uncertainty, information, and sequential experiments. Ann Math Stat 33:404–419

Diaconis P, Zabell S (1982) Updating subjective probability. J Am Stat Assoc 77:822–830

Jeffrey R (1965) The logic of decision. McGraw-Hill, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Lindley DV (1956) On the measure of information provided by an experiment. Ann Stat 27:985–1005

Shannon CE (1948) A mathematical theory of communication. AT&T Tech J 27(379–423):623–656

Sober E (2002) Instrumentalism, parsimony, and the Akaike framework. Philos Sci 69:S112–S123

Kurtosis: An Overview

Edith Seier
Professor
East Tennessee State University, Johnson City, TN, USA

Pearson (1905) defined $\beta_2 = m_4/m_2^2$ (where m_i is the ith moment with respect to the mean) to compare other distributions to the normal distribution, for which $\beta_2 = 3$. He called $\eta = \beta_2 - 3$ the "degree of kurtosis" and mentioned that it "measures whether the frequency towards the mean is emphasized more or less than that required by the Gaussian law." In Greek, *kurtos* means convex, and *kurtosis* had been previously used to denote curvature both in mathematics and medicine. Pearson's development of the idea of kurtosis during the years previous to 1905 is examined by Fiori and Zenga (2009). "Coefficient of kurtosis" is the name usually given to β_2.

A sample estimator of β_2 is $b_2 = (\sum(x - \bar{x})^4/n)/s^4$. Statistical software frequently include an adjusted version of the estimator of η:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\left[\sum(x-\bar{x})^4\right]}{s^4} - \frac{3(n-1)(n-1)}{(n-2)(n-3)}.$$

The adjustment reduces the bias, at least in the case of nearly normal distributions. Byers (2000) proved that $b_2 \leq n - 2 + 1/(n-1)$. Simulation results indicate that when β_2 is large for the population of origin, b_2 will be small on average if the sample size is small.

Currently the word kurtosis is understood in a broader sense, not limited to β_2. Balanda and MacGillivray (1988) conclude that kurtosis is best defined as "the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails." which can be formalized in many ways. Kurtosis is associated to both, the center and the tails of a distribution. Kurtosis is invariant under linear transformations or change of units of the variable. High kurtosis is linked to high concentration of mass in the center and/or the tails of the distribution. Heavy tails is a topic of interest in the analysis of financial data.

Several kurtosis measures have been defined. L-kurtosis (Hosking 1992) is popular in the field of hydrology. There are other measures defined in terms of distances between quantiles, ratios of spread measures, comparisons of sum of distances to the median, and expected values of functions of the standardized variable other than the fourth power that corresponds to β_2.

Ruppert (1987) proposed the use of the influence function to analyze kurtosis measures and points out that even those defined with the intention of measuring peakedness or tail weight alone, end up measuring both. There are measures that are more sensitive to the tails of the distribution than others: β_2 gives high importance to the tails because it is defined in terms of the fourth power of the deviations from the mean. For example, the value of β_2 is 1.8 for the uniform distribution and 3.53, 4.51, 36.2 and 82.1 for the $SU(0,\delta)$ distribution with $\delta = 3, 2, 1, 0.9$ respectively. For the same distributions, the values of L-kurtosis are 0,0.143,0.168, 0.293 and 0.329. The upper bound for L-kurtosis is 1, while β_2 is unbounded. The estimator b_2 is sensitive to ▸outliers; one single outlier can dramatically change its value.

Another approach to the study of kurtosis is the comparison of cumulative distribution functions. Van Zwet (1964) defined the convexity criterion $(<_S)$: two symmetric distributions with cumulative distribution functions F and G are ordered and $F <_S G$ if $G^{-1}(F(x))$ is convex to the right of the common point of symmetry. If $F <_S G$, the value of β_2 for F is not larger than its value for G. The following distributions are ordered according to the convexity criterion:

$$U\text{-shaped} <_S \text{Uniform} <_S \text{Normal} <_S \text{Logistic} <_S \text{Laplace}.$$

Some families of distributions are ordered according to the convexity criterion, with the order associated (either directly or inversely) to the value of their parameter. Among those families are $beta(\alpha,\alpha)$, $Tukey(\lambda)$, Johnson's $SU(0,\delta)$, and the symmetric two-sided power family $TSP(\alpha)$. Balanda and MacGillivray (1990) defined the spread-spread functions to compare non-necessarily symmetric distributions. Additional ordering criteria have been defined. Any new measure of kurtosis that is defined needs to order distributions in agreement with some ordering based on distribution functions. The numerical value of a kurtosis measure can be obtained for most distributions but not all distributions are ordered according to a CDF based ordering criterion. For example, the *Laplace* and *t-Student*$_{(6)}$ distributions have known values for β_2 (6 and 9 respectively). However, they are not $<_S$ ordered because $G^{-1}(F(x))$ is neither convex, nor concave for $x > 0$. In particular, not all the distributions are ordered with respect to the normal distribution according to the convexity criterion; but uniform $<_S$ unimodal distributions.

There are several ways of measuring kurtosis, there is also more than one way of thinking about peak and tails. One simple way of visualizing peak and tails in a unimodal probability distribution is to superimpose, on $f(x)$, a uniform density function with the same median and variance (Kotz and Seier 2008).

High kurtosis affects the behavior of inferential tools. Van Zwet (1964) proved that, when working with symmetric distributions, the median is more efficient than the mean as estimator of the center when the distribution has very high kurtosis. The variance of the sample variance is related to β_2. Simulations indicate that the power of some tests for the equality of variances diminishes (for small samples) when the distribution of the variable has high kurtosis.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Jarque-Bera Test
►Normality Tests
►Normality Tests: Power Comparison
►Statistical Distributions: An Overview

References and Further Reading

Balanda KP, MacGillivray HL (1988) Kurtosis: a critical review. Am Stat 42:111–119

Balanda KP, MacGillivray HL (1990) Kurtosis and spread. Can J Stat 18:17–30

Byers RH (2000) On the maximum of the standardized fourth moment. InterStat January #2 http://interstat.statjournals.net/YEAR/2000/articles/0001002.pdf

Fiori A, Zenga M (2009) Karl Pearson and the origin of Kurtosis. Int Stat Rev 77:40–50

Hosking JRM (1992) Moments or *L* moments? An example comparing two measures of distributional shape. Am Stat 46:186–199

Kotz S, Seier E (2008) Visualizing peak and tails to introduce kurtosis. Am Stat 62:348–352

Pearson K (1905) Skew variation, a rejoinder. Biometrika 4:169–212

Ruppert D (1987) What is Kurtosis? An influence function approach. Am Stat 41:1–5

Van Zwet WR (1964) Convex transformations of random variables. Mathematics Centre, Tract 7. Mathematisch Centrum Amsterdam, Netherlands