THE USE OF THE STANDARD SIGMOID AND THE LOGISTIC CURVES IN PAIRWISE COMPARISONS

By: Arpad E. Elo and Joseph Talacko

Marquette University
Milwaukee, Wisconsin 53233

ABSTRACT

Two practical rating systems for use in pairwise comparisons are described. For items measured on an interval scale the normal cumulative proportion curve is shown to apply and for items measured on a ratio scale the logistic curve is shown to apply. For the latter the merits of base $\sqrt{10}$ are discussed.

INTRODUCTION

The problem of pairwise comparison of the properties of physical and biological systems has recently received considerable attention in the literature of statistical theory. (Refs. 1, 2, 3, 4, 5, 6, 7) In general the problem here is to compare the varying properties of systems and to rank these systems on the basis of the properties. The ranking furthermore is not merely to be ordinal but one which would be expressed on an interval scale of measurements. The purpose of this paper is to present two of the principal forms of ranking or rating systems which may serve useful purposes in biological assay; to compare these systems; and to define their areas of application.

The mathematical problem met in pairwise comparisons is essentially the same as found in the ranking of individuals competing in a tournament. In the investigations cited the approach to the solution has been essentially along these lines and the present paper will therefore also use the same frame of reference. Thus the systems will be designated as the players and the properties compared as the strength or rating of the players. A rating itself may be expressed on an interval scale whereon the numbers serve to reflect differences or distances among the items or upon a ratio scale whereon the numbers serve to reflect ratios among the items. (Ref. 8) The type of scale used will determine the functional relation between the probability of one player winning from another or his percentage expectancy and their difference in rating.

RATINGS FOR MEASUREMENTS ON AN INTERVAL SCALE

This problem was investigated by one of the writers (Ref. 9) in connection with a study of performance and the aging factor. The random variable measured was the performance of chess players in tournaments and it was assumed that these were normally distributed around a mean μ_R and with a standard deviation σ_R. Then if the mean ratings of two players differ by D the difference in performance (designated by x) will also be normally distributed around D and with a standard deviation σ_D, which is equal to $\sqrt{2}\,\sigma_R$. (Fig. 1) If the individual performances are not truly normally distributed it nevertheless follows from the central limit theorem that the distribution of the differences will tend to be normal.

Referring now to Fig. 1, it is seen that the integral from $-\infty$ to 0, that is the shaded area under the curve, represents the probability of the difference going negative or the lower player winning and the integral from 0 to $+\infty$, or the unshaded area, the probability of the higher player winning. When the latter is plotted as a function of D there results the normal cumulative proportion curve or standard sigmoid with a standard deviation σ_D. This is shown in a useful form on normal probability paper in Fig. 2. This curve thus forms the basis of the normal rating scale and the rating of a player may then be expressed as:

$$\mu_R = R_c + D(p) \tag{1}$$

where: R_c is the competition rating i.e. that of the opponent or the average of the opponents, and $D(p)$ is the difference in rating based upon the percentage score p achieved in the tournament. Of course, at the outset of the tournament none of the ratings are known so that it is necessary to assign an arbitrary average rating to the group of players say R_c'. On this basis a tentative rating R_1' is determined for each player. Then these tentative ratings are used to determine a corrected competition rating for each player and this process of successive approximations is continued until a set of self consistent ratings is obtained for the entire group. As indicated above such ratings reflect only differences among the members. There is no absolute zero on the rating scale and as a general practice the average rating is given a number high enough so that no rating will go negative. In the field of biological assay an outstanding example of this type of rating scale is in probit analysis. (Ref. 10)

RATINGS FOR MEASUREMENTS ON A RATIO SCALE

This problem has been investigated by both Good and Buhlman & Huber. Good proceeded from the assumption that the odds of player #1 beating

#2 times the odds of #2 beating #3 is equal to the odds of #1 beating #3. Buhlman & Huber assumed that the logarithm of the probability ratio, namely the odds, is equal to the difference in the player parameter θ. that is, the rating difference, and then proceeded by the use of matrix algebra to derive the percentage expectancy function. Actually, if no draws between the players occur or if draws are treated as $\frac{1}{2}$ wins and $\frac{1}{2}$ losses the two assumptions are equivalent and lead immediately to the percentage expectancy function:

$$\text{If; } \log \frac{P_{12}}{P_{21}} = \theta_1 - \theta_2 = \frac{D}{\alpha} \qquad (2)$$

and since $P_{21} = (1 - P_{12})$ then it follows dropping subscripts that:

$$P(D) = \frac{1}{1 + e^{-D/\alpha}} \qquad (3)$$

where α is the natural scale interval on the rating scale, namely, that interval for which the odds or the probability ratio is just equal to e. This is the familiar logistic curve generally associated with growth processes. (Ref. 11) Its use as a distribution function is perhaps not as widely known although such uses have been recommended since 1929. (Ref. 12, 13) The function is shown in Fig. 3 drawn on logistic probability paper and serves as the basis for what will be designated as the logistic rating scale. From equation (3) the expression for the rating of a player in a tournament may be readily obtained as:

$$\mu_R = R_c + \log \frac{P}{1-P} \qquad (4)$$

where p is the actual percentage score achieved in the tournament.

At this point it is perhaps appropriate to inquire what is the nature of the distribution of the parent variable, that is, the player performance for which (3) is the cumulative proportion function. Differentiating (3) there is obtained:

$$\frac{dP}{dD} = \frac{e^{-D/\alpha}}{\alpha(1 + e^{-D/\alpha})^2} \qquad (5)$$

which is the frequency distribution function of the differences in performance. This is a bell shaped curve similar to the normal curve but with tails somewhat higher. The function is the Verhulst distribution and is one of the family of curves known as Perks' distributions. This family of distributions has been extensively investigated by the second author and his students. (Refs. 14, 15, 16, 17) It has been shown that if the distribution of the differences is represented by the Verhulst function then the distribution of the parent variable is also a Verhulst function with the natural parameter α_R such that $\alpha = \sqrt{2}\,\alpha_R$. Furthermore α_R has been shown to be related to the standard deviation of this distribution thus: $\sigma_R = (\pi/\sqrt{3})\alpha_R$. This is obviously a case for which the central limit theorem does not apply. The range of applicability therefore is to the statistics of a random variable involving a time series, that is, to dynamical statistics.

APPLICATIONS OF THE LOGISTIC TO DIFFERENT BASES

Returning to the basic assumption as stated in equation (2) it is evident that the logarithm of the odds may be taken to any arbitrary base B and then (2) may be written as:

$$\log_B \frac{P_{12}}{P_{21}} = \frac{D}{\beta} \qquad (6)$$

and the percentage expectancy function becomes:

$$P(D) = \frac{1}{(1 + B^{-D/\beta})} \qquad (7)$$

where now β is the natural scale interval for the new scale, namely, that interval for which the probability ratio is just equal to B. (If $B = e$, then $\beta = \alpha$.

If the odds are plotted on semi-log paper against rating differences in terms of β there is obtained a straight line and if different bases are used a family of straight lines as shown in Fig. 4. The line to the base 10 is recognized as the familiar accoustical curve with β = 1 bel or ten decibels. The line to the base $\sqrt[5]{100}$, which turns out to be about 2.5, is the astronomical rating curve with $\beta = 1$ stellar magnitude.

The line to the base $\sqrt{10}$, or 3.162, is of especial interest because of a fortuitous numerical relation. If we return to the normal rating curve it is seen that for a rating difference of σ_R of the normal curve the respective percentage expectancies are .7603 and .2397 and the odds 3.170. Thus using $\sqrt{10}$ as the base with its form of the percentage expectancy function:

$$P(D) = \frac{1}{(1 + 10^{-D/2\beta})} \qquad (8)$$

a very simple comparison is possible between the normal and logistic rating systems. This is shown in the short table below. It is seen that there is little difference between the logistic and normal distributions in the most important interval i.e. up to differences of two class intervals, and only beyond these do the odds reflect the differences in the two distributions.

CONCLUSION

In conclusion it should be pointed out that while the solutions of the problem of pairwise comparison have been presented in the framework of tournament rankings, the results can be applied to any other type of comparison. In addition to the two physical examples already cited other possible applications of these methods exist in both biological assay and in psychometrics. (Refs. 18, 19) In applying the rating methods, however, due consideration should be given to the type of measurements and random variables involved. For the normal rating scale is appropriate to interval measurements and to statical statistics, whereas the logistic is appropriate to ratio measurements and dynamical statistics.

COMPARISON OF THE NORMAL DISTRIBUTION AND LOGISTIC TO THE BASE $\sqrt{10}$

NORMAL DISTRIBUTION				LOGISTIC DISTRIBUTION -- BASE $\sqrt{10}$			
D in terms of σ_R	Percentage Higher	Expectancy Lower	Odds	D in terms of β	Percentage Higher	Expectancy Lower	Odds
0	.5000	.5000	1.000	0	.5000	.5000	1.000
.25	.5702	.4298	1.326	.25	.5713	.4287	1.334
.50	.6382	.3618	1.764	.50	.6400	.3600	1.778
.75	.7020	.2980	2.356	.75	.7034	.2966	2.371
1.00	.7602	.2398	3.170	1.00	.7598	.2402	3.162
1.25	.8116	.1884	4.309	1.25	.8083	.1917	4.217
1.50	.8555	.1445	5.920	1.50	.8495	.1505	5.645
2.00	.9228	.0772	11.95	2.00	.9091	.0909	10.00
2.50	.9614	.0386	24.9	2.50	.9467	.0533	17.8
3.00	.9830	.0170	58.0	3.00	.9694	.0306	31.6
4.00	.9977	.0023	434.	4.00	.9901	.0099	100.
5.00	.9988	.0012	834.	5.00	.9969	.0031	316.

REFERENCES

(1) Kendall, M.G. and Babington-Smith, B.: On the Method of Paired Comparisons. Biometrika 31, 324-341, 1945.

(2) Kendall, M.G.: Further Contribution to the Theory of Paired Comparisons. Biometrics 11, 43-62, 1955.

(3) Bradley, R.A. and Terry, M.: Rank Analysis of Incomplete Block Designs, a Method of Paired Comparison, Biometrics 39, 324-345, 1952.

(4) Good, I.J.: On the Marking of Chess Players, Mathematical Gazette, 39, 292-296, 1955.

(5) David, H.A.: Tournaments and Paired Comparison, Biometrika 46, 139-149, 1959.

(6) Trawinski, B.J. and David, H.A.: Selection of the Best Treatment in a Paired-Comparison Experiment, Annals of Mathematical Statistics 34: 75-91, 1963.

(7) Buhlman, H. and Huber, P.J.: Pairwise Comparison & Ranking in Tournaments, Annals of Mathematical Statistics 34: 501-510, 1963.

(8) Stevens, S.S.: On the Theory of Scales of Measurements. Science 103: 677-680, 1946.

(9) Elo, A.E.: Age Changes in Master Chess Performance, Journal of Gerontology 20: 289-299, 1965.

(10) Finney, D.J.: Probit Analysis. Cambridge University Press, New York, pp. 20-47, 1952.

(11) Pearl, R. and Reed, L.J.: On the Rate of Growth of the Population of the U.S. Since 1790 & its Mathematical Representation. Proc. Natl. Academy of Science, 6: 275-288, 1920.

(12) Berkson, J. and Reed, L.J.: Application of the Logistic Function to Experimental Data, J. of Physical Chem. 33: 760-779, 1929.

(13) Berkson, J.: Application of the Logistic Function to Bio-Assay, J. of the American Statistical Assn. 39: 357-65, 1944.

(14) Talacko, J.: Contribution to the Theory of Population Growth. Proceedings of the Actuarial Society, Prague, 149-221, 1939.

(15) Talacko, J.: A Note About a Family of Perks' Distribution. Sankhya, The Indian Journal of Statistics: 20, 323-328, 1958.

(16) Gardner, Rev. J.I.: The Normal Logistic Curve and its Generation and Use as a Probability Function. Unpublished Masters' Thesis, Marquette University, 1949.

(17) Swokowski, E.W.: The Verhulst Distribution Function as a Special Case of the Perks System. Unpublished Masters' Thesis, Marquette University, 1950.

(18) Stevens, S.S.: Ratio Scales & Category Scales for a Dozen Perceptual Continua, J. of Exp. Psychology 54: 377-411, 1957.

(19) Guilford, J.P.: Psychometric Methods. McGraw-Hill Book Co. New York, 597 pp. 2nd Ed. 1954.

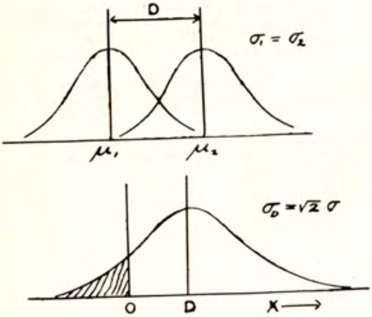

Fig. 1. Distribution of Individual Performances. (Upper); and Distribution of Differences in Performance. (Lower)

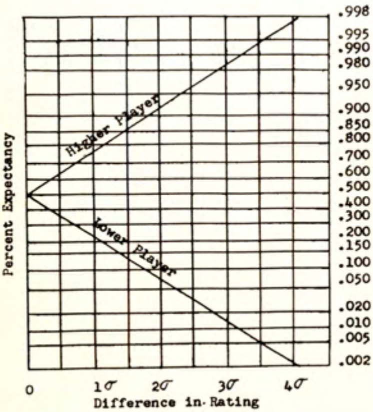

Fig. 2. Normal Rating Curve

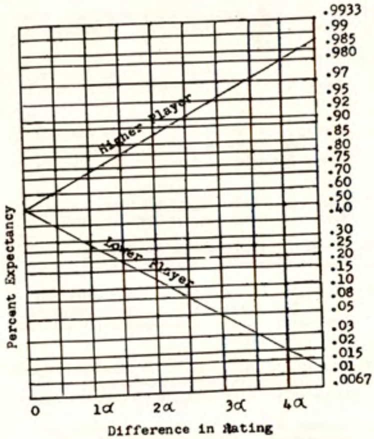

Fig. 3. Logistic Rating Curve
Base e

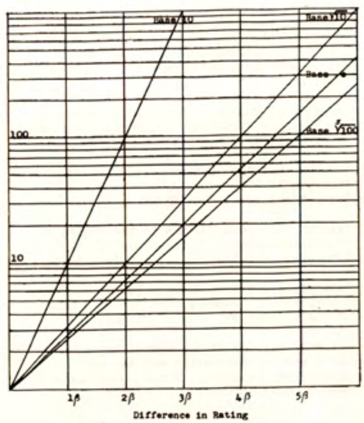

Fig. 4. Logistic Probability Ratios