Is ETSS really equitable?

L. Kalin

Meteorological and Hydrological Service, Croatia

working title...

- “Defectiveness of equitable skill scores for
a multicategorical table”

Outline

- a brief history of multicategory tables
- some thoughts on ETSS

...on multicategory tables...

- verification for standard 2x2 tables goes far into 19th century ("The Finley Affair": a 1884 paper by J. P. Finley in the *American Meteorological Journal)*
- when it comes to multicategory tables, history of development is not so long

...Vernon...

- Vernon, E. M., 1953: A New Concept Of Skill Score For Rating Quantitative Forecasts, *Mon. Wea. Rev.*

326 MONTHLY WEATHER REVIEW OCTOBER 1953

A NEW CONCEPT OF SKILL SCORE FOR RATING QUANTITATIVE FORECASTS

EDWARD M. VERNON

Weather Bureau Forecast Center, San Bruno Calif.

[Manuscript received November 8, 1951]

ABSTRACT

Skill scores for rating quantitative forecasts are proposed to take into account the deviations occurring between forecast and observed values. One score, the "deviation" skill score, weights the forecasts linearly according to the deviation; a second score, the "quadratic" skill score weights them according to the square of the deviation. These two scores are compared with the conventional skill score for two sets of forecasts, and for the same forecasts with bias introduced. It is concluded that use of either the deviation or the quadratic skill score is preferable to use of the conventional skill score in rating quantitative forecasts. Examples of the step-by-step computations of the two new scores are given.

THE DEVIATION SKILL SCORE

The skill score, as first proposed by Heidke [1] and used during recent years for certain forecast verification purposes, may be written

$$S=\frac{R-E}{T-E} \tag{1}$$

where S is the skill score, R the number of correct forecasts, T the total number of forecasts, and E the number of forecasts expected to be correct on some standard such as chance.

This method of computing a skill score places the same weight on each incorrect forecast regardless of the amount by which the observed condition deviates from the forecast. In other words, a deviation of say 10 class intervals has no more effect on the skill score than one of but 1 class interval. For some purposes it would be advantageous to have the skill score evaluate the actual amount by which forecast and observed conditions differ, i. e., take into account the magnitude of error. To accomplish this end, an analogous equation for skill score may be written

$$S_d=\frac{\sum d_e-\sum d_f}{\sum d_e} \tag{2}$$

where S_d is the skill score which considers magnitude of deviations, hereafter referred to as the "deviation skill score," Σd_f is the sum of deviations occurring between forecast and observed values, and Σd_e is the sum of deviations to be expected on some basis such as chance.

The value of Σd_f and Σd_e can best be expressed in terms of row, column, and cell totals in the typical contingency table, wherein the frequencies of forecast values are arrayed in columns and of observed values in rows, while a given cell is identified by the row and column to which it alone is common. When the standard of reference is chance, the summations become

$$\sum d_f=\sum(n_{re}d_{re}) \tag{3}$$

$$\sum d_e=\sum\left(\frac{n_r n_c}{T}d_{re}\right) \tag{4}$$

where n_r is the number of cases falling in a given row; n_c is the number of cases falling in a given column; n_{re} is the number of cases in the cell at the intersection of row r and column c; $n_r n_c/T$ is the number which would have fallen by chance in the cell representing the intersection of row r and column c; d_{re} is the deviation represented by that cell and is equal to the number of class intervals by which the cell is removed from the perfect hit cell for the same column.

When the standard of reference is climatological expectancy, according to one of the more common definitions of that standard, Σd_f remains as expressed in (3) while Σd_e becomes

$$\sum d_e=\sum\left(\frac{n_r n_{ce}}{T}d_{re}\right) \tag{5}$$

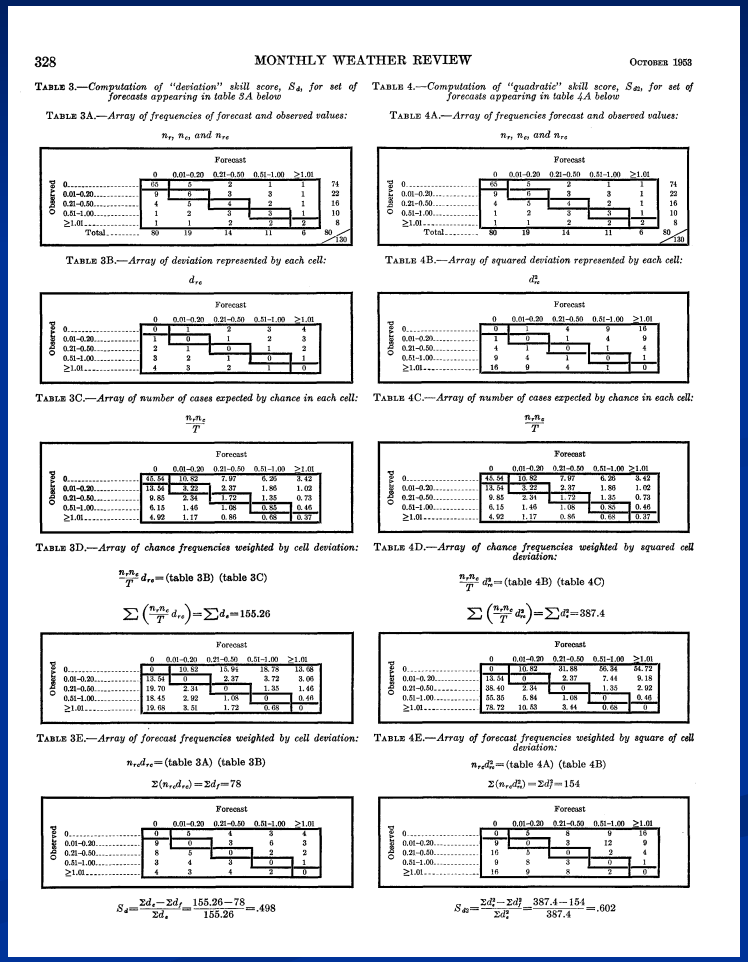
where n_{ce} represents the climatological expectancy for the column, i. e., the number of times which climatological averages would lead one to expect the observed conditions to fall in the particular class interval represented by the column. The other symbols in (5) remain as previously defined in (4) and (1).

THE QUADRATIC SKILL SCORE

In the foregoing equations all deviations are weighted linearly, a deviation of one class interval scoring as one unit deviation, two class intervals as two unit deviations,

...Vernon...

- proposed “deviation skill score” and “quadratic skill score” to take into account the amount of error (“accuracy”)
- such system would discourage forecaster to issue forecasts at extremes
- introduced weights based on chance.



328 MONTHLY WEATHER REVIEW OCTOBER 1953

TABLE 3.—*Computation of "deviation" skill score, S_d, for set of forecasts appearing in table 5A below*

TABLE 3A.—*Array of frequencies of forecast and observed values:*

n_f, n_o, and n_{fo}

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	
0	65	5	2	1	1	74
0.01–0.20	9	6	3	3	1	22
0.21–0.50	4	5	4	2	1	16
0.51–1.00	1	2	3	3	1	10
≥1.01	1	1	2	2	2	8
Total	80	19	14	11	6	80 / 130

TABLE 3B.—*Array of deviation represented by each cell:*

d_{fo}

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	1	2	3	4
0.01–0.20	1	0	1	2	3
0.21–0.50	2	1	0	1	2
0.51–1.00	3	2	1	0	1
≥1.01	4	3	2	1	0

TABLE 3C.—*Array of number of cases expected by chance in each cell:*

$\frac{n_f n_o}{T}$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	45.54	10.82	7.97	6.35	3.42
0.01–0.20	13.54	3.22	2.37	1.86	1.02
0.21–0.50	9.85	2.34	1.72	1.35	0.73
0.51–1.00	6.15	1.46	1.08	0.85	0.46
≥1.01	4.92	1.17	0.86	0.68	0.37

TABLE 3D.—*Array of chance frequencies weighted by cell deviation:*

$\frac{n_f n_o}{T} d_{fo}$ = (table 3B) (table 3C)

$\sum\left(\frac{n_f n_o}{T} d_{fo}\right) = \sum d_e = 155.26$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	10.82	15.94	19.05	13.68
0.01–0.20	13.54	0	2.37	3.72	3.06
0.21–0.50	19.70	2.34	0	1.35	1.46
0.51–1.00	18.45	2.92	1.08	0	0.46
≥1.01	19.68	3.51	1.72	0.68	0

TABLE 3E.—*Array of forecast frequencies weighted by cell deviation:*

$n_f d_{fo}$ = (table 3A) (table 3B)

$\Sigma(n_f d_{fo}) = \Sigma d_f = 78$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	5	4	3	4
0.01–0.20	9	0	3	6	3
0.21–0.50	8	5	0	2	2
0.51–1.00	3	4	3	0	1
≥1.01	4	3	4	2	0

$S_d = \frac{\Sigma d_e - \Sigma d_f}{\Sigma d_e} = \frac{155.26 - 78}{155.26} = .498$

TABLE 4.—*Computation of "quadratic" skill score, S_{d2}, for set of forecasts appearing in table 4A below*

TABLE 4A.—*Array of frequencies forecast and observed values:*

n_f, n_o, and n_{fo}

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	
0	65	5	2	1	1	74
0.01–0.20	9	6	3	3	1	22
0.21–0.50	4	5	4	2	1	16
0.51–1.00	1	2	3	3	1	10
≥1.01	1	1	2	2	2	8
Total	80	19	14	11	6	80 / 130

TABLE 4B.—*Array of squared deviation represented by each cell:*

d_{fo}^2

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	1	4	9	16
0.01–0.20	1	0	1	4	9
0.21–0.50	4	1	0	1	4
0.51–1.00	9	4	1	0	1
≥1.01	16	9	4	1	0

TABLE 4C.—*Array of number of cases expected by chance in each cell:*

$\frac{n_f n_o}{T}$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	45.54	10.82	7.97	6.26	3.42
0.01–0.20	13.54	3.22	2.37	1.86	1.02
0.21–0.50	9.85	2.34	1.72	1.35	0.73
0.51–1.00	6.15	1.46	1.08	0.85	0.46
≥1.01	4.92	1.17	0.86	0.68	0.37

TABLE 4D.—*Array of chance frequencies weighted by squared cell deviation:*

$\frac{n_f n_o}{T} d_{fo}^2$ = (table 4B) (table 4C)

$\sum\left(\frac{n_f n_o}{T} d_{fo}^2\right) = \sum d_e^2 = 387.4$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	10.82	31.88	56.34	54.72
0.01–0.20	13.54	0	2.37	7.44	9.18
0.21–0.50	38.40	2.34	0	1.35	2.92
0.51–1.00	55.35	5.84	1.08	0	0.46
≥1.01	78.72	10.53	3.44	0.68	0

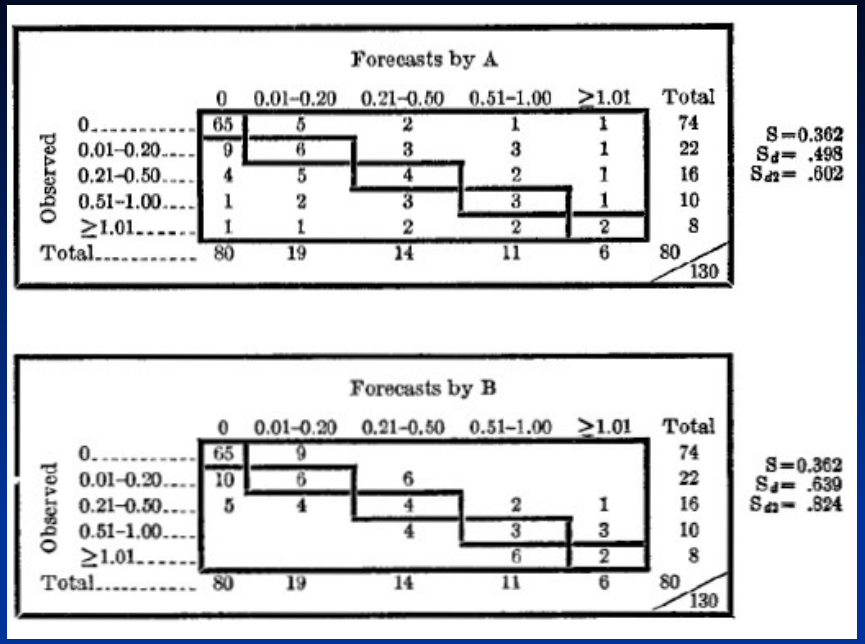
TABLE 4E.—*Array of forecast frequencies weighted by square of cell deviation:*

$n_f d_{fo}^2$ = (table 4A) (table 4B)

$\Sigma(n_f d_{fo}^2) = \Sigma d_f^2 = 154$

Observed \ Forecast	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01
0	0	5	8	9	16
0.01–0.20	9	0	3	12	9
0.21–0.50	16	5	0	2	4
0.51–1.00	9	8	3	0	1
≥1.01	16	9	8	2	0

$S_{d2} = \frac{\Sigma d_e^2 - \Sigma d_f^2}{\Sigma d_e^2} = \frac{387.4 - 154}{387.4} = .602$



Forecasts by A

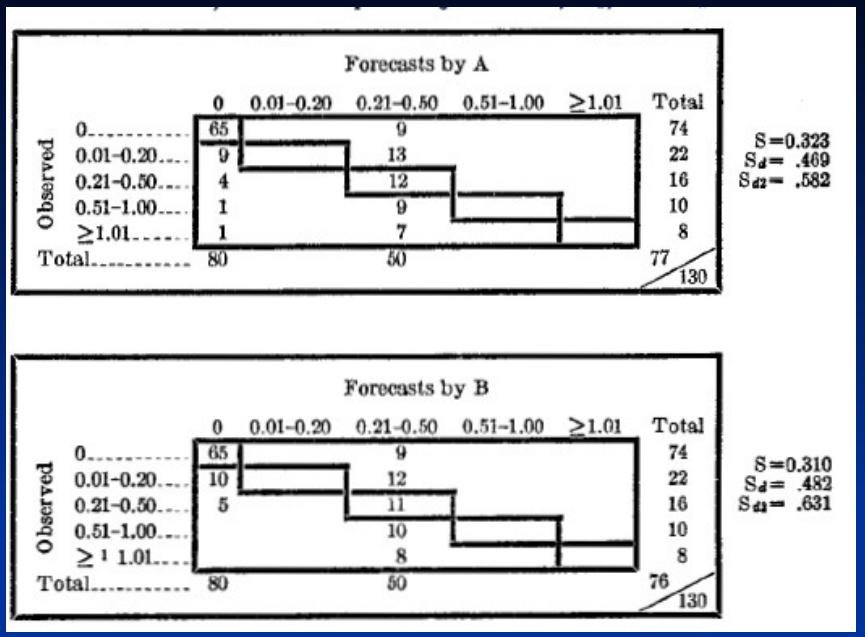
Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65	5	2	1	1	74
0.01–0.20	9	6	3	3	1	22
0.21–0.50	4	5	4	2	1	16
0.51–1.00	1	2	3	3	1	10
≥1.01	1	1	2	2	2	8
Total	80	19	14	11	6	80/130

$S = 0.362$
$S_d = .498$
$S_{d2} = .602$

Forecasts by B

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65	9				74
0.01–0.20	10	6	6			22
0.21–0.50	5	4	4	2	1	16
0.51–1.00			4	3	3	10
≥1.01				6	2	8
Total	80	19	14	11	6	80/130

$S = 0.362$
$S_d = .639$
$S_{d2} = .824$



Forecasts by A

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65		9			74
0.01–0.20	9		13			22
0.21–0.50	4		12			16
0.51–1.00	1		9			10
≥1.01	1		7			8
Total	80		50			77/130

$S = 0.323$
$S_d = .469$
$S_{d2} = .582$

Forecasts by B

Observed	0	0.01–0.20	0.21–0.50	0.51–1.00	≥1.01	Total
0	65		9			74
0.01–0.20	10		12			22
0.21–0.50	5		11			16
0.51–1.00			10			10
≥1.01			8			8
Total	80		50			76/130

$S = 0.310$
$S_d = .482$
$S_{d2} = .631$

- possibility that rating on the basis of the size of the deviations will lead forecasters to bias their forecasts towards the middle class interval (where is a minimum possible deviation), rather rather than rather than trying to catch extreme

...Bryan...

- Bryan, J. G., and I. Enger, 1967: Use of probability fo recasts to maximize various skill scores. *J. Appl. Meteor.*, 6, 762-769

by applying Eq. (3.11) to the same probability distribution used in Section 2C.

$v=0.0$

1.00	0.00	−1.00
−0.46	0.54	−0.46
−1.00	0.00	1.00

$v=0.2$

0.80	−0.20	−1.20
−0.57	0.43	−0.57
−1.20	−0.20	0.80

$v=0.4$

0.60	−0.40	−1.40
−0.68	0.32	−0.68
−1.40	−0.40	0.60

$v=0.6$

0.40	−0.60	−1.60
−0.78	0.22	−0.78
−1.60	−0.60	0.40

$v=0.8$

0.20	−0.80	−1.80
−0.89	0.11	−0.89
−1.80	−0.80	0.20

$v=1.0$

0.00	−1.00	−2.00
−1.00	0.00	−1.00
−2.00	−1.00	0.00

$v=0.2$

1.60	0.60	−0.40	−1.40	−2.40
−0.03	0.97	−0.03	−1.03	−2.03
−1.27	−0.27	0.73	−0.27	−1.27
−2.03	−1.03	−0.03	0.97	−0.03
−2.40	−1.40	−0.40	0.60	1.60

$v=0.5$

1.00	0.00	−1.00	−2.00	−3.00
−0.39	0.61	−0.39	−1.39	−2.39
−1.54	−0.54	0.46	−0.54	−1.54
−2.39	−1.39	−0.39	0.61	−0.39
−3.00	−2.00	−1.00	0.00	1.00

$v=0.8$

0.40	−0.60	−1.60	−2.60	−3.60
−0.76	0.24	−0.76	−1.76	−2.76
−1.82	−0.82	0.18	−0.82	−1.82
−2.76	−1.76	−0.76	0.24	−0.76
−3.60	−2.60	−1.60	−0.60	0.40

Gringorten, (1967)

- Gringorten score awards forecasts of extremes (too much?)
- compared to Bryan, bad forecasts are not equivalently punished

TABLE 3. Example of scores for the forecast of mutually exclusive events X_0, X_1, X_2, X_3 whose climatic frequencies are ${}_cP_0=0.05$, ${}_cP_1=0.10$, ${}_cP_2=0.10$, ${}_cP_3=0.75$.

Forecast event	Bryan Observed event				Gringorten Observed event			
	X_0	X_1	X_2	X_3	X_0	X_1	X_2	X_3
X_0	95	−5	−5	−5	20	0	0	0
X_1	−10	90	−10	−10	0	10	0	0
X_2	−10	−10	90	−10	0	0	10	0
X_3	−75	−75	−75	25	0	0	0	1.33

LEPS

- Linear Error in Probability Space
- originally introduced by Ward and Folland (1991)

$$LEPS = (1/n) \Sigma | Pv - Pf |$$

- corresponds to MAE transformed into probability space
- Revised, normalised LEPS (Potts et al., 1996)

LEPS...

- encourages forecasting of extremes of the cilimatological distribution (not as much as Gringorten)

TABLE 2. LEPS S'' scores for quints.

	Observation				
Forecast	Q1	Q2	Q3	Q4	Q5
Q1	1.28	0.52	−0.20	−0.68	−0.92
Q2	0.52	0.56	0.04	−0.44	−0.68
Q3	−0.20	0.04	0.32	0.04	−0.20
Q4	−0.68	−0.44	0.04	0.56	0.52
Q5	−0.92	−0.68	−0.20	0.52	1.28

ETSS

- proposed by Gandin and Murphy (1992)
- “equitability”
  - equitable score takes zero value (“no skill”)for random forecasts and for unvarying forecasts of a constant category
  - takes value 1 for perfect forecast

ETSS

- Gerrity (1992) showed that, for a multicategory table NxN, ETSS can be calculated as mean of N-1 values of Pierce Skill Score of collapsed 2x2 tables
- since PSS has some deficiences, ETSS will probably inherit them?

ETSS

- ECMWF 24-hour precipitation forecast (Green Book, 2007, Hungary)

- 4 categories
  - C0 less than 0.1 mm
  - C1 0.1 - 2 mm
  - C2 2 – 10 mm
  - C3 bigger than 10 mm

ETSS

F\O	C0	C1	C2	C3
C0	5812	179	46	5
C1	3292	1165	383	56
C2	608	735	971	279
C3	15	45	247	326

- ETSS = .561

ETSS – modified

F\O	C0	C1	C2	C3
C0	5812	179	46	5
C1	3292	1165	383	56
C2	0	0	0	0
C3	623	780	1218	605

- ETSS = .643

...scoring matrices...

- Based on marginal probabilities one can calculate the scoring matrix

P0=0.687

P1=0.150

P2=0.116

P3=0.047

F\O	C0	C1	C2	C3
C0	.233	-.251	-.650	-1
C1	-.252	.812	.414	.064
C2	-.650	. 414	2.455	2.105
C3	-1	.064	2.105	9.194

...scoring matrices...

- Based on marginal probabilities one can calculate the scoring matrix

P0=0.687

P1=0.150

P2=0.116

P3=0.047

F\O	C0	C1	C2	C3
C0	.233	-.251	-.650	-1
C1	-.252	.812	.414	.064
C2	-.650	. 414	2.455	2.105
C3	-1	.064	2.105	9.194

ETSS

- dependent on the number of categories
- if a NxN contingency table is reduced to

  (N-1)x(N-1)...

  ...and finally to 2x2

  ETSS is constantly decreasing...

ETSS

- favorizing rare events...
- dependent on bias, which may lead to “hedging”
- overforecasting of extreme (and rare) categories leads to increase of the score compared to the original forecast
- equitable
- proper?
- suitable for “hedging”

...on hedging...