United States
National Weather Service
Verification Program

International Verification Workshop

Equitable Skill Score Use in the U.S. National Weather Service

Chuck Kluepfel (Charles.Kluepfel@noaa.gov)
September 15, 2004

Contingency Tables

	FORECASTS 1	2	3	Totals
OBSERVED 1	Hit_1			O_1
2		Hit_2		O_2
3			Hit_3	O_3
Totals	F_1	F_2	F_3	

Equitable Scores

- Take into account weighted event probabilities (sometimes use climatology)
- Discourage forecast "hedging"
  - Penalizes over-forecasting the most climatologically likely events
  - Rewards correct forecasts of rare events
- Cannot be "gamed" by a clever forecaster.
- Boundaries of Score
  - Zero: no skill (climatology)
  - 1.00: perfect forecasts

Examples

- Heidke (1926) skill score
- Peirce (1884) skill score
  - True skill statistic
  - Hanssen-Kuippers discriminant
- Gandin and Murphy (MWR 1992)
- Gerrity (MWR 1992)

Heidke / Peirce Score

$$HSS = \frac{\sum_{i=1}^{k} H_i - \sum_{i=1}^{k} O_i F_i}{N - \sum_{i=1}^{k} O_i F_i}$$

$$PSS = \frac{\sum_{i=1}^{k} H_i - \sum_{i=1}^{k} O_i F_i}{N - \sum_{i=1}^{k} O_i O_i}$$

Gandin and Murphy Score (GM)
Gerrity Score (GS)

$$ESS = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} s_{ij}$$

Properties of scores

- Heidke and Peirce do not reward forecasts off the diagonal
- Gandin and Murphy (GM) provide partial credit to “near hits”
- Gerrity is a subset of GM
- Livezey (2003) compares the above scores and recommends Gerrity for ordinal categorical event forecasts

GM / Gerrity Equitable Skill Scores

Forecasts are scored in the following manner:

- **Each cell of the contingency table is multiplied by a scoring factor with relative levels of rewards and penalties**
- **Each of the multiplied cell values is summed for a total score.**
- **Graduated reward/penalty system**

GM / Gerrity Equitable Skill Scores

Graduated reward/penalty system:

- Forecast hits receive the most reward for each category of events
- A large forecast error is penalized more than a small error for a given category of events.
- A large reward for correct forecasts of rare events.
- A relatively small reward for correct forecasts of common events.
- Less penalty is assigned to an incorrect forecast of a rare event than a similar size error of a common event. “Near hits” of rare events receive a modest reward.

Example 1

P matrix

10	10	10
10	10	10
10	10	10

S matrix

1.25	- .25	- 1
- .25	0.5	- .25
- 1	- .25	1.25

Example 2

P matrix

22	5	3
7	19	4
0	0	1

S matrix

.52	- .49	- 1
- .49	1	- .02
- 1	- .02	30.5

Rare event – Impact on Score
7-category wind speed - Nation

Year	Gerrity score	N	>32 kt Hits/total	>32 kt % Corr	Delta
1995	.38	12,823	3 / 9	33	.02
1996	.47	13,145	11 / 17	65	.01
1997	.51	10,770	4 / 6	67	.03
1998	.53	12,717	19 / 24	79	.01
1999	.54	7371	6 / 7	86	.03
2000	.42	14,282	4 / 9	44	.02
2001	.38	28,064	1 / 9	11	.02
2002	.47	28,787	9 / 25	45	.01
2003	.44	30,751	6 / 15	40	.01
2004	.53	27,691	8 / 10	80	.02

Scoring Matrix

Select a formula for computing an equitable scoring matrix

- **Gandin and Murphy (ordinal, non-circular elements)**
- **Gerrity (ordinal, non-circular elements)**
- **We need a separate method for wind direction due to its circular properties**
  - **Burroughs (1993): assumes equal distribution of wind directions**
  - **New mathematical ideas?**

Scoring Matrix

How do we represent a random sample?

- **Establish independence from the distribution of forecasts being evaluated**
- **Build from distribution of observations**
  - **Long-term climatology (historical dataset)**
  - **Dataset from which skill score is computed**

Scoring Matrix

Problems with long-term history

- Climate change – Does long-term record represent current data?
- Lots of number crunching for specialized computations

Problems with using the sample (of score)

- Large datasets - Minimal problems
- Small datasets – volatile statistics

Scoring Matrix History in US NWS

- **Burroughs (EMC) – marine forecast verification:**
  - **Marine wind and wave verification (started 1994)**
  - **Used Gerrity score**
  - **Static climatology used for all areas/months**
- **NWS Headquarters took control - Sep 2002**
  - **NWS began computing scoring matrix from the sample of requested verification data**
  - **User of *Stats on Demand* selects the sample**
  - **Small samples (space or time) – volatile scores (NWS adds a "delta value" to address volatility)**
  - **Future – Return to climatology? If so, tailor climatology to area and time of year of sample**

Unanswered Questions

- Scoring matrix / climatology issue
- Gerrity score
  - Only works for ordinal non-circular elements
  - Circular element - wind direction

References

Gandin, L.S. and A.H. Murphy (1992). Equitable scores for categorical forecasts. *Mon. Weather Rev.,* **120,** 361-370.

Gerrity, J.P. Jr (1992). A note on Gandin and Murphy's equitable skill score. *Mon. Weather Rev.,* **120,** 2707-2712.

Burroughs, L.D. (1993). National marine verification program—verification statistics. *NMC Office Note 400,* OPC Contribution 79.

Heidke, P. (1926). Berechnung der erfolges und der gute der windstarkevorhersagen im sturmwarnungdienst. *Geogr. Ann.,* **8,** 301-349.

Joliffe, I.T. and D.B. Stephenson (2003). *Forecast verification: a practitioner's guide in atmospheric science.* John Wiley and Sons. See chapter 4 concerning categorical events, written by R. E. Livezey.

Peirce, C.S. (1884). The numerical measure of the success of predictions. *Science,* **4,** 453-454.