Forecast verification using scoring rules

Chris Ferro

Department of Mathematics
University of Exeter, UK

7th International Verification Methods Workshop
(Berlin, 9 May 2017)

Scoring rules

Forecasts $f_1, \ldots, f_n$
Outcomes $x_1, \ldots, x_n$

Definition: A **scoring rule**, $s(f, x)$, gives a numerical score to each forecast.

Example: $s(f, x) = (f - x)^2$

We measure performance by the mean score,

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(f_i, x_i).$$

Scoring rules

Forecasts $f_1, \ldots, f_n$
Outcomes $x_1, \ldots, x_n$

Definition: A **scoring rule**, $s(f, x)$, gives a numerical score to each forecast.

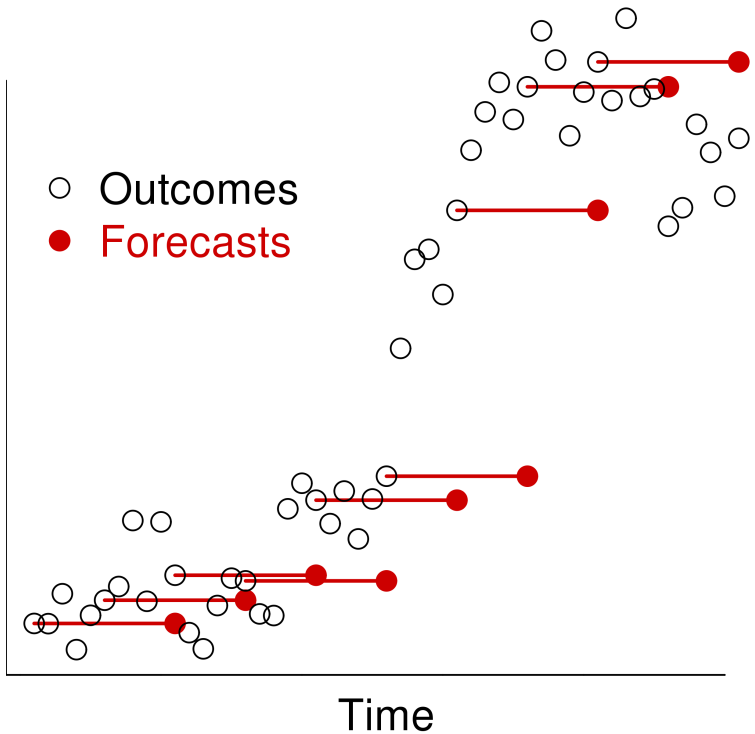
Example: $s(f, x) = (f - x)^2$

We measure performance by the mean score,

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(f_i, x_i).$$

Other measures can be spuriously inflated by trends.

Example: correlation = .87



Proper scoring rules

Let $x_1, x_2, \ldots$ have frequency distribution p.

Suppose we issue the same forecast, f, for all $x_1, x_2, \ldots$

The best choice is $f = p$.

Definition: A scoring rule is **proper** if the long-run mean score is optimized by $f = p$.

Example: Let $x \in \{0, 1\}$ and $f = \Pr(x = 1)$. Then $(f - x)^2$ is proper; $|f - x|$ is improper.

50th anniversary of 'proper'!

Proper scoring rules

Let $x_1, x_2, \ldots$ have frequency distribution p.

Suppose we issue the same forecast, f, for all $x_1, x_2, \ldots$

The best choice is $f = p$.

Definition: A scoring rule is **proper** if the long-run mean score is optimized by $f = p$.

Example: Let $x \in \{0, 1\}$ and $f = \Pr(x = 1)$. Then $(f - x)^2$ is proper; $|f - x|$ is improper.

50th anniversary of 'proper'!

Proper scoring rules

Let $x_1, x_2, \ldots$ have frequency distribution p.

Suppose we issue the same forecast, f, for all $x_1, x_2, \ldots$

The best choice is $f = p$.

Definition: A scoring rule is **proper** if the long-run mean score is optimized by $f = p$.

Example: Let $x \in \{0, 1\}$ and $f = \Pr(x = 1)$. Then $(f - x)^2$ is proper; $|f - x|$ is improper.

50th anniversary of 'proper'!

Lots Of Verification Excitement

Properties: interpretation

We can interpret proper scores as losses in decision problems.

Let $\mathcal{L}(a, x)$ be the loss following action a and outcome x.

Let a_f be the optimal (Bayes) action according to forecast f.

Then $s(f, x) = \mathcal{L}(a_f, x)$ is a proper scoring rule.

Properties: interpretation

We can interpret proper scores as losses in decision problems.

Let $\mathcal{L}(a, x)$ be the loss following action a and outcome x.

Let a_f be the optimal (Bayes) action according to forecast f.

Then $s(f, x) = \mathcal{L}(a_f, x)$ is a proper scoring rule.

Example: Suppose that we lose L if we do not act and $x = 1$, and lose C if we do act. If the cost-loss ratio, C/L, is uniformly distributed on $(0, 1)$ then our average loss from acting on f is $(f - x)^2 L/2$ more than from acting on a perfect forecast.

A Brier score of 0.1 means that a group of decision makers with a uniform distribution of cost-loss ratios will lose 5% of L more than they would do with perfect forecasts.

Properties: decomposition

Let p_f be the distribution of outcomes following forecast f.

Definition: Forecasts are **calibrated** if $p_f = f$ for all f.

Definition: Forecasts are **sharp** if p_f is concentrated for all f.

Properties: decomposition

Let p_f be the distribution of outcomes following forecast f.

Definition: Forecasts are **calibrated** if $p_f = f$ for all f.

Definition: Forecasts are **sharp** if p_f is concentrated for all f.

Proper scores measure calibration and sharpness:

$$\bar{s} = \mathcal{C} + \mathcal{S},$$

where $\mathcal{C}$ is optimized when the forecasts are calibrated and $\mathcal{S}$ is optimized when the forecasts are perfectly sharp.

Example: $\overline{(f - x)^2} = \overline{(f - p_f)^2} + \overline{p_f(1 - p_f)}$

Properties: sensitivity to distance

Definition: A scoring rule is **local** if it depends on only $f(x)$.

Definition: A scoring rule is **distance-sensitive** if the score improves whenever we move some probability nearer to x.

Local scores favour forecasts with more probability at x.

Distance-sensitive scores favour forecasts with more probability near x.

Properties: sensitivity to distance

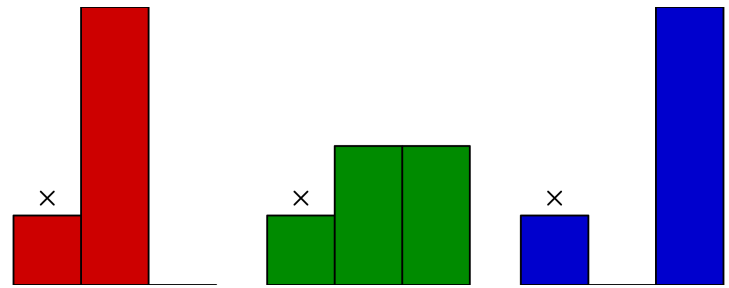
Definition: A scoring rule is **local** if it depends on only $f(x)$.

Definition: A scoring rule is **distance-sensitive** if the score improves whenever we move some probability nearer to x.

Local scores favour forecasts with more probability at x.

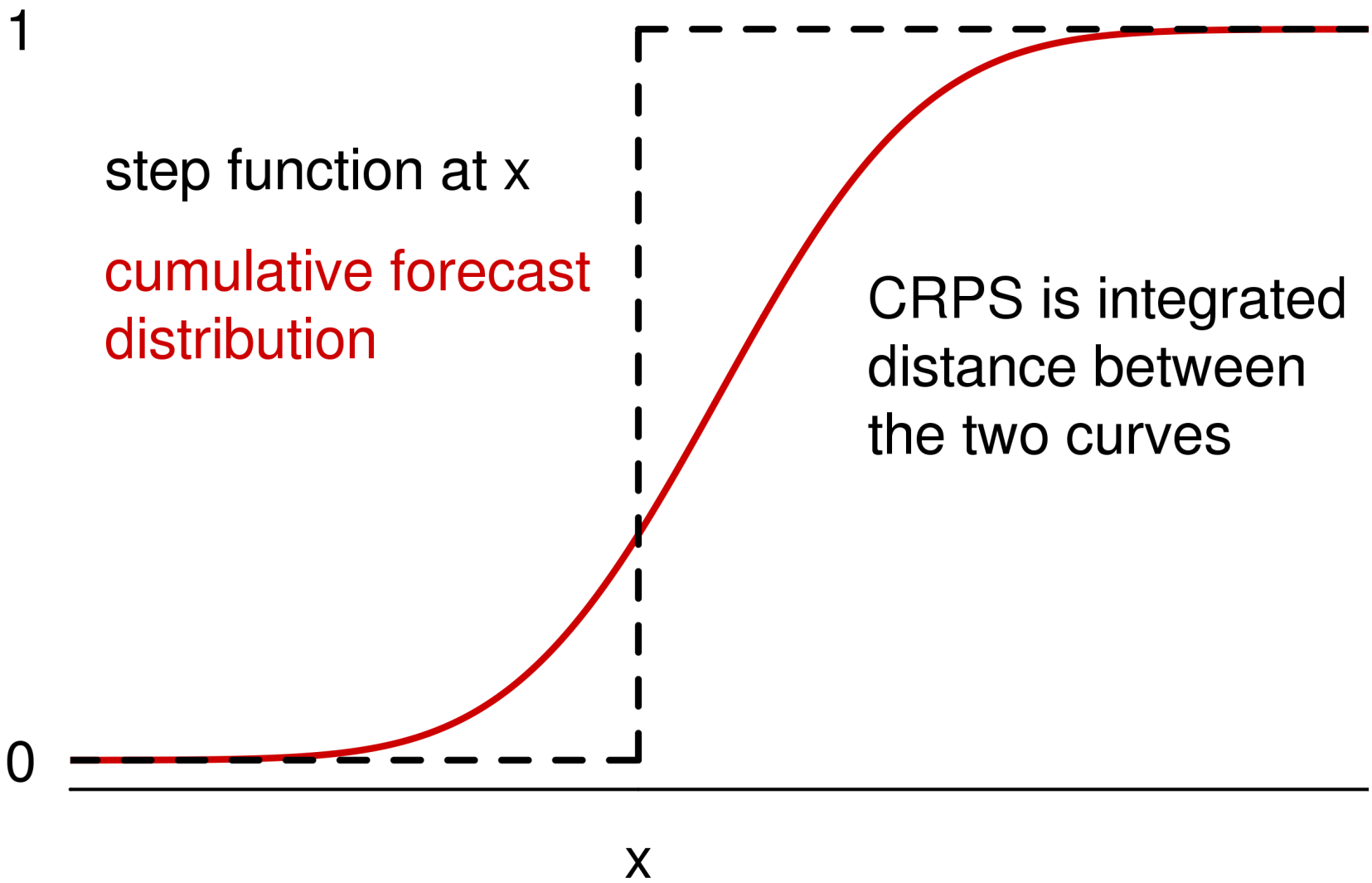
Distance-sensitive scores favour forecasts with more probability near x.

Example: Logarithmic (local), ranked probability (distance-sensitive) and quadratic (neither) scores for three forecasts.

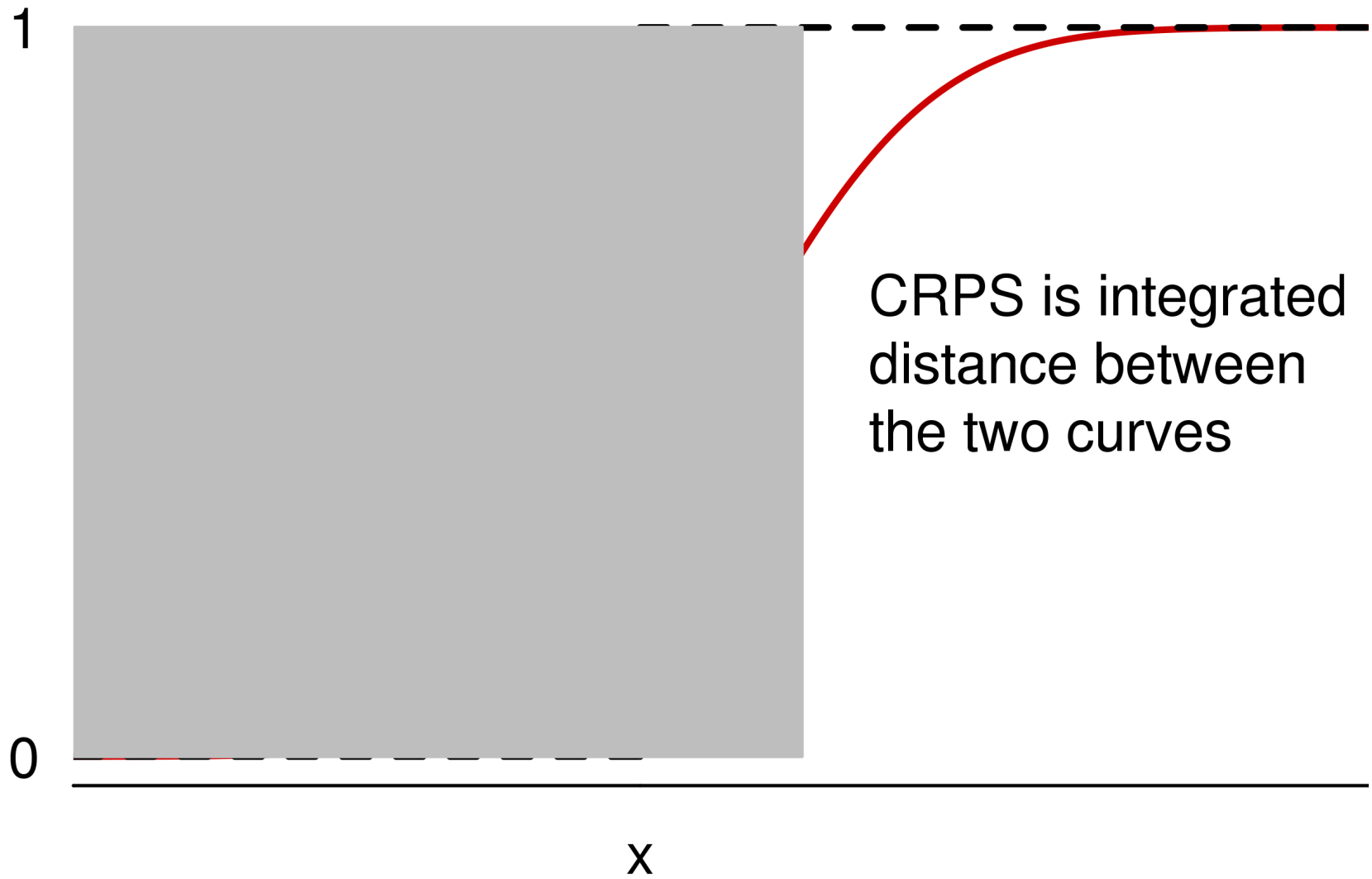


LS	**1.61**	**1.61**	**1.61**
RPS	**0.64**	0.80	1.28
QS	0.28	**−0.04**	0.28

Properties: sensitivity in region of interest



Properties: sensitivity in region of interest



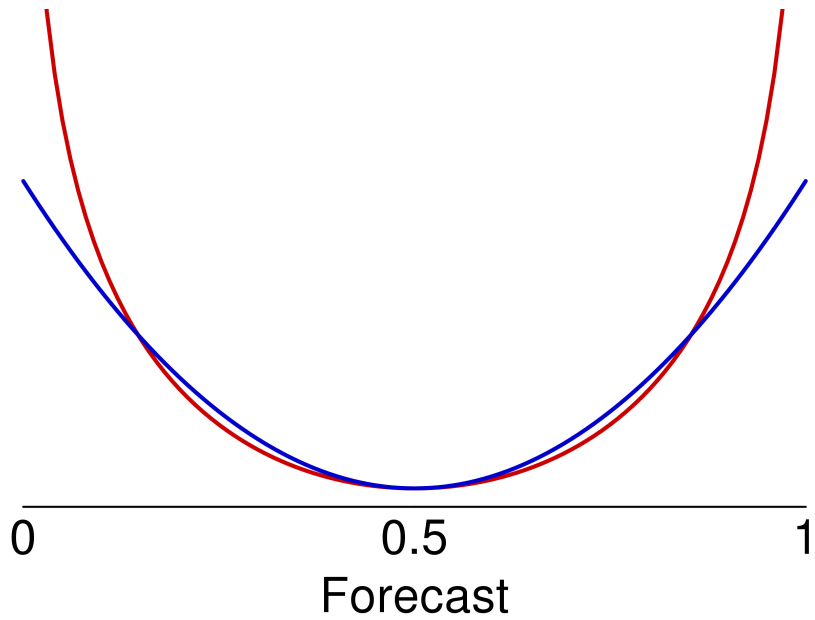
Properties: sensitivity to error

Suppose we issue the same forecast, f, for $x_1, x_2, \ldots$

Let $x_1, x_2, \ldots$ have frequency distribution p.

The long-run mean score is optimized by $f = p$ but how sensitive is it to $f \neq p$?

Example: Mean quadratic and logarithmic scores for different forecasts when $p = 0.5$. The log score is more sensitive to large errors and (on this scale) less sensitive to small errors.



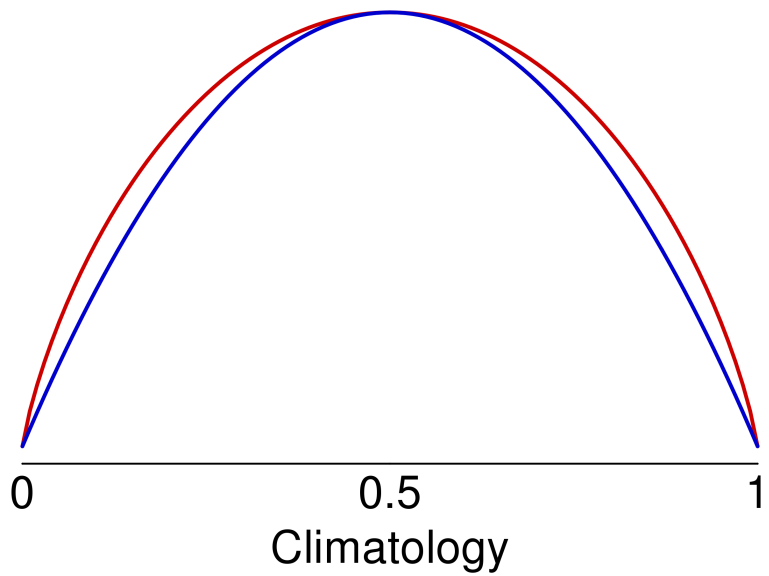
Properties: sensitivity to climatology

Let $x_1, x_2, \ldots$ have frequency distribution (climatology) p.

Suppose we issue climatology, $f = p$, for all $x_1, x_2, \ldots$

How does the long-run mean score vary with climatology?

Example: Mean logarithmic and quadratic scores (scaled) for different climatologies. Both scoring rules give their worst scores when $p = 0.5$.



Properties: sensitivity to climatology

Let $x_1, x_2, \ldots$ have frequency distribution (climatology) p.

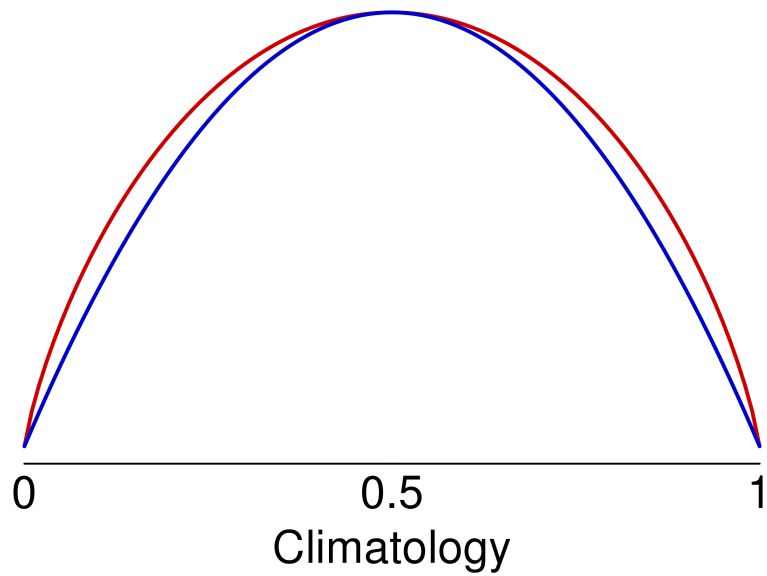
Suppose we issue climatology, $f = p$, for all $x_1, x_2, \ldots$

How does the long-run mean score vary with climatology?

Symmetric scoring rules give their worst scores when climatology is uniform.

Asymmetric proper scoring rules can be designed so that all climatologies score zero.

Example: Mean logarithmic and quadratic scores (scaled) for different climatologies. Both scoring rules give their worst scores when $p = 0.5$.



Inference

Assume $s(f_1, x_1), \ldots, s(f_m, x_m)$ are independent draws from a population with mean μ.

Assume $s(g_1, y_1), \ldots, s(g_n, y_n)$ are independent draws from a population with mean $\mu + \delta$.

If the outcomes, $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, are from the same population or are identical ($x = y$) then we may compare the two forecasters using standard two-sample inference for δ.

Inference

Assume $s(f_1, x_1), \ldots, s(f_m, x_m)$ are independent draws from a population with mean μ.

Assume $s(g_1, y_1), \ldots, s(g_n, y_n)$ are independent draws from a population with mean $\mu + \delta$.

If the outcomes, $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, are from the same population or are identical ($x = y$) then we may compare the two forecasters using standard two-sample inference for δ.

If the outcomes are from different populations then we cannot compare the forecasters: we don't know how well each would have forecast the other's outcomes.

We can compare their performances, though, especially using measures (e.g. ROC curves) that don't depend on climatology.

Some suggestions

Rank forecasts with proper scores.

Interpret scores using decision theory.

Measure calibration and sharpness by decomposing scores.

Use distance-sensitive or local scores if ‘distance’ has meaning.

Stress regions of interest by weighting scores.

Choose scores that represent relevant decision problems.

Use asymmetric scores to award zero to reference forecasts.

Understand the sensitivities of scores to different errors.

Beware that some scores are costly to compute.

Extensions: ensembles and probability intervals

Let $x_1, x_2, \ldots$ have distribution p and suppose that we sample ensembles, $w_1, w_2, \ldots$, from one distribution, f, for all $x_1, x_2, \ldots$

Definition: A scoring rule, $s(w, x)$, is **fair** if the long-run mean score is optimized when $f = p$.

Example: Let $x \in \{0, 1\}$ and $\bar{w}$ be the mean of an ensemble of size $m > 1$. Then $(\bar{w} - x)^2 - \bar{w}(1 - \bar{w})/(m - 1)$ is fair.

Extensions: ensembles and probability intervals

Let $x_1, x_2, \ldots$ have distribution p and suppose that we sample ensembles, $w_1, w_2, \ldots$, from one distribution, f, for all $x_1, x_2, \ldots$

Definition: A scoring rule, $s(w, x)$, is **fair** if the long-run mean score is optimized when $f = p$.

Example: Let $x \in \{0, 1\}$ and $\bar{w}$ be the mean of an ensemble of size $m > 1$. Then $(\bar{w} - x)^2 - \bar{w}(1 - \bar{w})/(m - 1)$ is fair.

Let forecasts be probability intervals (e.g. 0-5%, 5-15%, ...) and suppose we issue the same interval, I, for all $x_1, x_2, \ldots$

Definition: A scoring rule, $s(I, x)$, is **interval-proper** if the long-run mean score is optimized when I contains p.

Example: Let intervals $I_k = [c_{k-1}, c_k]$ partition $[0, 1]$. Then $s(I_k, x) = (c_{k-1} - x)(c_k - x)$ is interval-proper.

Extensions: observation error

x = true outcome
y = observed outcome
f = probability forecast for x

We would like to use $s_0(f, x)$ for a proper scoring rule s_0 but we don't know x.

Extensions: observation error

x = true outcome
y = observed outcome
f = probability forecast for x

We would like to use $s_0(f, x)$ for a proper scoring rule s_0 but we don't know x.

Definition: A scoring rule, s, is **error-corrected** if the long-run mean of $s(f, y)$ equals the long-run mean of $s_0(f, x)$.

The error-corrected score is an unbiased estimate of the score that would be awarded by s_0 if we knew the true outcome.

Extensions: observation error

x = true outcome
y = observed outcome
f = probability forecast for x

We would like to use $s_0(f, x)$ for a proper scoring rule s_0 but we don't know x.

Definition: A scoring rule, s, is **error-corrected** if the long-run mean of $s(f, y)$ equals the long-run mean of $s_0(f, x)$.

The error-corrected score is an unbiased estimate of the score that would be awarded by s_0 if we knew the true outcome.

Example: Let $x, y \in \{0, 1\}$ and $f \in [0, 1]$ with misclassification probabilities $r_x = \Pr(y \neq x \mid x)$. If $r_0 + r_1 \neq 1$ then

$$s(f, y) = s_0(f, y) + \frac{r_y\{s_0(f, y) - s_0(f, 1 - y)\}}{1 - r_0 - r_1}.$$

How can we get more from scoring rules?

More guidance on how to design scoring rules.

More guidance on how to interpret scores with decision theory.

More ways to interpret scores graphically.

Effective decompositions for calibration and sharpness.

Relate scores to meteorological errors/biases.

Extensions to time series, fields and multivariate features.

Better inference for rare events.

Formal analysis of decisions to update operational models.

References

Reviews

Gneiting, Raftery 2007. J Amer Statist Assoc 102: 359
Winkler 1996. Test 5: 1

Spurious skill

Fricker, Ferro, Stephenson 2013. Met Apps 20: 246
Hamill, Juras 2006. Q J R Meteorol Soc 132: 2905

Proper scoring rules

Brier 1950. Mon Wea Rev 78: 1
Good 1952. J R Statist Soc 14: 107

Decision theory

Dawid 1986. Encyclopedia Statist Sci 7: 210
Johnstone, Jose, Winkler 2011. Decision Analysis 8: 256
Murphy 1966. J Appl Meteorology 5: 534
Schervish 1989. Ann Stat 17: 1856

Decompositions

Bröcker 2009. Q J R Meteorol Soc 135: 1512
Candille, Talagrand 2005. Q J R Meteorol Soc 131: 2131
Sanders 1963. J Appl Meteorology 2: 191

Local/distance-sensitive

Bernardo 1979. Ann Stat 7: 686
Epstein 1969. J Appl Meteorology 8: 985
Staël von Holstein 1970. J Appl Meteorology 9: 360

Regions of interest

Diks, Panchenko, van Dijk 2011. J Econometrics 163: 215
Gneiting, Ranjan 2011. J Bus Econ Stat 29: 411
Lerch, Thorarinsdottir et al. 2017. Statist Sci 32: 106
Matheson, Winkler 1976. Management Sci 22: 1087

Sensitivity to errors

Bickel 2007. Decision Analysis 4: 49
Machete 2013. J Stat Plan Inf 1443: 1781
Merkle, Steyvers 2013. Decision Analysis 10: 292
Murphy, Winkler 1970. Acta Psychologica 34: 273

Asymmetric scoring rules

Jose, Nau, Winkler 2008. Operations Res 56: 1146
Jose, Nau, Winkler 2009. Management Sci 55: 582
Winkler 1994. Management Sci 40: 1395

Fair scoring rules

Ferro 2014. Q J R Meteorol Soc 140: 1917

Interval-proper scoring rules

Mitchell, Ferro 2017. Q J R Meteorol Soc (in press)

Observation error-corrected scoring rules

Ferro 2017. Submitted manuscript available on request