# *A test to compare KS statistics*

David J. Hand
Imperial College London
and
Winton Capital Management

and

Wojtek J. Krzanowski
University of Exeter

Reasons for comparing scorecards

- to compare: which is *better*, is the difference *substantively* significant
- to improve: does '*this change*' make things better?
  (one of the overwhelming merits of objective scoring methods)

Don't evaluate using design data

- use independent test set
- or bootstrap, l-o-o, cross-validation, ....

Note: testing by randomly splitting data into design and test sets is likely to overestimate future performance

Criteria for evaluating scorecards include:

- KS
- AUC, Gini
- H-measure
- and others . . . . .

And we *compare* scorecards by comparing the values of *test criteria*

*But simple **comparison** of such measures is insufficient*

Need to know if a difference is ***statistically significant***

That is, we need to answer

***Could any observed difference be due to chance?***

# Definition of KS statistic

$F_G(s)$ = cdf of scores of goods

$F_B(s)$ = cdf of scores of bads

$$KS = \sup_s \left| F_B(s) - F_G(s) \right|$$

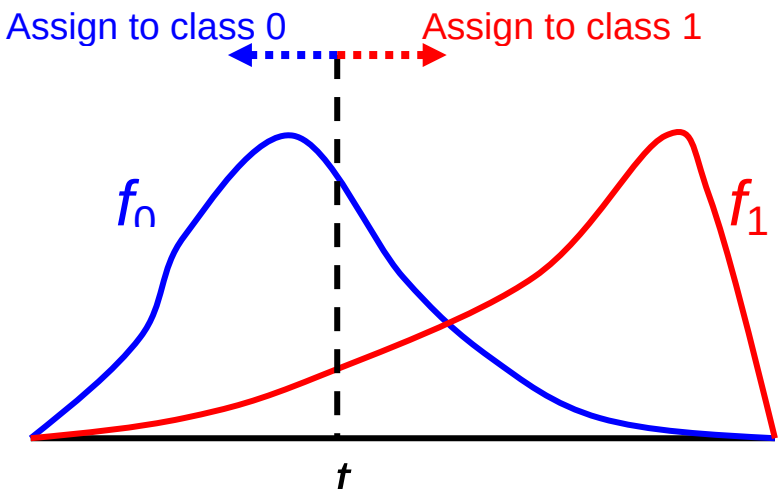
= *maximum absolute difference* between cdfs

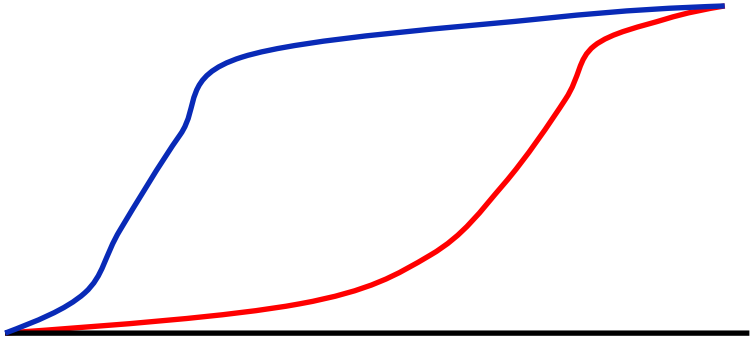
= *maximum vertical difference* between ROC curve and diagonal (assuming ROC curve is above diagonal)

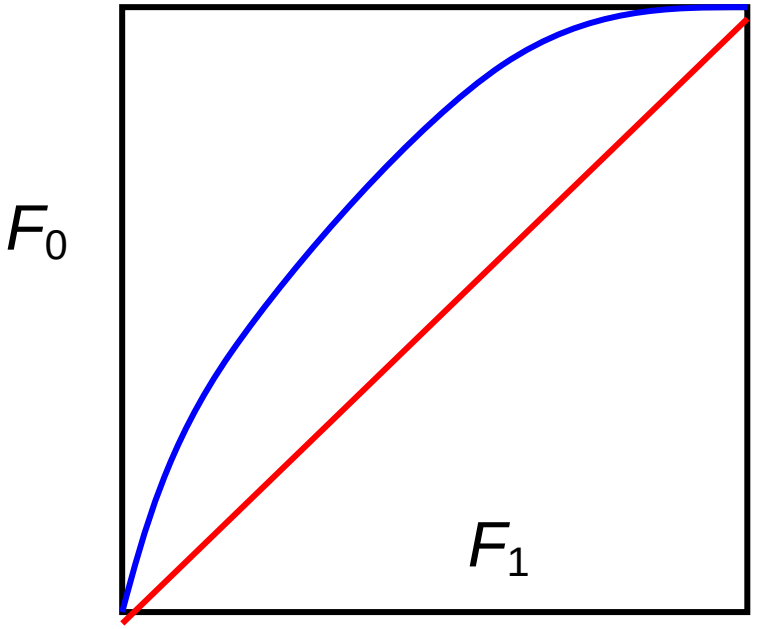
= *Youden statistic*

= 1 - "*Overlapping coefficient*"

= 1 - "*Proportion of similar responses*"


Assign to class 0
Assign to class 1
$f_0$
$f_1$
$t$




$F_0$
$F_1$

# History

**Kolmogorov (1933)** compared empirical score distribution $F_n(s)$ with hypothesised distribution $F_O(s)$

$$KS_1 = \sup_s \left| F_n(s) - F_O(s) \right|$$

Derived asymptotic distribution of $KS_1\sqrt{n}$

A.Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*,G. Inst. Ital.Attuari, 4 (1933), p83–91.

**Smirnov (1939)** simplified derivation and proposed *KS* for comparing two distributions and derived asymptotic distribution of $KS\sqrt{mn/(m+n)}$

N.V. Smirnov, *Estimate of derivation between distribution functions in two independent samples*, Bull. Mosc. Univ. 2 (1939), p3-16 (Russian).

***But this is a test to compare two cdfs***

*That is, a test to see if ROC curve is significantly different from the chance diagonal*

***We often want to compare ROC curves***

*That is, to know if one KS value is significantly different from another*

Note two different situations:

- scorecards are applied to different populations

- scorecards are applied to same population

Latter has induced correlation

(paired, matched, McNemar vs binomial,....)

# Test statistic

$$D = \left| KS^{(1)} - KS^{(2)} \right|$$

*Null case*: the true ROCs are the same

But: ***distribution of D will depend on ROC shapes***

So adopt robust binormal model:

$$f_G(s) = N\left(\mu_G, \sigma_G^2\right) \qquad f_B(s) = N\left(\mu_B, \sigma_B^2\right)$$

ROC has form: $\Phi^{-1}(y) = a + b\Phi^{-1}(x)$

with: $a = \left(\mu_G - \mu_B\right)/\sigma_B$ and $b = \sigma_G/\sigma_B$

This model is widely applicable:

- ROC curve is invariant to monotonic transformations
- undesirable behaviour normally restricted to very small parts of the curve

But even so: an exact analytic solution intractable and asymptotic results may not be useful

So use Monte Carlo simulation:

[wlg: $f_G(s) = N(0,1)$ $f_B(s) = N\left(a/b, 1/b^2\right)$]

# Case 1: independent samples
## (ROCs from different data)

Draw repeated samples:

$KS^{(1)}$ based on $n_1$ goods and $m_1$ bads

$KS^{(2)}$ based on $n_2$ goods and $m_2$ bads

With the distribution of goods being $N(0,1)$

and the distribution of bads being $N\left(a/b, 1/b^2\right)$

# Case 2: paired samples
## (ROCs from same data)

Repeatedly sample

$KS^{(1)}$ based on *n* goods and *m* bads

$KS^{(2)}$ based on the same *n* goods and *m* bads

*n* from bivariate normal

mean = $(0,0)$

sd = $(1,1)$

corr = *r*

*m* from bivariate normal

mean = $(a/b, a/b)$

sd = $(1/b, 1/b)$

corr = *r*

But *a* and *b* are unknown
So use *estimates*

For $KS^{(i)}$

From the $n_1$ scores in good class

$\bar{x}_{Gi}$ = sample mean

$s_{Gi}$ = sample s.d.

From the $m_1$ scores in bad class

$\bar{x}_{Bi}$ = sample mean

$s_{Bi}$ = sample s.d.

Then $\hat{a}_i = \dfrac{\overline{x}_{Bi} - \overline{x}_{Gi}}{s_{Bi}}$ and $\hat{b}_i = \dfrac{s_{Gi}}{s_{Bi}}$

Under null hypothesis the true values are the same

➔ use weighted sums

$$\hat{a} = \frac{(n_1 + m_1)\hat{a}_1 + (n_2 + m_2)\hat{a}_2}{n_1 + m_1 + n_2 + m_2}$$

$$\hat{b} = \frac{(n_1 + m_1)\hat{b}_1 + (n_2 + m_2)\hat{b}_2}{n_1 + m_1 + n_2 + m_2}$$

## Software in R available:

Krzanowski W.J. and Hand D.J. (2011) Testing the difference between two Kolmogorov-Smirnov values in the context of Receiver Operating Characteristic curves. *Journal of Applied Statistics*, **38**, 437-450.

To use specify

- sample sizes
- number of iterations
- *a* and *b* parameters
- *r* in the correlated case

# Example 1: Test to choose scorecard

1914 loan customers
- 1648 goods
- 266 bads (made no repayments)

Three scorecards
- logistic regression
- linear discriminant analysis
- support vector machine

Paired tests needed

# Summary statistics

| Statistic | LR | LDA | SVM |
|---|---|---|---|
| $\bar{x}_{Ni}$ | 0.8703 | 0.8785 | 0.8627 |
| $\bar{x}_{Pi}$ | 0.8042 | 0.7938 | 0.8517 |
| $s_{Ni}$ | 0.0879 | 0.1043 | 0.0354 |
| $s_{Pi}$ | 0.1234 | 0.1549 | 0.0620 |
| $\hat{a}_i$ | −0.5355 | −0.5471 | −0.1778 |
| $\hat{b}_i$ | 0.7121 | 0.6736 | 0.5714 |
| $K_2^{(i)}$ | 0.3422 | 0.3575 | 0.2180 |

# Pairwise scorecard comparisons

| Statistic | LR *vs* LDA | LR *vs* SVM | LDA *vs* SVM |
|---|---|---|---|
| Correlation | 0.9826 | 0.4838 | 0.4920 |
| $\hat{a}$ | −0.5413 | −0.3567 | −0.3625 |
| $\hat{b}$ | 0.6928 | 0.6418 | 0.6225 |
| $D = \lvert K_2^{(1)} - K_2^{(2)} \rvert$ | 0.0153 | 0.1242 | 0.1395 |
| 10% point | 0.0201 | 0.0562 | 0.0550 |
| 5% point | 0.0238 | 0.0671 | 0.0662 |
| 1% point | 0.0319 | 0.0868 | 0.0849 |

# Example 2: Test scorecard deterioration

UPL, outcome: at least 3 months in arrears

Training set

45,934 = 42012 goods + 3922 defaulters

Select ***first*** 500 goods and 500 bads

$$KS^{(1)} = 0.378$$

Test set

46,315 new customers

= 42063 goods + 4252 defaulters

Select ***last*** 500 goods and 500 bads

$$KS^{(2)} = 0.304$$

$$D = \left| KS^{(1)} - KS^{(2)} \right| = 0.074$$

***Suggesting fairly substantial deterioration***

But could it be chance?

Paired test needed

$\hat{a} = 0.5271 \quad \hat{b} = 0.8702$

| | |
|---|---|
| 10% | 0.064 |
| 5% | 0.076 |
| 1% | 0.100 |

➔ ***scorecard should be reviewed***

## Extension to marginal analysis

What if have only KS statistics and not raw scores

Can do a *marginal* analysis:

average over all plausible *a* and *b* values

The software also does this

# Conclusions

Evaluating scorecards is not enough

Need also to know if any difference is *real*
and not merely due to chance

Need a *significance test*

This talk has described such a test for the *KS* statistic and introduced easy-to-use software

Krzanowski W.J. and Hand D.J. (2011) Testing the difference between two Kolmogorov-Smirnov values in the context of Receiver Operating Characteristic curves. *Journal of Applied Statistics*, **38**, 437-450.

Krzanowski W.J. and Hand D.J. (2009) *ROC curves for continuous data*. Chapman and Hall.

# *thank you !*

***d.j.hand@imperial.ac.uk***