Measuring associations and evaluating forecasts of categorical and discrete variables

Andrei Sirchenko (Nyenrode Business University)
Jochem Huismans (University of Amsterdam)
Jan Willem Nijenhuis (Nedap NV)

July 20, 2023

To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it.

— K. A. C. Manderville, The Undoing of Lamia Gurdleneck

Measuring associations and correlations

Contingency table

	Male	Female	Total
Blonde	8	16	24
Brunette	14	18	32
Total	22	34	56

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Pearson correlation (Yule φ) coefficient: $\frac{n_{11} n_{22} - n_{12} n_{21}}{\sqrt{n_{1+} n_{+1} n_{+2} n_{2+}}} = -0.11$

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Pearson correlation (Yule φ) coefficient: $\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{+1}n_{+2}n_{2+}}} = -0.11$
- Michelet coefficient: $\frac{n_{11}^2}{n_{12}n_{21}}$

Measuring associations and correlations

Contingency table

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Pearson correlation (Yule φ) coefficient: $\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{+1}n_{+2}n_{2+}}} = -0.11$
- Michelet coefficient: $\frac{n_{11}^2}{n_{12}n_{21}}$
- Roux coefficient #1: $\frac{n_{11}+n_{22}}{\min(n_{12},n_{21})+\min(n-n_{12},n-n_{21})}$

Measuring associations and correlations

Contingency table

	Male	Female	Total
Blonde	$n_{11} = 8$	$n_{12} = 16$	$n_{1+} = 24$
Brunette	$n_{21} = 14$	$n_{22} = 18$	$n_{2+} = 32$
Total	$n_{+1} = 22$	$n_{+2} = 34$	$n = 56$

- Pearson correlation (Yule φ) coefficient: $\frac{n_{11} n_{22} - n_{12} n_{21}}{\sqrt{n_{1+} n_{+1} n_{+2} n_{2+}}} = -0.11$
- Michelet coefficient: $\frac{n_{11}^2}{n_{12} n_{21}}$
- Roux coefficient #1: $\frac{n_{11} + n_{22}}{\min(n_{12}, n_{21}) + \min(n - n_{12}, n - n_{21})}$
- Roux coefficient #2: $\frac{n - n_{11} n_{22}}{\sqrt{n_{1+} n_{+1} n_{+2} n_{2+}}}$

Evaluating categorical forecasts

Confusion matrix

		Actual values	
		Positive	Negative
Predicted values	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- Accuracy $= \frac{TP+TN}{n}$
- Hit rate $= \frac{TP}{TP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- Specificity $= \frac{TN}{FP+TN}$

Evaluating probabilistic forecasts

Diagnostic probability scores

- Brier score:

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\Pr\left(y_i = k\right) - \delta_{ik}\right)^2$$

- Spherical score:

$$1-\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{K}\delta_{ik}\Pr\left(y_i = k\right)}{\sqrt{\sum_{k=1}^{K}\left[\Pr(y_i = k)\right]^2}}$$

- Ranked probability score:

$$\frac{1}{n(K-1)}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left(\sum_{j=1}^{k}\Pr\left(y_i = j\right)-\sum_{j=1}^{k}\delta_{ij}\right)^2$$

Literature on measures of association

is poorly integrated across different fields

- a wide variety of scalar statistics have been developed and used in different fields
- a similarly wide variety of nomenclature has appeared in relation to these statistics
- some of these measures have been reinvented, duplicated and renamed on multiple occasions in other fields
- confusing terminology is confounded further by different notation

Literature on measures of association

is poorly integrated across different fields

- Cohen kappa coefficient (1960): $\frac{2(n_{11}n_{22}-n_{12}n_{21})}{n_{+1}n_{2+}+n_{1+}n_{+2}}$
- Heidke skill score (1926)
- Doolittle association ratio (1887)
- Galton coefficient (1892)
- Hubert–Arabie adjusted Rand index (Hubert and Arabie 1985)

Accuracy
Alternative terminology

- Accuracy
- Agreement rate
- Causal support
- Classification rate
- Count R^2
- Hit score
- Holsti $C.R.$ coefficient
- Kendall coefficient
- Osgood coefficient
- Proportion correct
- Rand coefficien
- Ratio test discriminant
- Simple matching coefficient
- Sokal-Michener coefficient

A catalog of probabilistic forecast evaluation metrics

1. Brier score, half-Brier score, probability score, quadratic score (Brier 1950; Toda 1963):

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\Pr(y_i = k) - \delta_{ik}\right)^2$$

2. Logarithmic score, ignorance score (Good 1952; Toda 1963; Winkler and Murphy 1968; Roulston and Smith 2002):

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\log(\Pr(y_i = k)) + (1-\delta_{ik})\log(1-\Pr(y_i = k))$$

3. Power score ($\beta > 1$, identical to the quadratic score at $\beta = 2$; Selten 1998):

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{\beta} - \sum_{k=1}^{K}\delta_{ik}\Pr(y_i = k)^{\beta-1} + \frac{\beta-1}{\beta}\sum_{k=1}^{K}\left[\Pr(y_i = k)\right]^{\beta}\right\}$$

4. Pseudospherical score ($\beta > 1$; identical to the spherical score at $\beta = 2$; Good 1971):

$$1-\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{K}\delta_{ik}\left[\Pr(y_i = k)\right]^{\beta-1}}{\left[\sum_{k=1}^{K}\left[\Pr(y_i = k)\right]^{\beta}\right]^{(\beta-1)/\beta}}$$

5. Ranked probability score (suitable only for ordinal variables; identical to the Brier score for binary variables; Epstein 1969; Murphy 1971):

$$\frac{1}{n(K-1)}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left(\sum_{j=1}^{k}\Pr(y_i = j) - \sum_{j=1}^{k}\delta_{ij}\right)^2$$

6. Spherical score (Toda 1963; Winkler 1967; Winkler and Murphy 1968; Friedman 1983):

$$1-\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{K}\delta_{ik}\Pr(y_i = k)}{\sqrt{\sum_{k=1}^{K}\left[\Pr(y_i = k)\right]^2}}$$

7. Two-alternative forced choice (2AFC) score #1 (Mason and Weigel 2009):

$$1-\frac{\sum_{l=1}^{K}\sum_{k\neq l}^{K}\sum_{i=1}^{n_{+k}}\sum_{j=1}^{n_{+l}}I\left[p_{k,i}(l), p_{l,j}(l)\right]}{\sum_{l=1}^{K}\sum_{k\neq l}^{K}n_{+k}n_{+l}},$$

where $p_{k,i}(l)$ is the forecast probability of category l for observation i in category k; and

$$I\left[p_{k,i}(l), p_{l,j}(l)\right] = \begin{cases} 0 & \text{if} \quad p_{l,j}(l) < p_{k,i}(l) \\ 0.5 & \text{if} \quad p_{l,j}(l) = p_{k,i}(l) \\ 1 & \text{if} \quad p_{l,j}(l) > p_{k,i}(l) \end{cases}$$

A catalog of association & correlation metrics

1. Accuracy, agreement rate, causal support, classification rate, count R^2, hit score, Holsti $C.R.$, Kendall, Osgood, proportion correct, Rand, ratio test discriminant, simple matching coefficient, Sokal-Michener (CTS; Finley 1884; Klein 1985; Zubin 1938; Sokal and Michener 1958; Osgood 1959; Holsti 1969; Rand 1971; Maddala 1992; Kodratoff 2001): $\frac{1}{n}\sum_{k=1}^{K} n_{kk}$
2. Added value, centered confidence, change of support (AS; Sahar and Mansour 1999; Tan et al. 2004; Geng and Hamilton 2007; Lallich et al. 2007): $\frac{n_{11}}{n_{1+}} - \frac{n_{+1}}{n}$
3. Adjusted noise-to-signal ratio (AS; Kaminsky et al. 1997; Kaminsky and Reinhart 1999): $\frac{n_{12}n_{+1}}{n_{+2}n_{11}}$
4. Alroy, corrected Forbes F (TS; Alroy 2015): $\frac{n_{11}(n+\sqrt{n})}{n_{11}(n+\sqrt{n})+\frac{3}{2}n_{12}n_{21}}$
5. Analyzing method patterns to locate errors (AMPLE) (CS; Dallmeier et al. 2005): $\left|\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right|$
6. Anderberg (TS; Anderberg 1973): $\frac{8n_{11}}{8n_{11}+n_{12}+n_{21}}$
7. Anderberg D (CTS; Anderberg 1973): $\frac{1}{2n}[max(n_{11}, n_{12}) + max(n_{21}, n_{22}) + max(n_{11}, n_{21}) + max(n_{12}, n_{22})$ - $max(n_{+1}, n_{+2})$ - $max(n_{1+}, n_{2+})]$
8. Appleman (CS; Appleman 1960): $\frac{n_{22}-n_{21}}{n_{12}+n_{22}}$ if $n_{11}+n_{21} > n_{12}+n_{22}$; $\frac{n_{11}-n_{12}}{n_{21}+n_{11}}$ if $n_{11}+n_{21} < n_{12}+n_{22}$
9. Atkinson (CTS; Atkinson 1970): $1-\left(\prod_{i=1}^{K}\prod_{j=1}^{K}\frac{n_{ij}}{n}K^2\right)^{1/K^2}$
10. Austin–Colwell (CTS; Goodall 1967; Austin-Colwell 1977): $\frac{2}{\pi}\arcsin\sqrt{\frac{1}{n}\sum_{k=1}^{K} n_{kk}}$
11. Balanced accuracy, balanced classification rate (CS; Brodersen et al. 2010; Urbanowicz and Moore 2015): $\frac{1}{K}\sum_{k=1}^{K}\frac{n_{kk}}{n_{+k}}$
12. Baroni-Urbani–Buser #1 (TS; Baroni-Urbani and Buser 1976): $\frac{\sqrt{n_{11}n_{22}}+n_{11}}{\sqrt{n_{11}n_{22}}+n_{11}+n_{12}+n_{21}}$
13. Baroni-Urbani–Buser #2 (TS; Baroni-Urbani and Buser 1976): $\frac{\sqrt{n_{11}n_{22}}+n_{11}-(n_{12}+n_{21})}{\sqrt{n_{11}n_{22}}+n_{11}+n_{12}+n_{21}}$

- a contingency table (confusion matrix)

A new Stata command classify

Input

- a contingency table (confusion matrix)
- the observed values of a categorical (discrete) variable and the predicted probabilities of each category

A new Stata command classify

Input

- a contingency table (confusion matrix)
- the observed values of a categorical (discrete) variable and the predicted probabilities of each category
- the values of two categorical (or discrete numerical) variables

- a contingency table (confusion matrix)
- 214 measures of association and correlation and 9 diagnostic scores of the accuracy of probabilistic forecasts
- the class-specific measures for each class as well as their simple and weighted averages

A new Stata command classify

Output in Results window

```
. matrix Confusion = (30,9,0 \ 25,163,26 \ 0,9,17)

. classify, mat(Confusion)

Contingency Table

     Actual |      1       2       3
------------+------------------------
  Predicted |
          1 |     30       9       0
          2 |     25     163      26
          3 |      0       9      17


Measures of association and correlation
Accuracy                                        =    0.7527
Goodman-Kruskal Lambda                          =    0.0769
Goodman-Kruskal Lambda weighted                 =    0.0874
Goodman-Kruskal Lambda r                        =    0.2959
Heidke skill score                              =    0.4629
Peirce skill score                              =    0.4127

See the Excel file 'Classify Metrics.xls' for the complete output
```

A new Stata command classify

Output in Results window

```
. classify x2 y2

Contingency Table

        x2= |      1       0
------------+----------------
        y2= |
          1 |     58     127
          0 |     40      54


Measures of association and correlation
Accuracy                                     =    0.4014
Goodman-Kruskal lambda                       =    0.0000
Goodman-Kruskal lambda weighted              =    0.0000
Goodman-Kruskal Lambda_r                     =   -0.7041
Heidke skill score                           =   -0.0912
Peirce skill score                           =   -0.1098
Adjusted noise to signal ratio               =    1.1856
Bias                                         =    1.8878
F1                                           =    0.4099
Hit rate                                     =    0.5918
Odds ratio                                   =    0.6165
Precision                                    =    0.3135

See the Excel file 'Classify Metrics.xls' for the complete output
```

A new Stata command classify

Output in Results window

```
. classify x y

Contingency Table

        x= |    -1      0      1
-----------+---------------------
        y= |
        -1 |    38     17      0
         0 |    74     54     53
         1 |     0     23     20


Measures of association and correlation
Accuracy                                          =    0.4014
Goodman-Kruskal lambda                            =    0.0000
Goodman-Kruskal lambda weighted                   =    0.0000
Goodman-Kruskal Lambda_r                          =    0.0000
Heidke skill score                                =    0.0958
Peirce skill score                                =    0.0965

See the Excel file 'Classify Metrics.xls' for the complete output
```

A new Stata command classify

Output in Results window

```
. quietly oprobit y bias house gdp spread

. predict p1 p2 p3
(option pr assumed; predicted probabilities)

. classify y, probs(p1 p2 p3)

Confusion Matrix

     Actual |     -1       0       1
------------+-----------------------
  Predicted |
         -1 |     30       9       0
          0 |     25     163      26
          1 |      0       9      17


Diagnostic scores for probabilistic forecasts
Brier score               =     0.1679
Ranked probability score =     0.0847
Spherical score           =     0.1882

Measures of association and correlation
Accuracy                                         =     0.7527
Goodman-Kruskal lambda                           =     0.0769
Goodman-Kruskal lambda weighted                  =     0.0874
Goodman-Kruskal Lambda_r                         =     0.2959
Heidke skill score                               =     0.4629
Peirce skill score                               =     0.4127

See the Excel file 'Classify Metrics.xls' for the complete output
```

A new Stata command classify

Output in Excel file

No.	**Score name**	**Value**
1	Brier score	0.16788
2	Logarithmic score	1.06605
3	Power score (beta = 1.5)	0.13741
4	Pseudospherical score (beta = 1.5)	0.14317
5	Ranked probability score	0.08471
6	Spherical score	0.18818
7	Zero-one score	0.24731

A new Stata command classify

Output in Excel file

No.	Coefficient name	Symmetry	Value	Class-specific values: Class -1	Class 0	Class 1	Macro average	Weighted average
1	Accuracy	CTS	0.7527					
3	Adjusted noise to signal ratio	AS		0.0737	0.5779	0.0965	0.40428	0.24933
72	F1-score	TS		0.6383	0.8253	0.4928	0.73719	0.65212
73	F_beta-score (beta = 1.5)	AS		0.5991	0.8527	0.4501	0.74066	0.63397
74	Ganascia	AS		0.5385	0.5234	0.3077	0.49310	0.45651
76	Gilbert	TS		0.4688	0.7026	0.3269	0.59859	0.49942
77	Gilbert skill score	TS		0.3962	0.2594	0.2707	0.28812	0.30878
80	Gini #2	CS		0.5053	0.3801	0.3572	0.40128	0.41421
81	Gini #3	CTS		0.1257	0.0659	0.0664	0.07774	0.08598
82	G-mean	CS		0.7236	0.6572	0.6167	0.66403	0.66580
86	Goodman-Kruskal lambda	TS	0.0769					
87	Goodman-Kruskal lambda weighted	CS	0.0874					
88	Goodman-Kruskal lambda_r	CS	0.2959					
89	Goodman-Kruskal tau	CS		0.336	0.1843	0.1969	0.21613	0.23906
90	Goodman-Kruskal #1	CTS		0.2766	0.1534	-0.014	0.15179	0.13849
91	Goodman-Kruskal #2	CTS		0.2766	0.1534	-0.014	0.15179	0.13849
92	Goodman-Kruskal #3	CTS		0.2766	0.1779	0.1159	0.18782	0.19015
101	Heidke skill score	CTS	0.4629					
102	Hit rate	AS		0.5455	0.9006	0.3953	0.75269	0.61379
144	Odds ratio	CTS		28.667	8.3453	16.491	13.60681	17.83448
150	Peirce skill score	CS	0.4127					
157	Precision	AS		0.7692	0.7617	0.6538	0.74655	0.72825

Binary confusion matrix

	$x = 1$	$x = 0$
$y = 1$	TP	FP
$y = 0$	FN	TN

- Accuracy $= \frac{TP+TN}{n}$
- Hit rate $= \frac{TP}{TP+FN}$
- Specificity $= \frac{TN}{FP+TN}$

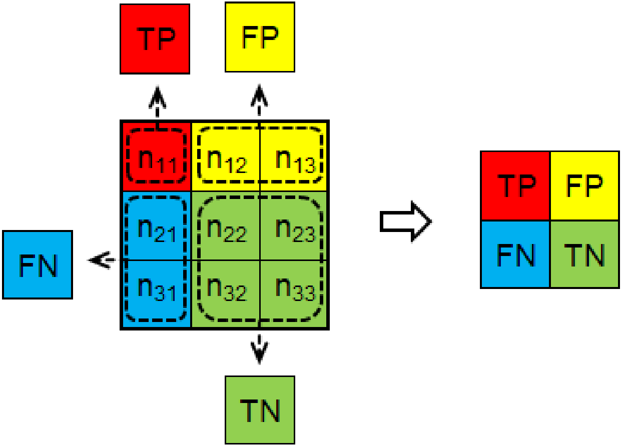
Multy-class confusion matrix

	$x = 1$	...	$x = K$
$y = 1$	n_{11}	...	n_{1K}
...	...	...	...
$y = K$	n_{K1}	...	n_{KK}

$$Accuracy = \frac{1}{n}\sum_{k=1}^{K} n_{kk}$$

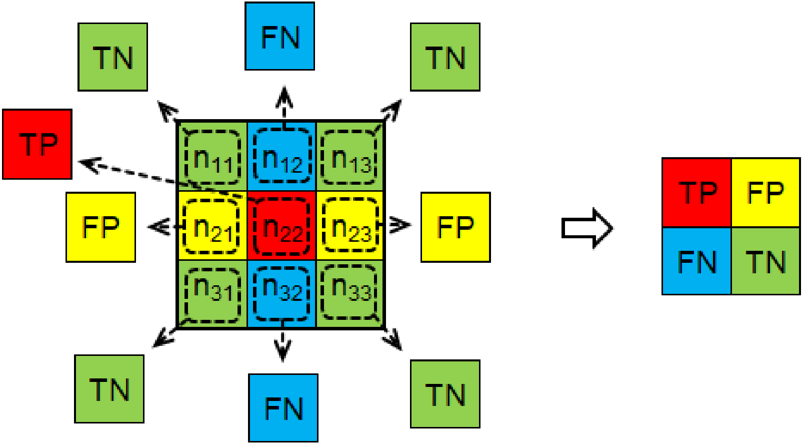
Class-specific measures

Class 1



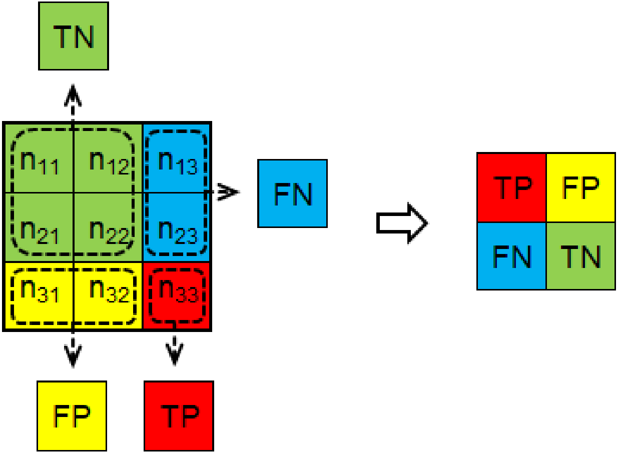
Class-specific measures

Class 2



Class-specific measures

Class 3



Class-specific measures

Arithmetic and weighted averages

- The `classify` command also computes the simple arithmetic and weighted arithmetic averages of all class-specific measures as:

$$
\begin{aligned}
Measure_{macro} &= \frac{1}{K}\sum_{k=1}^{K} Measure_k \\
Measure_{weighted} &= \sum_{k=1}^{K} Measure_k \frac{n_{+k}}{n}
\end{aligned}
$$

- The macro-averaged measures calculate unweighted (arithmetic) mean of class-specific coefficients.
- The weighted-averaged measures take a weighted mean. The weights for each class are the total number of observations of that class.

Stata command classify

Symmetric measures: two types of symmetry

- A measure is transpose symmetric if it treats both variables equivalently, and so it is invariant to relabelling of them — it remains unchanged if the row variable and column variable are interchanged.

- A measure is complement symmetric if it treats all categories equivalently, and so it is invariant to relabelling of them — it remain unchanged if any two columns and the corresponding two rows are swapped.

Classify and be happy

" . . . there is no absolutely general measure of the degree of dependence. Every attempt to measure a conception like this by a single number must necessarily contain a certain amount of arbitrariness and suffer from certain inconveniences."

— Cramér (1924)