Measures Of Population Stability and Instability

Tom, Daniel

1 January 2022

Measures Of Population Stability and Instability

Daniel Tom, Ph.D.
https://orcid.org/0000-0003-4853-2498

DTom Computing
https://www.linkedin.com/in/DanielTom

January 1, 2022
July 4, 2023 (rev.)

Abstract

A popularly used PSI treats all variables as categorical, regardless of bin ordering. Also, bin boundaries and the number of bins could affect the PSI quantity. We build our PSI without requiring binning, distinguishes numeric shifts from categorical redistribution, and unify the two for mixed numeric/categorical variables.

Introduction

In modeling and analytics we are concerned with the stability of the data. Data instability may impact the generalization of a model or conclusion or inference in the presence of unforeseen data and distribution. A population stability index (PSI) helps quantify the stability of a data attribute or variable.

One PSI in popular use is based on a symmetrized version of the information theoretic Kullback-Leibler (1951) divergence, nowadays called the Jeffreys (1948) divergence (JD). This divergence is defined as a continuous integral for ideal analytic distribution functions. For discrete data samples, there is a corresponding summation over contiguous ranges of an attribute. Binning those ranges is a chore, and having a different number of bins could produce substantially varying JD PSI which does not help to quantify stability.

Following our success to find a replacement for an information value (IV) measure of separation which is based on the same divergence measure, we would like to find a replacement for JD PSI. In addition to avoiding binning, we would also like our PSI to have a few other desirable characteristics. In particular, we would like our PSI to distinguish numeric shifts from redistribution in categories. The discrete form of JD PSI essentially treats all binned ranges as categories irrespective of any ordering.

In this article, we revisit our search for a separation measure to replace IV. We compare and contrast IV with other measures of separation, note the advantages and disadvantages, and make our choice. We take the learnings from this exercise in searching for a replacement for JD PSI. We list desirable qualities, and embark on reviewing a number of measures and metrics. We build our PSI for numeric variables separate from categorical variables, and show a simple shift test and an overlap test. We create a mechanism to link the quantities for mixed variables (e.g., numeric with missing).

Distance and Divergence

Given two numbers we can get the distance between them by subtraction. Given two random variables we can compute a statistical distance between their probability distributions. The Kullback-Leibler (KL) divergence is an information theoretic statistical distance (logarithmic difference) between two analytic (idealized) probability distribution functions $p(x)$ and $q(x)$:

$$\int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx \qquad (1)$$

Jeffreys divergence (JD) is a symmetrized version of the information theoretic Kullback-Leibler divergence:

$$\int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx + \int_{-\infty}^{\infty} q(x) \ln\left(\frac{q(x)}{p(x)}\right) dx \qquad (2)$$

$$= \int_{-\infty}^{\infty} \left(p(x) - q(x)\right) \ln\left(\frac{p(x)}{q(x)}\right) dx \qquad (3)$$

As our data come in samples, the integral in JD is replaced by a discrete summation:

$$\sum_{k} \left(p_k - q_k\right) \ln\left(\frac{p_k}{q_k}\right) \qquad (4)$$

Here p_k and q_k are the densities of the two probabilities. JD PSI is a measure of instability of the distribution, where p_k and q_k are the densities at two different points in time. With the same formulation, information value (IV) is a measure of separation of two distributions, where p_k and q_k are their respective densities. As summation is commutative, JD is irrespective of numeric rank ordering, but otherwise is fine with categorical variables.

Frequently either p_k or q_k could be zero, making the quotient of the log term and therefore JD undefined. A dubious practice zeros out the term so it would not ruin the whole summation. Another fix utilizes contiguous ranges, making the bins fewer and wider until the numerator and the denominator are both non-zero. It may work with one data sample, but could break again with another sample. Neither remedy is satisfactory. We know the number of bins affect the JD quantity – collapsing to one single bin would make JD exactly zero. We learned that other metrics, e.g., the Hosmer and Lemeshow (HL) statistic, could also be impacted by binning. Recent software that allows changing the ranges and/or the number of bins shows highly varying HL statistics as a result.

Prior Research

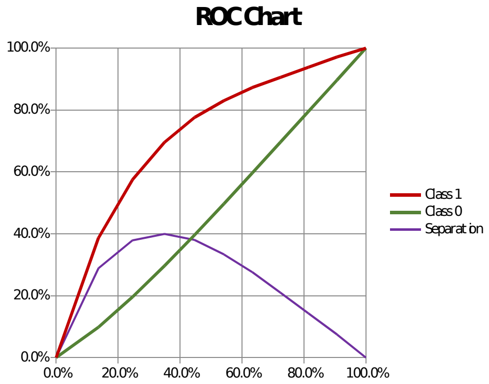
Due to the challenges of computing JD, we have been seeking a replacement for IV as a measure of separation for a modeling attribute. Often, decile ranking procedures fail to produce ten bins as many of our attributes are highly skewed. Manual binning is a chore. Fundamentally binning is a noise generating irreversible quantization. We want our measure of separation to require no binning, and be sensitive to rank ordering of numerical modeling attributes.

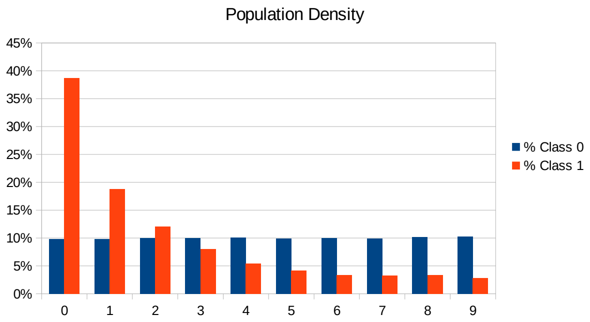
Our breakthrough comes when we encounter an existing score among the modeling attributes. We can incorporate an existing score in a new model. Typically, for a score we compute the KS (Kolmogorov-Smirnov) statistic as a measure of separation. When a score is used as a modeling attribute, there is no reason why we cannot compute IV. Score or attribute, they are just numeric variables. If we can compute IV, we can compute KS for an attribute too. In the same vein, we can also compute for an attribute or a score the Gini coefficient (GC, same as Somers' D) often used in classification models.

When we compare IV, KS, and GC, we find desirable properties in some but not others. KS fairs slightly better than IV in that no binning is required. However, it is insensitive to rank ordering on either side of the maximum separation, so it is a poor candidate to replace IV. In contrast, the GC is sensitive to rank ordering everywhere, and requires no binning. GC handles numeric data well and can even handle categorical variables as ranked log odds. GC can be readily obtained from the CDFs (cumulative distribution functions) in a ROC (Receiver Operating Characteristics) chart, along with other association statistics like concordance, discordance, ties, and the c statistic. GC is now our preferred measure of separation.

Examples below illustrate the similarities and differences among IV, KS, and GC. Our first example shown in Exhibit A looks like a fairly good attribute with IV=0.860, KS=0.399, and GC=0.495:

Exhibit A

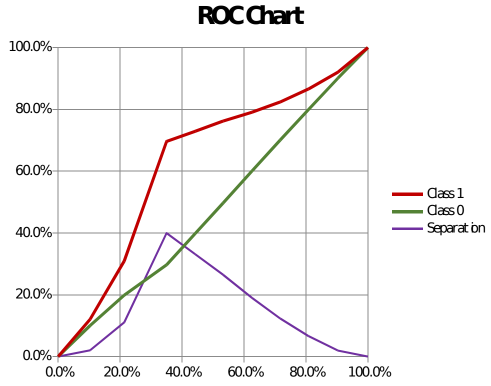


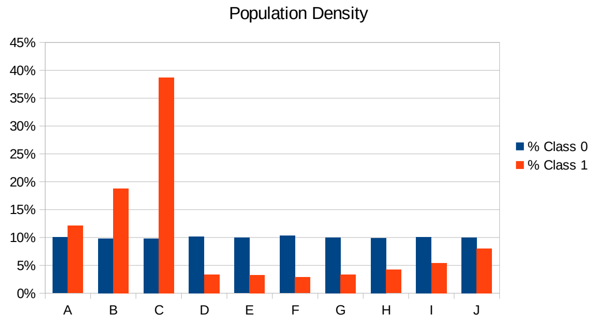


Attribute	% Class 0	% Class 1
0	9.8%	38.7%
1	9.8%	18.8%
2	10.1%	12.1%
3	10.0%	8.0%
4	10.1%	5.4%
5	9.9%	4.2%
6	10.0%	3.3%
7	9.9%	3.3%
8	10.1%	3.3%
9	10.3%	2.9%
All	100.0%	100.0%

The second example in Exhibit B has IV=0.860, KS=0.399, and GC=0.306.

Exhibit B



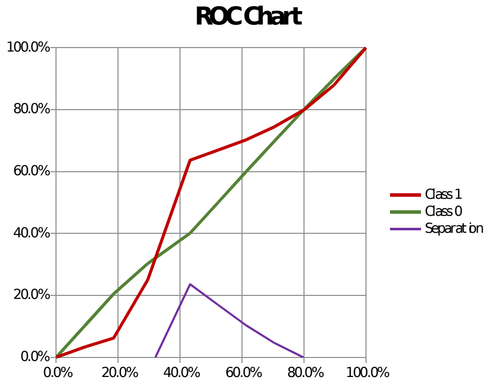


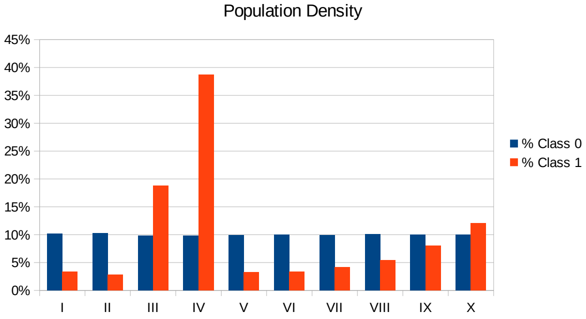
Attribute	% Class 0	% Class 1
A	10.1%	12.1%
B	9.8%	18.8%
C	9.8%	38.7%
D	10.1%	3.3%
E	9.9%	3.3%
F	10.3%	2.9%
G	10.0%	3.3%
H	9.9%	4.2%
I	10.1%	5.4%
J	10.0%	8.0%
All	100.0%	100.0%

The attribute in the second example does not have as fat a separation in the ROC chart as the first. GC shows decrease, yet IV and KS remain the same.

The third example in Exhibit C has IV=0.860, KS=0.236, and GC=0.053.

Exhibit C





Attribute	% Class 0	% Class 1
I	10.1%	3.3%
II	10.3%	2.9%
III	9.8%	18.8%
IV	9.8%	38.7%
V	9.9%	3.3%
VI	10.0%	3.3%
VII	9.9%	4.2%
VIII	10.1%	5.4%
IX	10.0%	8.0%
X	10.1%	12.1%
All	100.0%	100.0%

The attribute in the third example looks even worse in the ROC chart. However, IV remains the same. KS now shows a decrease. GC decreases further.

The reader will notice the second and third examples are just row-permutations of the first. IV staying the same means it is insensitive to any rank ordering. Between the first two examples, KS unchanging means it is insensitive to rank ordering on either side of the maximum separation. GC is the separation measure that is sensitive to rank ordering everywhere. This is the reason we choose GC and not IV or even KS.

Design of a PSI

Following our prior success to replace IV with GC, we now turn our focus onto designing a PSI that does not use JD or KL divergence. We list desirable properties, and search the literature for existing measures and metrics. If nothing meets our requirements exactly, we adapt, modify, and/or adjust until we have a PSI that suits our needs.

Here are some desirable properties to have with our PSI: We like its computation to require no binning. This also means it needs to work well with skewed distributions, and zeros densities which may often occur with smaller data samples or out-of-time data. IV has no upper bound, but having one would be useful so that it may be rescaled or normalized. Along that line we would need ranges that are comparable to the conventional definition of minor/medium/major instability. We want the PSI to indicate redistribution in a categorical variable to be distinct from a shift in a numeric variable. For a mixed ordinal/categorical variable (e.g., numeric with missing) we need to be able to unify the shift and redistribution concepts and generate a combined measure of population stability.

Here are a few of the distance measures we researched (in no particular order): Kullback-Leibler divergence, Jeffreys divergence, Fisher information metric, Hellinger distance, energy distance, earth movers distance, Wasserstein metric, total variation distance, Levy metric, Bhattacharyya distance, Mahalanobis distance, Minkowski distance, Kolmogorov-Smirnov statistic, Cramér-von Mises criterion, Anderson-Darling test, Wald test, chi-squared tests, Gini coefficient (Somers' D), mean absolute difference, Jensen-Shannon divergence.

A number of the above statistical distance measures are either Bregman or f-divergences, or both in the case of KL divergence. Notably the squared Euclidean distance is a Bregman distance but not an f-divergence. Quite a few of them are cousins of KL and Jeffreys divergence, requiring a logarithm of a ratio so we have to be careful. Earth movers and energy distances have a physical analog to shifts in a numeric variable, which may be a useful concept to borrow.

PSI for a Categorical Variable

We choose a scaled version of the Jensen-Shannon (JS) divergence (Lin, 1991) as the PSI for a categorical variable. JS divergence is the average of the KL divergences between distributions p and m and between q and m, where the mixture m is the average of p and q. In discrete summation form, JS divergence is:

$$\frac{1}{2}\sum_k p_k \ln\left(\frac{p_k}{m_k}\right)+\frac{1}{2}\sum_k q_k \ln\left(\frac{q_k}{m_k}\right) \tag{5}$$

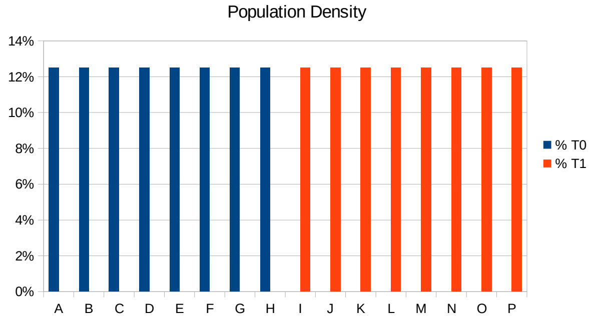
$$m_k=\frac{1}{2}\left(p_k+q_k\right) \tag{6}$$

In (5), the denominator of the quotient could only be zero if both densities p_k and q_k are zero. When the numerator p_k (or q_k) is zero, the multiplier outside the logarithm p_k (or q_k) ensures the term is zero. Therefore, JS divergence is well-defined for data samples that may have zero densities.

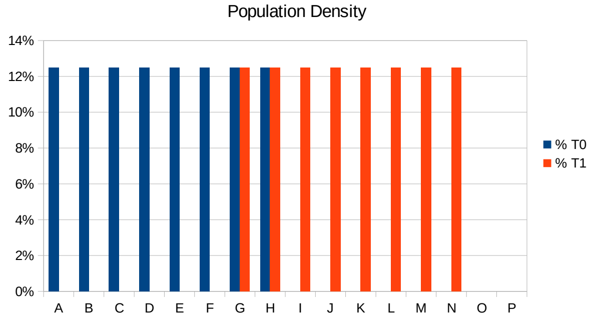
Unlike JD, JS divergence is bounded from below and above. JS divergence is zero when $p = q$. When p and q are mutually exclusive, $m=1/2$, and JS divergence attains the maximum value of $\ln(2)$. To normalize, we define JS PSI to be JS divergence divided by $\ln(2)$, so that 0 < JS PSI < 100%. For small values, we observe JS divergence ≈ JD PSI/8.

We display below a few overlap tests with a uniformly distributed categorical variable. Different degrees of overlap between two time points (e.g., T0, T1) give different JS PSI values. Full overlap means the distribution is stable, in which case JS PSI should be 0. No overlap means completely unstable – JS PSI should be 100%. These are the boundary cases. The categories are not ordered, so there could be different permutations of a partial overlap with the same JS PSI.

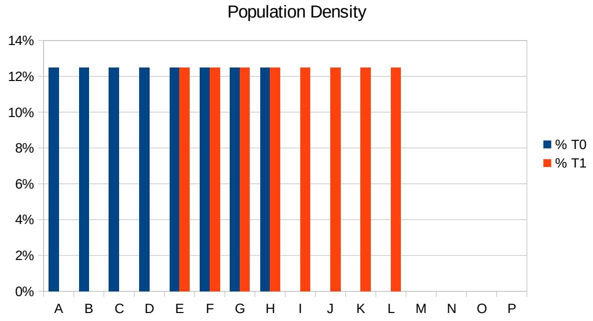
(a) No categories overlap: JS PSI = 100%



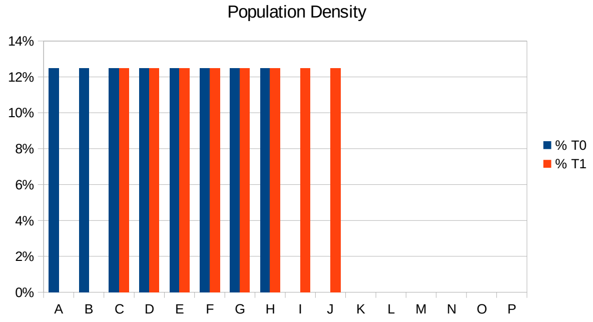
(b) 2 of 8 categories overlap: JS PSI = 75%



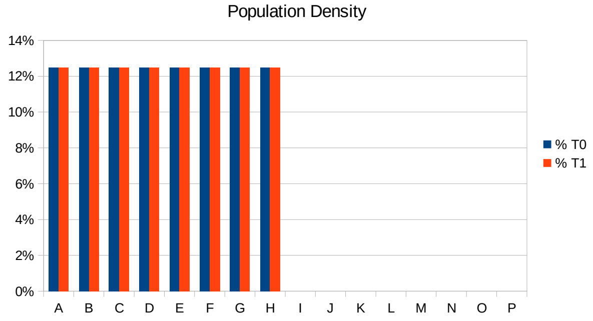
(c) 4 of 8 categories overlap: JS PSI = 50%



(d) 6 of 8 categories overlap: JS PSI = 25%



(e) All categories overlap: JS PSI = 0%



PSI for a Numeric Variable

There are a number of inspirations behind our definition of PSI for a numeric variable. First, we follow our success using GC to replace IV. GC is based on the ROC. GC requires no binning, and so it is good for numeric variables. GC measures the area under ROC curve, but could be neutralized or negated by negative area of flipped CDFs. We want to measure upshifts and downshifts. The Cramér-von Mises (CVM) criterion, an alternative to the KS statistic, uses the squared distance between the ROC curves (Anderson, 1962). CVM is not impacted by flipped CDFs, but the squaring skews it for larger differences. The earth mover's distance or the first Wasserstein (Vaserstein, 1969) distance (W_1) has a physical analog of shifting a pile of earth shaped like one distribution to the other with minimum energy. W_1 uses the absolute distance between the ROC curves which may suit our need. We also need our PSI for numeric variable to be symmetric, therefore we employ the mid-CDF (Lancaster, 1961) popularized by Parzen (2009). The mid-CDF of a distribution with cumulative distribution $P(x)$ and density $p(x)$ is $P(x) - p(x)/2$.

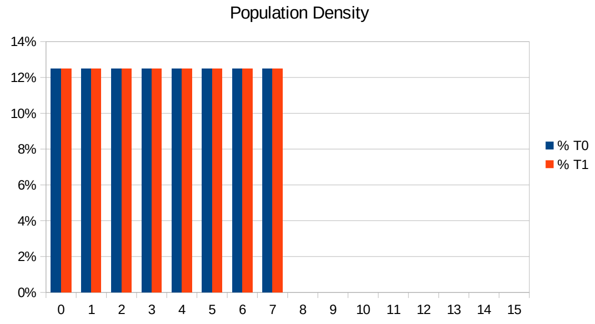
Our definition of PSI for numeric data is Absolute Area Between mid-CDFs (AABC PSI):

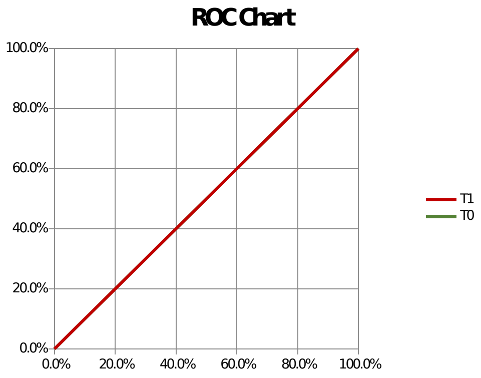
$$2\sum_k m_k \left|\mathring{P}_k - \mathring{Q}_k\right| \tag{7}$$

Here $\mathring{P}$ and $\mathring{Q}$ are mid-CDFs of the numeric variable's distributions at two time points. The mixture density m_k is the same as (6). Already scaled, we will see that AABC PSI is bounded by 0 and 100%.

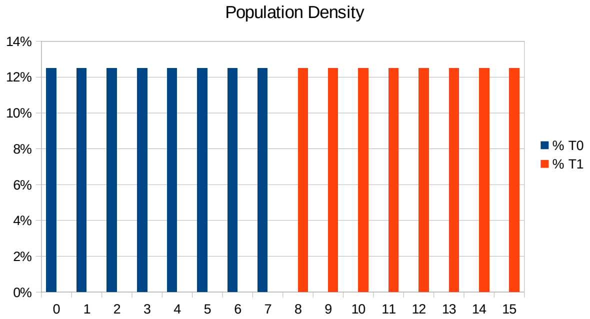
Shift tests are shown below with different degrees of shift. With no shift, the distributions are identical (i), and AABC PSI = 0. When all shifted out (ii), we get AABC PSI = 100%. These are the boundary cases. We show three others (iii), (iv), and (v) with AABC PSI at 75%, 50%, and 25%. Note that (iii) may look like (c) for JS PSI shown above, however JS PSI is for a categorical variable. If we have (iii), (iv), and (v) as categorical distributions, JS PSI would have the same 50% value.

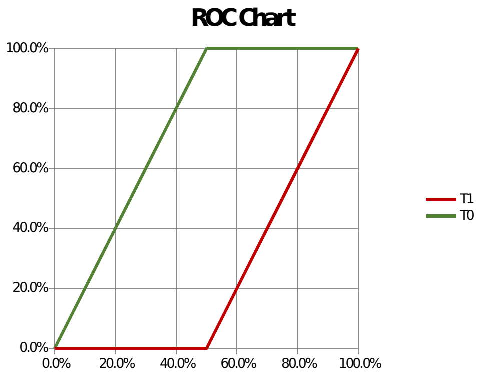
(i) AABC PSI = 0%



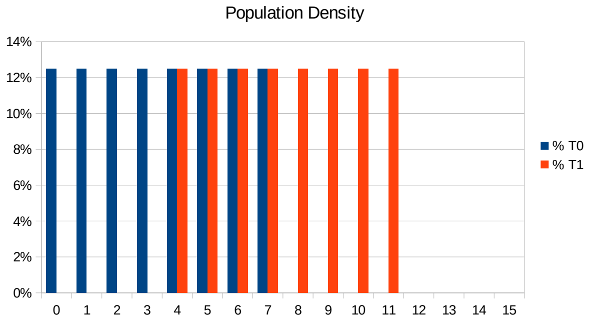


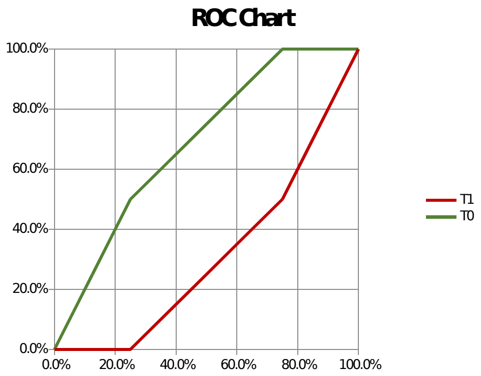
(ii) AABC PSI = 100%



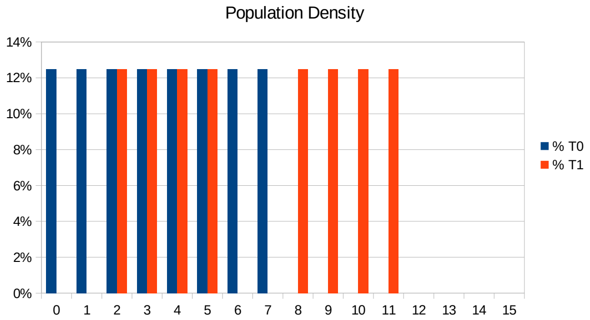


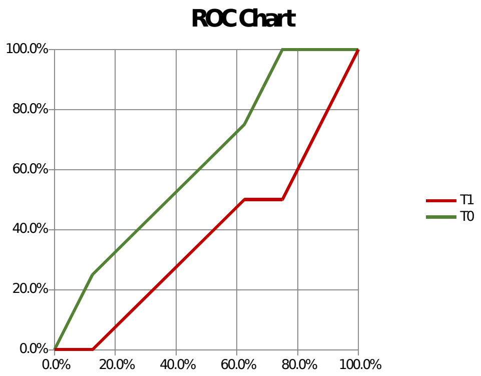
(iii) AABC PSI = 75%



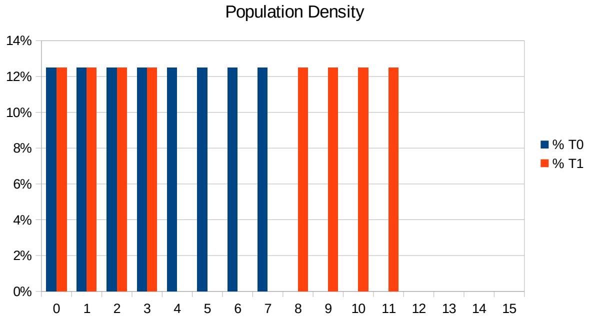


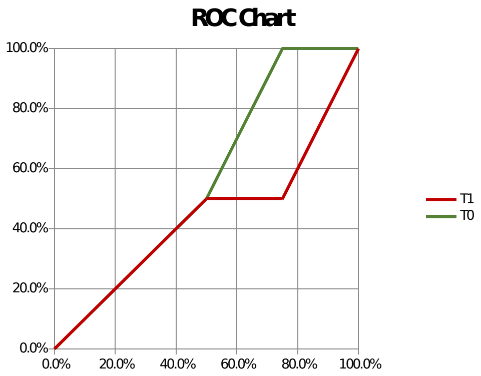
(iv) AABC PSI = 50%





(v) AABC PSI = 25%





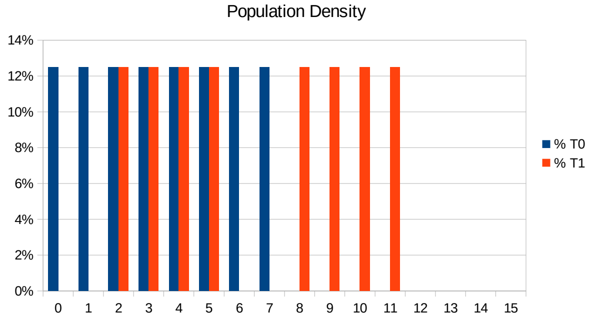
PSI for a Mixed Ordinal/Categorical Variable

Mixed ordinal/categorical variable occur frequently in our data as numeric variables with missing values. Considerable number of missing values may arise from a default in derivation (e.g., computing a ratio when encountering zero in either the numerator or the denominator or both). Redistribution of non-missing vs missing categories can be measured by JS PSI. Shift in a numeric variable may be measured by AABC PSI. Since both AABC PSI and JS PSI are scaled between 0 and 100%, we can unify the two. We define a PSI for a mixed ordinal/categorical variable to be an interpolation between two categorical scenarios by the actual degree of ordinal shift.

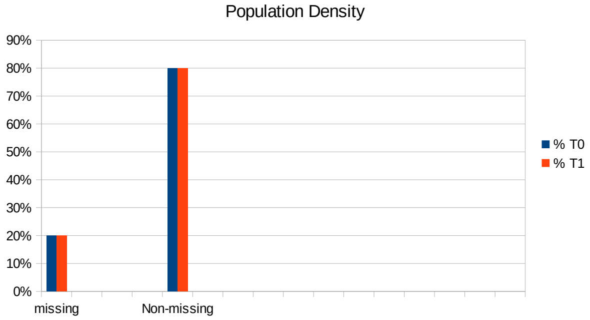
Take the case of a numeric variable with missing. AABC PSI measures the degree of shift a in the numbers. In the base scenario, JS PSI measures the stability b of the non-missing (numbers) and missing categories. In the contrasting scenario, we create a new category to house all the numbers at time T1. Here JS PSI measures the stability c with this new distribution. With interpolation, PSI for a numeric variable with missing is $b+a(c-b)$.

In the example below we calculate the PSI of a numeric variable with 20% missing unchanged from T0 to T1, and the non-missing bears a shift in example (iv) above with AABC PSI = 50% ($a=0.5$). Since there is no redistribution between the missing and non-missing, $b=0$ in the base scenario. For the contrasting scenario, we put all the numbers in a new category at T1, giving $c=0.8$.

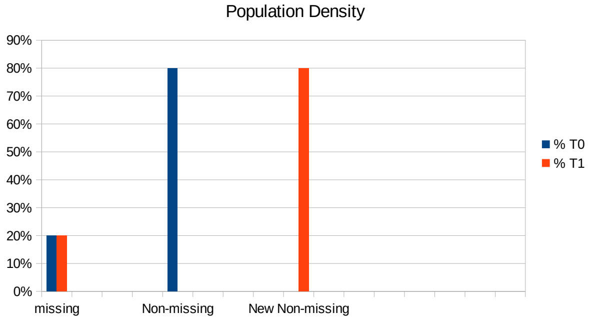
$a=0.5$ (AABC PSI)



$b=0$ (JS PSI base scenario)



$c=0.8$ (JS PSI contrasting scenario)



Post interpolation, PSI for this example numeric variable with missing is $b+a(c-b)=0+0.5(0.8-0)=0.4$ = 40%. A pseudo-spreadsheet is in the Appendix.

After reviewing hundreds of actual numeric variables, we establish 15% and 30% as our reference boundaries between minor, medium, and major instability. These are not direct equivalents of the customary 0.1 and 0.25 JD PSI cutoffs, but they provide roughly the same trigger frequencies. We hope our readers find these references useful in their population stability work.

Conclusion

In this article we analyze the disadvantages of the conventional population stability index, list the qualities of a desirable one, research the literature for a suitable substitute, and end up building our own. We define our PSI to require no binning. For ordinal variables we have a shift sensitive AABC PSI as: $2\sum m_k\left|\mathring{P}_k-\mathring{Q}_k\right|$ where $\mathring{P}$ and $\mathring{Q}$ are mid-CDFs of the variable's distributions at two time points, and m_k is the average mixture density $(p_k+q_k)/2$. For categorical variables we have JS PSI between distributions p and q as: $\left(\sum p_k(\ln p_k-\ln m_k)+\sum q_k(\ln q_k-\ln m_k)\right)/(2\ln 2)$. For mixed ordinal/categorical variables (e.g., numeric variable with missing) we interpolate between the base categorical scenario JS PSI (b) and contrasting new category scenario JS PSI (c) by the ordinal AABC PSI (a) in the numbers. The resulting interpolated PSI is $b+a(c-b)$.

References

1. Kullback, S.; Leibler, R.A. (1951). "On information and sufficiency" Annals of Mathematical Statistics. 22.

2. Jeffreys, H. (1948). Theory of Probability, 2nd ed. The Clarendon Press, Oxford.

3. Lin, J. (1991). "Divergence measures based on the Shannon entropy" IEEE Transactions on Information Theory. 37 (1): 145–151.

4. Vaserstein, L. N. (1969). "Markov processes over denumerable products of spaces, describing large systems of automata" Problemy Peredači Informacii. 5 (3): 64–72.

5. Anderson, T. W. (1962). "On the Distribution of the Two-Sample Cramer–von Mises Criterion" Annals of Mathematical Statistics. Institute of Mathematical Statistics. 33 (3): 1148–1159.

6. Lancaster, H. O. (1961). "Significance tests in discrete distributions" Journal of the American Statistical Association, 56, 223-234.

7. Parzen, E. (2009). "United Applicable Statistics: Mid-Distribution, Mid-Quantile, Mid P Confidence Intervals Proportion p" University of Maryland.

Appendix

Here we provide an example to calculate PSI for a numeric variable with missing values.

We first construct a table with missing and sorted numeric values {1}. We count the population at T0 {2}, and also at T1 {3}. For the non-missing numeric portion we compute the distribution density at T0 {4}, and at T1 {5}. We compute the missing and non-missing {6} categorical densities at T0 {7}, and at T1 {8}.

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
Value	Count #T0	Count #T1	Numeric dist. %T0	Numeric dist. %T1	Value	Categorical dist. %T0	Categorical dist. %T1
Missing	2	2	N/A	N/A	Missing	20%	20%
0	1		12.5%		Non-Missing	80%	80%
1	1		12.5%				
2	1	1	12.5%	12.5%			
3	1	1	12.5%	12.5%			
4	1	1	12.5%	12.5%			
5	1	1	12.5%	12.5%			
6	1		12.5%				
7	1		12.5%				
8		1		12.5%			
9		1		12.5%			
10		1		12.5%			
11		1		12.5%			
Total	10	10	100%	100%	Total	100%	100%

Compute AABC PSI

With the non-missing portion we compute the numeric shift AABC PSI. We obtain the cumulative distributions at T0 {9} and at T1 {10}. Then we compute the mid-CDFs at T0 {11} and at T1 {12}, and the absolute difference between the mid-CDFs {13}. The average mixture density {14} is calculated based on the distribution densities at T0 and at T1. The AABC PSI {15} is computed element-wise per numeric value and then summed up.

{1}	{4}	{5}	{9}	{10}	{11}	{12}	{13}	{14}	{15}
Value	Numeric dist. %T0	Numeric dist. %T1	Cumulative dist. %T0	Cumulative dist. %T1	Mid-CDF T0	Mid-CDF T1	Mid-CDFs Absolute Difference	Average Mixture Density	AABC PSI by numeric value
			accum. {4}	accum. {5}	{9} - {4}/2	{10} - {5}/2	abs({11}-{12})	({4}+{5})/2	2*{13}*{14}
0	12.5%		12.5%	0%	6.25%	0%	6.25%	6.25%	0.78125%
1	12.5%		25.0%	0%	18.75%	0%	18.75%	6.25%	2.34375%
2	12.5%	12.5%	37.5%	12.5%	31.25%	6.25%	25.00%	12.5%	6.25000%
3	12.5%	12.5%	50.0%	25.0%	43.75%	18.75%	25.00%	12.5%	6.25000%
4	12.5%	12.5%	62.5%	37.5%	56.25%	31.25%	25.00%	12.5%	6.25000%
5	12.5%	12.5%	75.0%	50.0%	68.75%	43.75%	25.00%	12.5%	6.25000%
6	12.5%		87.5%	50.0%	81.25%	50.00%	31.25%	6.25%	3.90625%
7	12.5%		100%	50.0%	93.75%	50.00%	43.75%	6.25%	5.46875%
8		12.5%	100%	62.5%	100%	56.25%	43.75%	6.25%	5.46875%
9		12.5%	100%	75.0%	100%	68.75%	31.25%	6.25%	3.90625%
10		12.5%	100%	87.5%	100%	81.25%	18.75%	6.25%	2.34375%
11		12.5%	100%	100%	100%	93.75%	6.25%	6.25%	0.78125%
							Total of {15}	**AABC PSI**	***a* = 50.0%**

Compute Base JS PSI

Next we compute the base scenario categorical redistribution JS PSI (b):

{6}	{7}	{8}	{16}	{17}	{18}	{19}
Value	Categorical dist. %T0	Categorical dist. %T1	Average Mixture Density ({7}+{8})/2	{7}* ln({7}/{16})	{8}* ln({8}/{16})	JS PSI by catetory ({17}+{18}) /(2*ln(2))
Missing	20%	20%	20%	0.000	0.000	0%
NonMissing	80%	80%	80%	0.000	0.000	0%
		Total of {19}	**Base**	**Scenario**	**JS PSI**	***b* = 0.0%**

Compute Contrasting Scenario JS PSI

We construct a contrasting scenario by moving the non-missing at T1 into a new category to compute categorical redistribution JS PSI (c):

{6}	{7}	{8}	{16}	{17}	{18}	{19}
Value	Categorical dist. %T0	Categorical dist. %T1	Average Mixture Density ({7}+{8})/2	{7}* ln({7}/{16})	{8}* ln({8}/{16})	JS PSI by catetory ({17}+{18}) /(2*ln(2))
Missing	20%	20%	20%	0.000	0.000	0%
NonMissing	80%		40%	0.555		40%
New		80%	40%		0.555	40%
		Total of {19}	**Contrasting**	**Scenario**	**JS PSI**	***c* = 80.0%**

{17} is blank where {7} is blank, based on multiplier before ln(); similarly for {18}.

Compute Composite PSI

PSI for a numeric variable with missing is obtained by interpolating between the base and contrasting JS PSI scenarios using AABC PSI as the interpolation factor:

a	b	c	
AABC PSI	JS PSI base scenario	JS PSI contrasting scenario	**Composite PSI b+a*(c-b)**
50.0%	0.0%	80.0%	**40.0%**