A Note on Decomposing the Difference of Two Ratios

Daniel R. Corro

A Note on Decomposing the Difference of Two Ratios

Dan Corro
Director of Claims Research
National Council on Compensation Insurance, Inc.

August, 1999

Abstract

Consider a ratio statistic (e.g. the mean) built from observations assigned into classes. An example would be losses=L, claim counts=C, and exposures=E each aggregated by rating class with the applicable statistic being either case severity=L/C or case frequency=C/E. The note discusses comparing two observed values for such a statistic. The difference is expressed as a sum of two components. One component measures the change due to the change in class mix. The other measures the change "holding the class mix constant". It is shown that a T-test on each component can assess whether it represents a nonzero difference. A simple numeric example is presented and an Appendix provides a SAS routine to perform the calculations.

Introduction

One of the ongoing assignments of the claims research department at the National Council on Compensation Insurance, Inc. (NCCI) is to monitor the experience of states which have introduced reforms in their workers compensation (WC) systems. These state specific post reform monitor (PRM) reports analyze WC costs by breaking them down into their frequency and severity components. That experience is compared with a benchmark determined from the pooled experience of a collection of other states, selected from among states that did not undergo major changes in their WC systems over the time frame of the study. The time frame of the study is, in turn, broken down into two subsets, corresponding to the pre- and post-reform experience of the PRM state. This approach leads to a number of two-way comparisons: PRM state vs. benchmark states; pre- vs. post-reform time period.

A primary PRM data source is the NCCI unit record system (URS), used for WC ratemaking and experience rating. A key feature of URS data is its capture according to the job classification system used for pricing WC insurance. When comparing frequency and severity, over time or between jurisdictions, it is obviously important to be able to account for differences in exposure mix. This short note describes the technique developed specifically for the PRMs but which clearly has general application.

The Decomposition

The idea comes from simple arithmetic. Consider any paired comparison, indexed by $j \in \{1,2\}$, in which a "numerator" $N_j = \sum_i n_{ji}$ and "denominator" $D_j = \sum_i d_{ji}$ are determined by summing over a common set of disjoint classes, indexed by i.

The difference of the ratios $\rho_j = {}^{N_j}\!/_{D_j}$ can be decomposed as:

$$\rho_2 - \rho_1 = \alpha + \beta \quad \textit{where, letting } r_{ji} = {}^{n_{ji}}\!/_{d_{ji}},$$

$$\alpha = \sum_i r_{1i}\left(d_{2i}/D_2 - d_{1i}/D_1\right) \text{ and } \beta = \sum_i (r_{2i} - r_{1i})\left(d_{2i}/D_2\right).$$

The component α is referred to as the *class mix component* and β, which is a weighted sum of ratio differences by class, as the *matched difference component.* Observe that when the two denominators share a common class decomposition, i.e. $d_{2i} = a \cdot d_{1i}$ for some fixed number a, then $\alpha = 0$. By the same token, if the two ratios are the same for all classes ($r_{2i} = r_{1i}$ for all i) then $\beta = 0$.

Observe that the matched difference component

$$\beta = \sum_i \left(n_{2i}/d_{2i}\right)\cdot\left(d_{2i}/D_2\right) - \sum_i r_{1i}\cdot\left(d_{2i}/D_2\right) = \rho_2 - \sum_i r_{1i}\left(d_{2i}/D_2\right) = \rho_2 - \hat{\rho}_1$$

may be interpreted as the difference between ρ_2 and the result of reweighting the observed ratios $\{r_{1j}\}$, which yielded the first ratio ρ_1, to match ρ_2's denominator distribution $\{d_{2i}\}$.

Of course, the ratios ρ_1 and ρ_2 can be regarded as weighted means. Indeed, we will regard the d_{ji} both as individual observed "denominators" as well as weights. It is natural to consider testing whether the difference of means $\rho_2 - \rho_1 = \alpha + \beta$ is significantly different from 0. The usual test for this is the conventional T-test of mean difference. In its customary formulation, however, that test is not suited for weighted observations. For example, the SAS TTEST procedure does not support a weight variable, even though the SAS package is most accommodating of weighted data. It is well known, however, that the customary T-test of mean difference is a special case of the OLS regression calculation. Indeed, the coefficient parameters are routinely tested for significance using a T-test. The Appendix illustrates, using SAS, a simple and general way to test whether

the difference $\rho_2 - \rho_1$ of weighted means is significantly different from 0, via weighted OLS.

In this regard, consider the set of matched pairs $\{(r_{1i}, r_{2i}) \mid 1 \le i \le n\}$ in which the paired observations are assigned weights according to the second denominator distribution $\{d_{2i}\}$. Those familiar with what the SAS documentation refers to as a "matched T-test" to determine whether the ratios are different, will note that it is in fact the matched difference component $\beta = \rho_2 - \hat{\rho}_1$ that is being tested. To see this, first recall that, unlike the conventional T-test of mean difference in PROC TTEST, SAS recommends the use of PROC MEANS to perform a matched T-test. The SAS PROC MEANS directly accommodates weighted data (although one has to choose a weight, here we chose the $\{d_{2i}\}$). The idea is to consider the set of matched differences $\{x_i = r_{2i} - r_{1i} \mid 1 \le i \le n\}$ to determine whether its mean is different from 0. The value being tested is thus the weighted mean:

$$\sum_i (x_i) \cdot \left(\frac{d_{2i}}{D_2}\right) = \sum_i (r_{2i} - r_{1i}) \cdot \left(\frac{d_{2i}}{D_2}\right) = \beta, \text{ as claimed.}$$

A more formal statement of our result (which allows for nonnegative weights, rather than the strictly positive weights demanded by the $r_{ji} = {}^{n_{ji}}\!/_{d_{ji}}$ formulation) is provided below:

Proposition: *Given any ordered set of 2N nonnegative real numbers:*

$$\{r_{ji}, d_{ji} \mid j = 1,2; 1 \le i \le N\}$$

Set

$$D_j = \sum_i d_{ji},\ R_j = \sum_i r_{ji} \quad \textit{and} \quad \rho_j = \sum r_{ji}\left(\frac{d_{ji}}{D_j}\right)$$

and assume $D_j > 0$ *and* $R_j > 0,\ j = 1,2.$

Define

$$\alpha = \sum_i r_{1i}\left(d_{2i}/D_2 - d_{1i}/D_1\right) \quad \textit{and} \quad \beta = \sum_i (r_{2i} - r_{1i})\left(d_{2i}/D_2\right).$$

α is referred to as the **class mix component** *and β as the* **matched difference component.**

Then:

(i) $\rho_2 - \rho_1 = \alpha + \beta$.

(ii) *An appropriate test of the hypothesis $\alpha = 0$ is a matched T–test on the pairs* $\{(\frac{d_{1i}}{D_1}, \frac{d_{2i}}{D_2}) \mid 1 \le i \le N\}$ *using* $\{r_{1i}\}$ *as weights.*

(iii) *An appropriate test of the hypothesis $\beta = 0$ is a matched T–test on the pairs* $\{(r_{1i}, r_{2i}) \mid 1 \le i \le N\}$ *using* $\{d_{2i}\}$ *as weights.*

Proof: Everything follows directly from earlier remarks except (ii) on testing the hypothesis $\alpha = 0$. In that regard, set

$$\hat{r}_{2i} = d_{2i}, \hat{r}_{1i} = d_{1i}, \hat{d}_{2i} = r_{1i} \textit{ and } \hat{d}_{1i} = r_{2i}$$

and apply the established part of the proposition to the hatted numbers, noting that:

$$\alpha = R_1\hat{\beta}.$$

Since a T-test is unaffected by multiplication by a positive constant, the result follows from (iii) as applied to $\hat{\beta}$. This establishes the proposition.

An (unmatched) T-test for the difference of means $\rho_2 - \rho_1$ involves $2N - 2 = 2(N - 1)$ degrees of freedom, which the proposition suggests can be split evenly between the two matched T-tests for α and β, each involving $N - 1$ degrees of freedom.

A Numeric Example

This note concludes with a simple numeric example, designed to illustrate the calculation as well as the need to account for class mix when making comparisons. Think of the r-

values as a cost measure (frequency or severity) and the d-values as the exposure base (payrolls or cases). The Group 1 data is meant to suggest a starting point situation in which much of the exposure lies in high cost classes (I and J). This changes into the Group 2 situation with most of the exposure assumed to move into the lower cost classes (A, B and C). The cost within class is fairly similar between the two groups but note that for every class, the Group 2 cost equals or exceeds that for Group1. The Appendix provides a SASLOG and listing of the routine used to make the calculations.

Data Table				
	Group 1		Group 2	
Class	r_1	d_1	r_2	d_2
A	95	2	98	30
B	100	2	100	30
C	105	1	106	20
D	295	5	298	2
E	300	10	300	5
F	305	10	308	2
G	310	10	310	2
H	495	10	500	1
I	500	25	505	5
J	505	25	505	3

Decomposition of Ratio Difference				
Component	Value	T-Test	T-Value	P-Value
α	-254.15	Matched	-1.9456	0.0836
β	1.52	Matched	2.9135	0.0172
$\rho_2-\rho_1$	159.32 - 411.95 = - 252.63	Unmatched	-4.409	0.0003

All of the decline in overall mean cost from Group 1 to Group 2 is attributed to the change in class mix component α. The matched difference component β works in the opposite direction, due to the higher class costs for Group 2. Observe that the overall

difference is the most statistically significant finding, as measured by the lowest P-value. It is interesting to note that the dominating component numerically, the change in the class mix, is of marginal statistical significance. On the other hand, the numerically smaller matched difference component reflects a statistically significant increase in the by class costs.

APPENDIX

SASLOG

```
*************************************************************************;
 373        DATA ONE;
 374        INPUT  R1 D1 R2 D2;
 375        CARDS;

 NOTE: The data set WORK.ONE has 10 observations and 4 variables.
 NOTE: The DATA statement used 3248K.

 386        ;
 387        PROC PRINT   DATA=ONE;

 NOTE: The PROCEDURE PRINT printed page 1.
 NOTE: The PROCEDURE PRINT used 3381K.

 388        PROC SUMMARY DATA=ONE;
 389        VAR D1 D2 R1 R2; ;
 390        OUTPUT OUT = SUMM SUM = SD1 SD2 SR1 SR2;
 391        *DEFINE DIFFERENCES;

 NOTE: The data set WORK.SUMM has 1 observations and 6 variables.
 NOTE: The PROCEDURE SUMMARY used 3521K.

 392        DATA ONE;
 393        SET ONE;
 394        KEEP R1 R2 D1 D2 R2_R1 D2_D1;
 395        RETAIN SD1 SD2 SR1 SR2;
 396        IF _N_ = 1 THEN SET SUMM;
 397        D2_D1 = (D2/SD2 - D1/SD1)*SR1;
 398        R2_R1 = R2 - R1;

 NOTE: The data set WORK.ONE has 10 observations and 6 variables.
 NOTE: The DATA statement used 3558K.

 399        PROC MEANS DATA=ONE;
 400        VAR R1;
 401        WEIGHT D1;
 402        TITLE2 'WEIGHTED MEAN OF R1, WEIGHT D1';

 NOTE: The PROCEDURE MEANS printed page 2.
 NOTE: The PROCEDURE MEANS used 3572K.

 403        PROC MEANS DATA=ONE MEAN STDERR T PRT;
 404        VAR R2 R2_R1;
 405        WEIGHT D2;
 406        TITLE2 'WEIGHTED MEAN OF R2 AND MATCHED T-TEST,WEIGHT D2';

 NOTE: The standard error of the mean is computed as sqrt( weighted
sample variance / sum of weights ).
 NOTE: The PROCEDURE MEANS printed page 3.
 NOTE: The PROCEDURE MEANS used 3572K.

 407        PROC MEANS DATA=ONE MEAN STDERR T PRT;
```

```
 408         VAR D2_D1;
 409         WEIGHT R1;
 410         TITLE2 '  MATCHED WEIGHTED T-TEST D2 - D1 MEANS WEIGHT R1';
13 The SAS System
11:51 Monday, August 23, 1999

 411         *SET UP TO DO WEIGHTED TTEST OF R1-R2 USING OLS';

 NOTE: The standard error of the mean is computed as sqrt( weighted
sample variance / sum of weights ).
 NOTE: The PROCEDURE MEANS printed page 4.
 NOTE: The PROCEDURE MEANS used 3572K.

 412         DATA TWO;SET ONE;
 413         KEEP R D C;
 414         R = R1;D = D1;C = 0;OUTPUT;
 415         R = R2;D = D2;C = 1;OUTPUT;

 NOTE: The data set WORK.TWO has 20 observations and 3 variables.
 NOTE: The DATA statement used 3572K.

 416         PROC REG DATA=TWO;
 417         MODEL R = C;
 418         WEIGHT D;
 419         TITLE2 'UNMATCHED WEIGHTED T-TEST USING OLS';
 NOTE: 20 observations read.
 NOTE: 20 observations used in computations.
 NOTE: The PROCEDURE REG printed page 5.
 NOTE: The PROCEDURE REG used 4071K.

 NOTE: The SAS session used 4071K.
 NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
```

OUTPUT LISTING

Page 1:

OBS	R1	D1	R2	D2
1	95	2	98	30
2	100	2	100	30
3	105	1	106	20
4	295	5	298	2
5	300	10	300	5
6	305	10	308	2
7	310	10	310	2
8	495	10	500	1
9	500	25	505	5
10	505	25	505	3

Page 2: WEIGHTED MEAN OF R1, WEIGHT D1

Analysis Variable : R1

N	Mean	Std Dev	Minimum	Maximum
10	411.9500000	391.7262011	95.0000000	505.0000000

Page 3: WEIGHTED MEAN OF R2 AND MATCHED T-TEST,WEIGHT D2

Variable	Mean	Std Error	T	Prob>³T³
R2	159.3200000	41.8235681	3.8093354	0.0042
R2_R1	1.5200000	0.5217066	2.9135150	0.0172

Page 4: MATCHED WEIGHTED T-TEST D2 - D1 MEANS WEIGHT R1

Analysis Variable : D2_D1

Mean	Std Error	T	Prob>³T³
-254.1500000	130.6297110	-1.9455758	0.0836

Page 5:UNMATCHED WEIGHTED T-TEST USING OLS

Model: MODEL1

Dependent Variable: R

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	3191095.845	3191095.845	19.436	0.0003
Error	18	2955334.51	164185.25056		
C Total	19	6146430.355			

Root MSE	405.19779	R-square	0.5192
Dep Mean	285.63500	Adj R-sq	0.4925
C.V.	141.85859		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > ³T³
INTERCEP	1	411.950000	40.51977919	10.167	0.0001
C	1	-252.630000	57.30362127	-4.409	0.0003