Pitfalls in using the relative standard deviation of win percentages to measure competitive balance in sports leagues

P. Dorian Owen

ABSTRACT
The relative standard deviation of win percentages is by far the most commonly used measure of within-season competitive balance in sports leagues. It compares the actual (ex post) standard deviation of win percentages, across teams in the league within a season, to the standard deviation of win percentages in the ‘ideal’ case in which each team has an equal chance of winning each game. The popularity of the relative standard deviation stems from the belief that, unlike the actual standard deviation, it controls for both the number of teams and the number of rounds in a league competition; hence, the relative standard deviation is usually advocated for comparisons of competitive balance over time and/or across different leagues. However, this paper shows that the relative standard deviation is much more sensitive to changes in the number of teams and the number of games played than is the case for the actual standard deviation. This can lead to misleading comparisons of within-season competitive balance across leagues or over time if the numbers of teams and/or games played are not constant, which in practice is usually the case. Using a normalized standard deviation measure avoids these problems.

Keywords:
Competitive balance, Relative standard deviation, Idealized standard deviation

JEL classification:
D63, L83

Contact details:
Dorian Owen
Department of Economics
University of Otago
PO Box 56
Dunedin 9054
New Zealand
email: Dorian.Owen@otago.ac.nz
tel: (64) 3 479 8655
fax: (64) 3 479 8174
1. Introduction

Competitive balance in sports leagues, i.e., how evenly teams are matched, is reflected in the degree of inequality in match and championship outcomes. Because of its pivotal role in the economic analysis of professional sport, considerable effort has gone into measuring competitive balance. By far the most commonly used measure is the relative standard deviation of win percentages. This compares the actual (ex post) standard deviation of win percentages with the standard deviation of win percentages in the ‘ideal’ case in which each team has an equal chance of winning each game.

The relative standard deviation of win percentages is widely regarded as the most useful measure of competitive balance “because it controls for both season length and the number of teams, facilitating a comparison of competitive balance over time and between leagues” (Fort, 2007, p. 643). This widely held view is misplaced. For comparable levels of competitive balance, the relative standard deviation of win percentages is sensitive to changes in the number of teams or games played. This can lead to misleading comparisons of within-season competitive balance across leagues or over time if the numbers of teams and/or games played are not constant, which in practice is usually the case. These problems can be avoided by using a normalized standard deviation measure that takes into account variations in the relevant upper bound.

2. Measuring competitive balance with actual and relative standard deviations

Competitive balance in a sports league is a multi-faceted concept. The different dimensions include the distribution of wins across teams in the league within a single
season, the persistence of teams’ record of wins across successive seasons over time, and
the degree of concentration of overall championship wins reflected in teams’ shares of
championship wins over a number of seasons (Kringstad and Gerrard, 2007).

The ex post or ‘actual’ standard deviation (ASD) of teams’ win ratios (or, equivalently,
win percentages) in a single season is a natural measure for the first of these dimensions.
This can be represented as

\[ ASD = \sqrt{\frac{ \sum_{i=1}^{N} \left[ \frac{w_i}{G_i} - 0.5 \right]^2 / N } {N}} \]  

in which \( N \) equals the number of teams in the league, and \( w_i \) and \( G_i \) are, respectively, the
number of wins and the number of games played by team \( i \) in a season. A smaller standard
deviation of win ratios across teams in a season indicates a more equal competition.

However, when comparing values of ASD, either for the same league over time or across
different leagues, \( N \) and/or \( G \) are typically not constant. Other things equal, ASD tends to
decrease as \( G \) increases, so it is common to compare ASD to a benchmark ‘idealized
standard deviation’ corresponding to an ex ante representation of a perfectly balanced
league in which each team has an equal probability of winning each game.\(^1\) In the absence
of ties (draws), the idealized standard deviation, ISD = 0.5/G^{0.5} can be derived as the
standard deviation of a binomially distributed random variable with a (constant) probability
of success of 0.5 across independent trials (Fort and Quirk, 1995).\(^2\) As \( G \) increases, there is
likely to be less random noise in the final outcomes and, hence, the idealized standard

\(^1\) The use of a relative measure involving a benchmark standard deviation corresponding to an ex ante
perfectly balanced league is attributable to Noll (1988) and Scully (1989), but became popular following its
use by Quirk and Fort (1992) and Fort and Quirk (1995).

\(^2\) If ties are possible, ISD can be applied to absolute total points or the percentage of points, with amendments
to account for different possible points allocations for wins, ties and losses (e.g., Fort, 2007).
deviation will be smaller. The relative standard deviation, \( RSD \) is thus expressed as \( \frac{ASD}{ISD} \).

\( RSD \) is a ‘static’ measure based on the variation of (final) win ratios across teams in a single season. Its evolution can be plotted over time, but it does not capture championship concentration or persistence of performance of individual teams over successive seasons. Given the multidimensional nature of competitive balance, it is generally considered unrealistic to expect any single measure to reflect all of its different dimensions. This apart, the \( RSD \) measure has met with widespread acceptance. It is the most widely used competitive balance measure in the sports economics literature; e.g., see the list of competitive balance measures used in studies of the big four North American professional sports leagues cited in Fort (2006a, Table 10.1).

However, despite its resounding endorsement as “the tried and true” measure of within-season competitive balance (Utt and Fort, 2002, p. 373), \( RSD \) has properties that can lead to invalid conclusions in comparisons of competitive balance involving different numbers of teams and/or games.

\( RSD \) has an upper bound, because teams can not win games in which they do not play. The complications this causes for interpretation of the Gini coefficient and the Herfindahl-Hirschman index applied to wins are documented by Utt and Fort (2002) and Owen et al. (2007) respectively, but this well-known feature of the distribution of wins in sports leagues also has implications for \( RSD \) that have not been recognized.

Another distinctive feature of \( RSD \) is the different measures of ‘sample size’ that appear in its numerator \( (N, \) the number of teams) and denominator \( (G, \) the number of games played by each team). If each team plays the other teams more than once in a season, then \( N \) and \( G \)
usually differ markedly. These characteristics can invalidate exactly the sorts of comparisons of competitive balance (involving scenarios with different $N$ and/or $G$) for which $RSD$ is usually advocated (e.g., Fort, 2006b, pp. 175-177; Leeds and von Allmen, 2005, pp. 160-161).

3. **The sensitivity of the relative standard deviation to $N$ and $K$**

The degree of sensitivity of $RSD$ to variations in $N$ and $K$ can most easily be seen by considering the ex post ‘most unequal distribution’ of win ratios (Fort and Quirk, 1997; Horowitz, 1997; Utt and Fort, 2002). This involves one team wining all its games, the second team wining all except its game(s) against the first team, and so on down to the last team, which wins none of its games. For ease of exposition, consider balanced schedules of games in which each team plays every other team the same number of times, $K$. Each team plays $G_i = G = K(N - 1)$ games. The upper bound for $ASD$, denoted $ASD_{ub}$, is given by:  

$$ASD_{ub} = \left[\frac{(N + 1)/\{12(N - 1)\}}{0.5/[K(N - 1)]^{0.5}}\right]^{0.5}$$  \hspace{1cm} (2)$$

Substituting $G = K(N - 1)$ into the expression for $ISD$, and noting that the ex ante $ISD$ measure is unaffected by the actual outcome for $ASD$, gives:

$$RSD_{ub} = \frac{ASD_{ub}}{ISD} = \frac{[(N + 1)/\{12(N - 1)\}]^{0.5}}{0.5/[K(N - 1)]^{0.5}} = 2[K(N + 1)/12]^{0.5}$$  \hspace{1cm} (3)$$

3 With unbalanced schedules, additional assumptions about specific teams’ performances are required to specify the most-unequal distribution and, hence, the upper limit of the standard deviation of win ratios.

4 See the Appendix for a derivation. Eq. (2) can also be obtained by substituting the expression for the corresponding upper bound of the Herfindahl-Hirschman index (HHI) applied to wins, derived in Owen et al. (2007, Appendix, pp. 300-301), into Depken’s (1999) Eq. (6), which expresses the variance of wins in terms of the HHI.
The upper bound of $RSD$ in Eq. (3) depends not only on the number of teams in the league, $N$, but also on the number of times they play against each other, $K$. This is due to the dependence of $ISD$ on $G$ and hence $K$, in contrast to $ASD^{ub}$, which is invariant to $K$. Increases in $N$ and/or $K$ lead to increases in $RSD^{ub}$. This is conventionally interpreted as implying a decrease in competitive balance, even though, given we are considering the upper bound of $RSD$, wins are initially perfectly unequally distributed and remain that way. The upper bound of $ASD$ in Eq. (2) also depends on $N$, with expansions in $N$ leading to a decrease in $ASD^{ub}$. However, $RSD^{ub}$ is much more sensitive than $ASD^{ub}$ to variations in $N$. This is illustrated in Fig. 1, and is apparent from a comparison of Eqs (2) and (3). For large $N$, the $(N + 1)$ and $(N - 1)$ terms approximately cancel out, so that, in the limit, $ASD^{ub}$ tends to $(1/12)^{0.5} = 0.289$. For smaller values of $N$, as in most sports leagues, the dependence on $N$ is not removed entirely, but is relatively modest, with, for example, $ASD^{ub}$ varying from 0.327 for $N = 8$ to 0.298 for $N = 30$ (a decrease of approximately 8.8%). In contrast, $RSD^{ub}$ → $\infty$ as $N$ → $\infty$ and, for commonly observed values of $N$, the increase in $RSD^{ub}$ is more dramatic than for $ASD^{ub}$, varying (for $K = 1$) from 1.732 for $N = 8$ to 3.215 for $N = 30$ (an increase of approximately 85.6%). Rather than purging $ASD$ of its dependence on the number of teams and the number of games played, $RSD$’s use of $ISD$ as a benchmark actually makes it more sensitive than $ASD$ to variations in $N$ (and $G$).

Variation in the upper bounds can be explicitly incorporated in a normalized measure of competitive balance, such as

$$ASD^* = \frac{ASD}{ASD^{ub}}$$

It is straightforward to show that $\frac{\partial ASD^{ub}}{\partial N} < 0$ if $N \geq 2$.

This asymptotic result is consistent with $ASD$ corresponding more closely to a pure inequality measure, such as $IGE(2)$, a member of the family of generalized entropy measures of inequality (Bajo and Salas, 2002). $IGE(2) = CV^2/2$, where $CV$ is the coefficient of variation. If the mean of the win ratios in a season equals 0.5, variation in $CV$ applied to win ratios corresponds to variation in $ASD$. 

---

5 It is straightforward to show that $\frac{\partial ASD^{ub}}{\partial N} < 0$ if $N \geq 2$. 
6 This asymptotic result is consistent with $ASD$ corresponding more closely to a pure inequality measure, such as $IGE(2)$, a member of the family of generalized entropy measures of inequality (Bajo and Salas, 2002). $IGE(2) = CV^2/2$, where $CV$ is the coefficient of variation. If the mean of the win ratios in a season equals 0.5, variation in $CV$ applied to win ratios corresponds to variation in $ASD$. 

---

5

---
The normalized measure, $ASD^*$, lies in the interval $[0, 1]$, with 0 representing perfect parity and 1 maximum imbalance.\(^7\) This measure is also suggested by Goossens (2006), although she calculates $ASD^{ub}$ numerically for different values of $N$ rather than deriving a general expression. Goossens’s argument for preferring a normalized $ASD$ measure to $RSD$ is that $RSD$ can be less than 1 (if $ASD < ISD$). This is not surprising given the ex ante probabilistic nature of $ISD$, in comparison to the ex post minimum (0) and maximum ($ASD^{ub}$) values for $ASD$. However, the problems emphasized above, which make $RSD$ unsuitable for comparisons involving different $N$ and $K$ values, provide a more fundamental justification for the use of a normalized standard deviation measure.

Problems with $RSD$ are easiest to demonstrate in cases in which $N$ or $K$ is varied and the degree of imbalance is controlled, as in the case of the upper bounds. Anomalous results are, however, not confined to the case of perfect inequality. For example, consider increasing $K$ while maintaining the original level of competitive balance by ‘scaling up’, i.e., reproducing exactly the corresponding set of results for the original league. As a concrete example, consider a league with balanced schedules ($N = 16$, $K = 2$, $G = 30$) in which $ASD$, is 0.037 and $RSD$ is 0.408.\(^8\) If $K$ is increased to 4 (e.g., each team playing every other team twice at home and twice away) and assuming the additional home and away results are identical to those in the actual two rounds played, then, clearly, this scaled up set of results will display unchanged win ratios (and the same value for $ASD$), reflecting an unchanged competitive balance situation. However, $RSD$ increases to 0.577 (due to the decrease in $ISD$). Increasing $K$ to 10 and again reproducing the original results five times,

\(^7\) Trivially, the corresponding measure for $RSD$, i.e., $RSD^* = RSD/RSD^{ub} = ASD^*$, so this adjustment also removes the dependence of the upper bound of $RSD$ on $K$.

\(^8\) These results apply to what is widely believed to be one of the most competitive league outcomes in association football: the 1983-84 season in the somewhat obscure Romanian Divizia C, Seria a VIII-a league (see [http://www.rsssf.com/miscellaneous/even.html](http://www.rsssf.com/miscellaneous/even.html)).
leaves win ratios and \( ASD \) unchanged but \( RSD \) increases to 0.913, despite no change in competitive balance. However, because the upper bound for \( RSD \) increases as \( K \) increases (as in Eq. (3)), the normalized measure, \( RSD^* (= ASD^*) \) remains unchanged.

If \( N \) varies across time for a given league, similar anomalies can occur. For example, in the First Division of New Zealand Rugby Union’s National Provincial Championship (NPC) \( K = 1 \) in round-robin play (prior to the semi-finals) but \( N \) has varied over time. The \( RSD \) of win ratios (with draws, which are relatively rare in rugby, counting as 0.5 of a win) was 1.763 in 1990 compared to 1.633 in 1994, suggesting a higher level of competitive imbalance in 1990. However, the upper bounds for \( RSD \) were, respectively, 2.000 \((N = 11)\) and 1.826 \((N = 9)\); \( RSD \) was therefore closer to the upper bound reflecting complete inequality in 1994 \((ASD^* = 0.894)\) than in 1990 \((ASD^* = 0.882)\).

Even more misleading results can occur if \( RSD \) is used to compare competitive balance across different leagues, because the differences in \( N \) and \( K \) are often considerably greater than for a single league over time. Because of the sensitivity of \( RSD^{ub} \) to \( K \) and \( N \), it is feasible for the observed values of \( RSD \) in one league (with \( 0 < RSD < RSD^{ub} \)) to be greater than the feasible value of \( RSD^{ub} \) in another. Interpreted literally this suggests that a less than completely unbalanced league (with larger \( K \) and/or \( N \)) is more unbalanced than a completely unbalanced league, which makes little sense. For example, the upper bound for \( RSD \) for the NPC noted above (1.825 to 2) is less than many of the calculated \( RSD \) values for the ‘big four’ US leagues reported in Fort’s (2006b) Table 6.3, especially basketball for which all reported values are greater than 2 and many are greater than 3. For this sort of comparison, the normalized measure, \( ASD^* \), is more appropriate.
4. Conclusion

In the sports economics literature there is a widespread belief that the idealized standard deviation provides a ‘common standard’ (Leeds and von Allmen, 2005, p. 160) against which to compare the actual standard deviation of win ratios. However, the resulting relative standard deviation measure has an upper bound and hence a range of feasible values which vary markedly in response to variation in the number of teams and/or number of games played. This can lead to anomalous results and misleading conclusions when comparing competitive balance across leagues or over time (involving scenarios with different \(N\) and/or \(G\)), i.e., exactly the sorts of comparisons of competitive balance for which \(RSD\) is advocated. The sensitivity of \(RSD\) to the number of teams and games provides a much more compelling reason than previously advanced for using a normalized standard deviation measure, rather than \(RSD\), to measure within-season competitive balance.
Appendix: Derivation of upper bounds for ASD and RSD

The upper bounds of ASD and RSD applied to win ratios are derived on the assumption of a perfectly unbalanced league of teams playing a balanced schedule with no ties (draws) or with ties (draws) treated as half a win. In a balanced schedule, in which each of the $N$ teams plays every other team the same number of times, $K$, each team plays $G = K(N - 1)$ games.

The actual (ex post) variance of win ratios ($AVAR$) across the $N$ teams in a season (with the mean win ratio equal to 0.5 for any degree of competitive balance) is given by:

$$AVAR = \left[ \sum_{i=1}^{N} \left( \frac{w_i}{G_i} \right)^2 / N \right] - (0.5)^2$$

In a perfectly unbalanced league

$$AVAR^{ub} = \frac{1}{N} \left[ \frac{K^2(N - 1)^2}{K^2(N - 1)^2} + \frac{K^2(N - 2)^2}{K^2(N - 1)^2} + \ldots + \frac{K^2(N - N)^2}{K^2(N - 1)^2} \right] - (0.5)^2$$

Note that the $K^2$ terms cancel, implying that $AVAR^{ub}$ and hence $ASD^{ub}$ are invariant to the number of rounds played if schedules are balanced. Simplifying,

$$AVAR^{ub} = \frac{1}{N(N - 1)^2} \left[ \frac{N(2(N - 1)(N - 1))^2}{6} \right] - (0.5)^2$$

$$= \frac{1}{N(N - 1)^2} \left( \frac{2N - 1}{6} \right) = \frac{N + 1}{12(N - 1)}$$

Taking the square root, $ASD^{ub} = (AVAR^{ub})^{0.5}$ gives the result in Eq. (2).

$$ISD = 0.5 / \sqrt{K(N - 1)}$$

$ISD$ does not depend on the actual outcome for ASD; hence

$$RSD^{ub} = ASD^{ub} / ISD$$

giving the result in Eq. (3).
References


Fig.1. Variation in the upper bounds of $RSD$ and $ASD$ with $N$ and $K$