SPORTS TEAM QUALITY AS A CONTEXT FOR UNDERSTANDING VARIABILITY

K. Laurence Weldon
Simon Fraser University, Canada
weldon@sfu.ca

When instruction in statistical concepts can be tied to practical sports issues, students are motivated to understand the statistical concepts. In this paper we describe an issue that would be relevant to discussions of many different sports leagues, and would also be a vehicle for teaching statistical concepts such as simulation, graphical displays, illusions of randomness, measurement of variability, and the logic of hypothesis testing. In addition to motivating a keen interest in the effects of random variation, these examples provide students with a way to verbalize what they learn in statistics classes to their lay acquaintances. Moreover, examples like these have the potential for engaging instructors who have been focused on more traditional approaches. Programs in the software language R are provided and their use with introductory classes is discussed.

INTRODUCTION

Many students are involved in, or at least interested in, sports. When statistical concepts can be tied to practical sports issues, there is motivation to also understand the statistical concepts. In this paper we describe an issue that would be relevant to the current status of many different sports leagues, and would also be a vehicle for teaching statistical concepts such as simulation, graphical displays, illusions of randomness, measurement of variability, and the logic of hypothesis testing. By using simulation for the demonstrations, it is possible to incorporate examples like these into statistics courses at any level. By using the demonstrations to study real and current data, we avoid the criticism of constructing artificial, historical, or irrelevant data to try to demonstrate the current utility of our statistical strategies. In addition to motivating a keen interest in the effects of random, unexplained variation, these examples provide students with a way to verbalize what they learn in statistics classes to their lay acquaintances.

The issue we address is: How can we separate chance and quality in team performance? Is the English Premier League soccer team Chelsea with 95 league points really a better team than Liverpool with 58 points? The performance of teams in a sports league may reflect both the quality of the team and other chance factors. However, it can be shown that the range of game points earned by the various teams in a sports league is not much greater than might be obtained assuming all teams have an equal chance of success at each match. Similarly, the goals scored for and against will occur with a range of ratios across the league, and this ratio is not much greater than would occur if all teams had the same distribution of “for” goals. To judge team quality on the basis of past performance in typical professional leagues is very difficult. Since such judgment is clearly important for awarding trophies, advising gamblers, arranging draft sequences, and even in our modern culture, it is important to detect those instances of performance that exceed the influence of chance factors. Some tools to study this problem are provided, and examples are demonstrated with some current leagues.

The immediate goal is to provide instructors with tools to simplify demonstrations of this kind of randomness in sports leagues. The ultimate goal is to encourage active discussion among students about everyday issues, like sports, that depend on statistical knowledge, thus enhancing the relevance of the subject to a wide variety of students.

EXAMPLES

We illustrate the methods of analysis with three examples; each example uses a slightly different approach, although each uses the common logic of hypothesis testing.

Example 1: Basketball

Consider a typical sports league status such as the thirty teams of the U.S. National Basketball Association: At the time of writing, each team had played approximately 38 games. The three teams with the best record had won 0.865, .769, and .744 of their games. The three teams at the bottom of the standings had won .270, .282, and .324 of their games. A partial league table showing these extremes is shown here (NBA, 2005):

	Wins	Losses	Win Rate
Detroit	32	5	.865
San Antonio	30	9	.769
Dallas	29	10	.744
(24 other teams)	13-25	13-26	.333-.658
Houston	12	25	.324
Charlotte	11	28	.282
Atlanta	10	27	.270

Each team played approximately 38 games but the fraction of games won varied from .270 to .865. Can we conclude that this ranking has some natural relationship to the quality of the team? Here the quality of team A relative to team B is defined as the probability that team A wins a match against team B. As a simplification, we assume that these relative qualities remain constant over the season, but this seems a reasonable approximation.

One way to approach the quality question for a given league is to simulate the results based on an assumption of equal quality for all teams and compare the vector of win rates (wins per game) in the simulated league, with the actual league outcomes. We call the hypothesis of "Equal Quality" EQ: each team in a match has the same probability to win the match. If we simulate 38 games for each team in the above league with the EQ hypothesis, each game being a 50-50 game, the result is:

Win Rate: Actual versus Equal-Quality Simulation

Rank	1	2	3	4	...	27	28	29	30
Actual wins/game	0.87	0.77	0.74	0.66	...	0.33	0.32	0.28	0.27
Sim'd wins/game	0.66	0.63	0.61	0.59	...	0.40	0.39	0.37	0.34
SD(sim'd w/g)	0.04	0.03	0.03	0.03	...	0.03	0.03	0.03	0.04

The third row is the expected rate for the team of the indicated rank when all 38 matches are decided by the toss of a fair coin, the EQ hypothesis. This is computed as the average for the team in this rank, over 1000 repetitions. The last row is the SD around the second row that would apply to the Actual wins/game under EQ. In this case the table shows that the actual wins/game is more variable than would be expected under EQ, suggesting that there is a difference in the quality of the thirty teams. But we should ask how strong this indication needs to be before we must reject the EQ hypothesis. Using SD of the fraction of wins across the teams (i.e., across the second row) as an index of team variability, the value of the index in this case is 0.14, and by simulation under EQ, we can show that anything over about 0.10 rejects the EQ hypothesis at the .05 level.

The above simulation of the EQ hypothesis can provide a bit more information. The expected proportion of wins under the EQ hypothesis, along with the SD of the proportion of wins under EQ, provides an interval with which to compare the actual proportion of wins. Using a 2 SD interval, we see that teams in ranks 1-4 exhibit superior win rates. It can also show (but not shown here) that many middling teams ranked 6-18 do not have win rates different from what the EQ hypothesis would predict. The underlying idea here is that there will always be a spread of the teams' win rates in a league, but some of that spread has nothing to do with differentials of team quality, but rather is ascribable to chance. The comparison with the equal teams hypothesis allows a separation of the influences, which is a common goal of statistical inference.

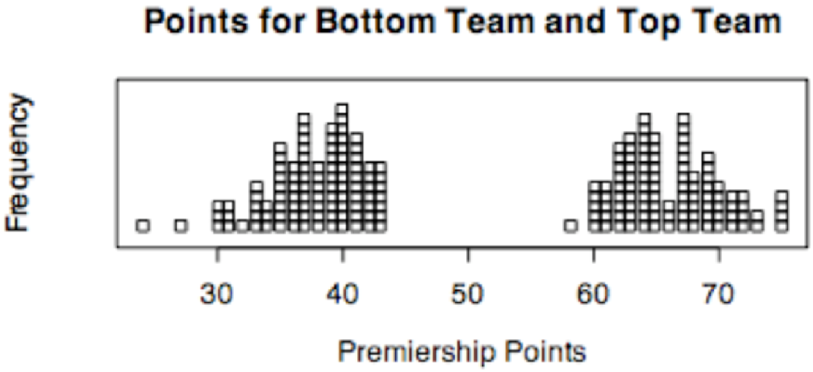
Example 2: Soccer (Association Football)

Many leagues involve games that can end in a tie. Usually for these, league rankings use a point system: e.g., 3 for a win, 1 for a tie, and 0 for a loss. For example, soccer leagues have 20-30 percent tie games, so the win-percentage does not capture the full outcome. The English

Premier soccer league in 2004-2005 had 38 games for each of the 20 teams in the league, and 28 percent of the games ended in a tie. The league points (0-1-3) earned ranged from 32 to 95 at the close of the regular season. Of interest in this case is the spread of league points under the EQ hypothesis: the probabilities for each team under EQ are, for W-D-L, .36,.28,.36 respectively. Again simulation can provide information that helps in interpreting the real data: The SD of league points across the 20 teams is 17.1 whereas the simulation shows that it would usually (probability .95) be less than 9 for teams of equal quality.

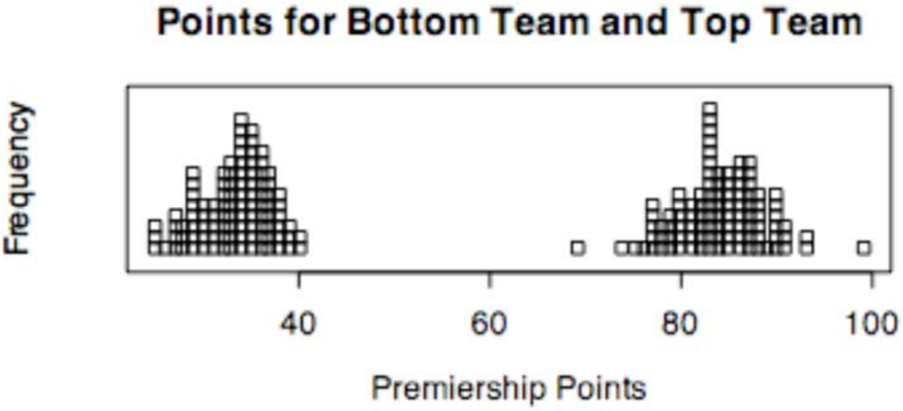
	W	D	L	Pts
Chelsea	29	8	1	95
Arsenal	25	8	5	83
Manchester United	22	11	5	77
Everton	18	7	13	61
Liverpool	17	7	14	58
Bolton	16	10	12	58
Middlesbrough	14	13	11	55
Manchester City	13	13	12	52
Tottenham	14	10	14	52
Aston Villa	12	11	15	47
Charlton	12	10	16	46
Birmingham	11	12	15	45
Fulham	12	8	18	44
Newcastle	10	14	14	44
Blackburn	9	15	14	42
Portsmouth	10	9	19	39
West Bromwich	6	16	16	34
Crystal Palace	7	12	19	33
Norwich	7	12	19	33
Southampton	6	14	18	32

The “Pts” vector from the above table suggests that the quality differential may exist only between the top three teams and the other 17 teams. One way to investigate this suggestion is to look at the distribution of max and min league points under the EQ hypothesis.



It does look as though the top teams, Chelsea, Arsenal and Manchester United, with 95, 83, 77 points league points respectively, have more points than one would expect if they had the same chance of winning as the other teams. However, the teams at the bottom of the league are not worse than the bottom teams would be under the EQ hypothesis.

We can do a bit more to investigate the apparent inequality of the top three teams as compared to the bottom 17 teams. If we postulate that the top three teams are in a class by themselves, but equal in that class, and similarly for the bottom 17 teams, simulation reproduces the kind of pattern of league points that the Premier League experienced in 2004-5. The three teams at the top, Chelsea, Arsenal and Manchester United, can be given a simulation edge in games with the other 17 teams: the probability is .63, instead of .36, that the elite team wins a given contest with the non-elite team (remember that the tie rate is still set at .28). It turns out that this edge is what is needed to replicate the seasons' outcomes. As a partial demonstration of this, the following figure can be presented based on 100 simulated seasons:



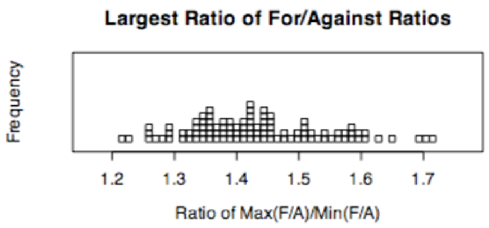
Another way to describe the outcome of this modified EQ simulation, is that the 17 lower-quality teams must increase their win-rate against the top three teams from 0.36 towards 0.63 to be competitive with the best teams in this league.

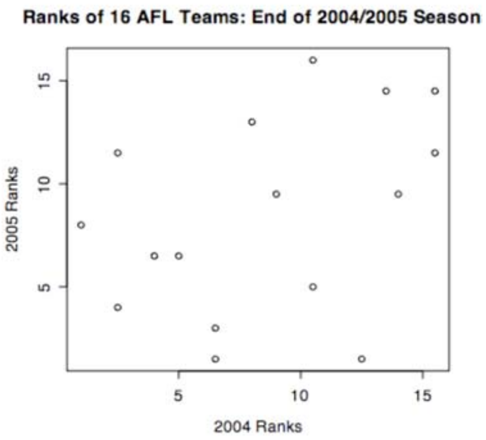
A way to validate our conclusion here is to look at the season-to-season variation. The "top-three-elite" proposal is validated by the fact that, in this 20 team league, only three teams have ended the season at the top of the league over the last ten years, and the three teams are exactly the same ones that were in the top three in the 2004-5 season: Manchester United, Arsenal and Chelsea (Hutchison, 2004b). Chelsea is at the top so far in the 2005-6 season. It is perhaps surprising that the English Premier League remains so popular with such persistent quality differentials over several years. The EQ hypothesis might be proposed as a property of a healthy league, but this league would be an exception.

Example 3: Football

A third example of a league analysis using simulation is the 2005 season of Australian Rules Football. There are 16 teams, and tens of thousands of fans turn out to watch each game. There is a keen interest in forecasting the outcomes of games. Although we could use the same approach here as with soccer, instead we use a different strategy. Teams tend to score about 95 points with a SD of about 25 in games. The EQ hypothesis in this case allows us to ignore which particular pair of teams is playing – we just simulate two scores and see which is greatest. Ties are rare (about 1 percent) and we can ignore them. A good summary of the teams' performance in a season is the ratio of points "for" to points "against." We use this rather than the actual league points since it is expected to be more sensitive to team quality differences, if they exist, than the league points earned (Clarke, 2005).

The best teams should have the largest for/against point ratios. The ratio of max(for/against)/min(for/against) compares the best team with the worst team. The question of interest is whether the observed ratio of ratios is greater than what one would expect if all teams were of equal quality. "EQ" here means that all teams have the same distribution of points attained in any given game. The chart below shows the result of the 16 team league under the Gamma(mean=95, sd=25) distribution of points for each team. A Normal model could have been used but there is evidence of a right skew in points earned so the Gamma is preferred. Specifying the mean and standard deviation sets the Gamma parameters. The full league consisted of 22 games, so the round was not complete; however, this is not a problem for our simulation using the EQ hypothesis.





The top team, Adelaide, had a for/against ratio of 1.36 and the bottom team Carlton had a for/against ratio of .75, so the ratio of ratios is 1.36/.75=1.81. The simulation of 100 seasons shows top values in the range 1.5-1.7 (1.72 in the chart above) or a little less than what was observed in the 2005 season. So again, we see some evidence of team differential quality in this league. However, the differential is quite small considering that we used such a sensitive measure. So our conclusion in this case is that the EQ hypothesis explains most of the variation in the team ladder. When one compares the 2004 season with the 2005 season, one gets a very different ranking of the teams, which support the credibility of the EQ hypothesis in this league.

DISCUSSION

These examples aim to accomplish several goals when used in the classroom:

- They illustrate one kind of demonstration that could be used to interest students in statistical thinking.
- They show that these simple methods can reveal some surprising and useful facts about current events in professional sport.
- They show that modern statistics involves simulation and graphics in ways that are understandable by statistical novices and even the lay public.
- They put into a familiar context abstract ideas like “probability,” “distribution,” “variability,” and “hypothesis testing.”

We have argued in Weldon (2005) that simulation and graphics are becoming an increasingly important part of undergraduate statistics courses. For example, the discussion in elementary courses of nonparametric smoothing, resampling, and robustness is a fairly new development that is now feasible with modern software. The suggestion here is that sports contexts involving relative team quality in a league is a useful one to introduce some modern statistical methods in applications of current interest to many students. While many papers and web pages contain sports examples for teaching statistics, and some include extensive use of simulation, such as Andrews (2005) and Cobb (2005), the context described here is new, adaptable to many team sports, and rich in the potential to demonstrate the surprising consequences of randomness.

Of course simulation and graphics, and in fact most modern statistical methods, require the use of statistical software, and so software needs to be used at least by instructors, and ideally by students as well. Most universities support the subscription to the software needed for teaching. But when students gain employment outside the university, the availability of the software they know may not be supported. However, a back-up strategy in this situation is to use software that is freely downloadable, such as *R* (*R* Development Core Team, 2005). The programs used for the examples described in this paper are all programs in *R* and are freely available from the web page www.sfu.ca/~weldon. *R* is not so easy to learn to program, but prepared programs can be run, and modified, fairly easily. In addition, when students want to make more serious use of *R*, they will have some experience with its broad capabilities. Of course, *R* is not the only freely available software source: See the wonderful compilation produced and maintained by John Pezzulo (2005).

Do the examples illustrate the traditional material of statistics courses? There is certainly use of the standard deviation in measuring variability, use of probability in the EQ hypothesis idea, use of the idea of distribution in the dotplots, and the whole EQ hypothesis approach is exactly the conventional hypothesis test, although executed in a modern way. The use of simulation and graphics as bone fide methods for analyzing data is not traditional but is an approach that is gaining popularity, partly because it is so readily understood, a point reinforced by Cobb (2005). So these examples are typical of the kind of deviation from a traditional course that nudge it in the direction of a modern computer-based discipline.

SUMMARY

A suggestion has been made for the use of pre-programmed simulations in introductory courses. The context of sports team comparisons has been proposed since it can be described with reference to "live" data at the time of instruction, in a context of interest to many students. The results of the analyses are somewhat surprising given the general illusion of quality differential that seem to be portrayed by widely varying win-rates in a sports league. This should help to convince students of the vitality of our discipline, and simultaneously teach them some ideas that they can use right away in their real life.

REFERENCES

Andrews, C. (2005). The ultimate flow. *Journal of Statistics Education*, 13(1), www.amstat.org/publications/jse/v13n1/andrews.html

Clarke, S. R. (2005). Personal communication.

Cobb, G. W. (2005). The introductory statistics course – A sabre tooth curriculum? Presentation at USCOTS, Columbus, Ohio. http://www.causeweb.org/uscots/plenary/

Hutchison, B. (2004a). English Premier League Final Standings. Reuters, http://sportsillustrated.cnn.com/2004/soccer/09/30/englandtable.0304/.

Hutchison, B. (2004b). English soccer – All time league winners list. About, Inc., http://worldsoccer.about.com/od/england/a/engleaguewins.htm.

NBA. (2005). Division standings. NBA Media Ventures, LL, http://www.nba.com/standings/team_record_comparison/conferenceNew_Std_Alp.html.

Pezzullo, J. (2005). StatPages.net website. http://statpages.org/javasta2.html.

R Development Core Team (2005). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria.

Weldon, K. L. (2005) Modern introductory statistics using simulation and data analysis. *Proceedings of the 55th Session of the International Statistical Institute*, Vol. LI, Sydney. Voorburg, The Netherlands: International Statistical Institute.