Statistical Analysis of Soccer League Differences

SHEN Feng, MAO Shaokun, ZHENG Zequ
Department of Mathematics, City University of Hong Kong

Abstract
This project utilizes statistics of on-pitch performances of soccer teams from three different leagues in Europe to do analysis. One-Way ANOVA, Two-Way ANOVA and Tukey’s Test are employed to test the style differences between teams from different countries. In addition, the variable month indicating the time during which the match is played is taken into account to examine the influence of schedule to different leagues. The results show that every league has quite different tactic styles and the schedule has some impacts to different leagues considered in this project. However, the arrangement of schedule has different impacts at different periods of time to all the leagues and to specific technical statistics, such as tackles attempted and aerial duels.

Introduction
Tactic has become the core of modern soccer. This project targets to examine the relationship between the tactic style of a club and the country where it comes from. Statistics of on-pitch performances, including goals, shots, passes, dribbles, tackles attempted and aerial duels are taken into account to make the comparison.

Preliminaries
Data Collection & Assumption
On-pitch performance statistics from matches of English Premier League, Spanish La Liga and Germany Bundesliga in 2014/15 season are collected from WhoScored.com. Data from 380 matches of Premier League, 380 matches of La Liga and 366 matches of Bundesliga is included in this project for analysis.

Assumption 1 The level of strength of clubs from the three leagues included in the test are similar.
Assumption 2 The statistics for 2014/15 season are a good representative of all the recent seasons.
Assumption 3 The performance of a team in a specific match does not influence its performance in the other matches.
Assumption 4 The distribution of total number of goals, shots, passes, dribbles, tackles attempted and aerial duels follow normal distribution $N(\mu, \sigma^2)$, where $\mu$ is estimated by the sample mean $\bar{\mu}$ and $\sigma$ is estimated by standard deviation $s$.

Methods
One-Way ANOVA
One-Way ANOVA is employed to test the impacts of league in this project. According One-Way ANOVA model, within-group variation (SSW) and among-group variation (SSA) are required.

We will reject $H_0$ at the $\alpha$ level of significance if

$$F = \frac{MSA}{MSW} > F_{\alpha(r-1,n-r)};$$

The One-way ANOVA table can be constructed as below:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Square</th>
<th>df</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>SSA</td>
<td>$r-1$</td>
<td>$MSA = \frac{SSA}{r-1}$</td>
<td>$F = \frac{MSA}{MSW}$</td>
</tr>
<tr>
<td>Within groups</td>
<td>SSW</td>
<td>$n-r$</td>
<td>$MSW = \frac{SSW}{n-r}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>$n-1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Two-Way ANOVA
In Two-Way ANOVA, we totally have three factors (Effects of factor A (League): $\alpha_i$, effects of factor B (Month): $\beta_j$, effects of interaction of factor A and factor B: $\alpha\beta_{ij}$) to be tested.

The Two-Way ANOVA Table can be constructed as below:

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Square</th>
<th>df</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>SSB</td>
<td>$a-1$</td>
<td>$MSA = \frac{SSB}{a-1}$</td>
<td>$F = \frac{SSB}{SSA}$</td>
</tr>
<tr>
<td>Factor B</td>
<td>SSB</td>
<td>$b-1$</td>
<td>$MSB = \frac{SSB}{b-1}$</td>
<td>$F = \frac{SSB}{SSA}$</td>
</tr>
<tr>
<td>Interaction $SS(AB)$</td>
<td>$(a-1)(b-1)$</td>
<td>$ab(n-1)$</td>
<td>$MSE = \frac{SS(AB)}{(a-1)(b-1)}$</td>
<td>$F = \frac{MSE}{SSA}$</td>
</tr>
<tr>
<td>Error</td>
<td>SSE</td>
<td>$ab(n-1)$</td>
<td>$MSE = \frac{SSE}{ab(n-1)}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>$n-1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tukey’s Range Test
Tukey’s Range Test is applied simultaneously to the set of all pairwise comparisons of factors (Month and League) and their interaction.

Conclusion
- Premier League, La Liga and Bundesliga show different tactic styles and the result contradict with common impressions toward their soccer styles.
  - Bundesliga plays with the fastest tempo.
  - Premier League’s clubs try to control the pace of matches through increasing passes rather than merely depend on physical engagement.
  - La Liga adores personal skills of breakthrough more.
- The schedule (Winter Break) has little impacts to clubs from different leagues.
- Teams show less competence after Christmas but become more competent during the home stretch of the season.

Reference

Acknowledgement
This work originates from a course project of MA3518 Applied Statistics at CityU (2015), instructed by Dr. ZHOU Xiang.