Erasmus University Rotterdam

Master Thesis

The effect of a crowd on the home advantage in European football; An application and simulation of the permutation test on small specific samples

Antonie Louter (474356),

Supervisor: Dr. N.W. Koning

Second Assessor: R. Paap

September 27, 2021

Abstract

This paper investigates the use of the permutation test to determine significance of the effect of the presence of a crowd on the home advantage in football and its comparison to existing literature. This is done by performing a simulation study. This showed that the size of the testing methods in existing literature is distorted for small sample sizes, whereas the size of the permutation test remains valid in all sample sizes. However, this is accompanied by a loss of power. Furthermore, we have found that it is easier to detect the home advantage in a setting with equally strong teams in comparison to a setting with unequally strong teams and that the goal difference between two teams is a better measure for the home advantage than the amount of points won. Additionally, the permutation test is applied in the testing of the crowd-effect on the home advantage in multiple European leagues, for both the average as well as team-specific. Both the Premier League and La Liga showed a significant crowd-effect on the home advantage, where the crowd-effect for the 'top' teams seemed to be the driving factor.

Contents

1 Introduction

In professional sports there is only one thing that is important: Winning. The clubs, and everyone that is connected to these clubs will do anything for this. Supporters from all over the world cheer their favorite teams on and hope to lead them to victory. In the 2018/2019 season of the Premier League, the last season that supporters were allowed in the stadiums during the entire year, the average attendance was 38.188. All these supporters hope to give their team just that little edge in the game. The influence of these supporters, and in particular their presence or absence, on the home advantage is the subject of this paper. This will be researched with the use of a permutation test.

The home advantage, which we will refer to as HA from now on, is defined by Neave and Wolfson (2003) as the consistently better performance of professional teams in multiple sporting contexts when they play at home. This better performance usually results in, for example, the winning probability, amount of goals scored or the amount of chances created. As the performance is hard to measure directly, this paper will focus on the amount of points won and the goal difference in home- and away games. This is the same as in Jiménez Sánchez and Lavín (2021).

The HA is a often researched topic, for example as in Courneya and Carron (1992), which gives a literature review on the HA in multiple sports. The papers written on this topic are countless, but some examples are Gómez et al. (2011), Jones (2013) and Nevill and Holder (1999), which all show a positive influence of playing at home across all sports.

Jamieson (2010) and Pollard and Pollard (2005) show that the HA in football seems to be one of the largest among all sports. However, the findings of the papers written on the topic do not all agree on the outcome. There are differences found for leagues across different countries, as is shown in Pollard (2006), while Pollard and Gómez (2014) and Pollard et al. (2017) show that there is also a difference between women-and men football leagues. These showed a bigger HA for the men's leagues. This, while keeping in mind that the attendance at men's games is higher than for women's games, might be another indication that the crowd influences the outcome of matches.

Although there have been done a lot of theory studies on this subject it is hard to actually measure it in an experiment. This is mainly due to the difficulty of recreating good environments to

test the theory. This has to be done either by simulating an event or organizing an official sporting event where an audience is not allowed to enter(Jiménez Sánchez and Lavín, 2021). This would be a good way to separate the HA from the possible incidence of the audience. If one, for example, would consider a "normal" season with switching attendances to measure this you might develop some bias in your estimates. For big games, or evenly matched games, there will be a higher crowd attendance, so one would expect a higher winning probability for these games. However, the opponent will also be of higher quality, which would decrease the winning probability. In the normal setting it would be very hard to determine exact effects. Johnston (2008) states that these kinds of comparisons are insensitive to variation within leagues and have very low power.

The reason that this subject is especially good to be researched at this point in time is the COVID-19 pandemic. Due to the pandemic, there are no more supporters allowed in the stadium and it offers an excellent opportunity to research the influence of the crowd. This has started in March 2020 and from this period on the competitions, while shutting down for a while, have restarted and there are already thousands of games played in the meantime. We will refer to these games as *ghost games*. This gives us a perfectly randomized experiment, as teams can't choose themselves whether they let supporters enter their stadiums or not. With the assumption that not much else has changed in football, "there are no new rules, no major trends and no new leagues added in football", it is possible to isolate the effect of the crowd on the HA.

Multiple papers have been written with the use of these ghost games already, for example Fischer and Haucap (2020) speaks of the influence of crowd attendance in the German top 3 leagues. They found a decrease of HA for the ghost games, but only in the highest division, where the amount of supporters is highest. In the other leagues the effect was not found to be significant. Jiménez Sánchez and Lavín (2021) has also made use of the ghost games played during the COVID-19 pandemic and investigates the effect of crowd size in the top 8 leagues. Only in the top leagues of Spain and Germany the results turn out to be significant.

Other examples are Bryson et al. (2021), Cueva (2020) and Sors et al. (2020), which used the bias of refereeing decisions as the driving factor behind the HA and found a significant decrease in HA for games without a crowd. Deutscher and Winkelmann (2020), Dilger and Vischer (2020), Ferraresi et al. (2020), Fischer and Haucap (2020), McCarrick et al. (2020), Reade et al. (2020),

Scoppa (2021), and Tilp and Thaller (2020) all use the difference in points and goals scored as indication for the HA, which found a significant decrease in HA for games without a crowd. This is a lot of papers for such a small period of time. However, all these papers have something in common, which is the use of parametric testing, such as the t-test. This is a big trend in the testing of the crowd-effect on the HA and the HA as a whole.

However, it is shown in Tversky and Kahneman (1971) that people tend to believe that small samples are overly representative for showing that two populations are significantly different. They also show that in practice people need far larger samples to obtain a good power for their test when they use a parametric test. The size of this test can be distorted as well. This is why we will consider the permutation test. The permutation test can even with a small sample give a test that controls size. Therefore, the permutation test would be very good to research certain smaller subsamples of games, such as Champions League games or matches between rivals, without losing their validity. This will be an addition to the literature and will give a more reliable conclusion about the influence of the crowd on the HA in football.

There is a paper written on this topic that does make use of permutations. This is Hill and Van Yperen (2021). However, this paper uses another way of permuting the observations, where the observations are not permuted under the null hypothesis. What they do is that they permute the values over the last four seasons, where there was still a crowd present in the stadiums. This is done to reduce year specific effects for the games with a crowd. These year specific effects are shown to be present by Clarke and Norman (1995). For all permutations they compare the means of variables as goals and points of the "new" sample of games with a crowd with the games without a crowd. However, this does not take variation of the games played without a crowd into account. Our way of permuting does do this. By permuting under the null hypothesis it is possible to swap observations between the groups of the games with a crowd and games without a crowd. This is completely different from what Hill and Van Yperen (2021) does.

The main question, with corresponding sub-questions, that will be answered in this paper is: *In what way can the permutation test be an addition to the existing literature on the testing of the effect of crowd size on the HA?*

- *Does the permutation test outperform classical tests, which assume normality, that correspond*

to relevant settings for the HA, in terms of power and size? And how does the use of goals and specific sub-sampling influence the according power and size?

- *In what way does the use of the permutation test in significance testing of effect of crowd size on the HA qualitatively differ to the results that are found in existing literature, which uses classical testing?*
- *How does the effect of crowd size on the HA differ for different subsets of games?*

The first question, where we ask if the permutation test outperforms classical tests that correspond to relevant settings for the HA, in terms of power and size, will be answered with the use of a simulation study, which we will generate ourselves. In this simulation study, we analyze the power of the permutation test and the t-test for multiple values of the HA. We generate a model for the goals using two independent Poisson distributions. One for the home team and one for the away team. This is done for set-ups including two teams, with equal or unequal strength, and a competition for multiple sample sizes. Goals as well as points will be used for the detection of the HA. We found an increase in power for the HA of equally strong teams, in contrast to that of unequally strong teams. The use of goal difference also showed an increase of power in comparison to the use of points. The t-test showed an increasing distortion of the size as the sample size decreased, where the permutation test remained valid throughout all sample sizes. We found a higher power for the t-test for small samples in comparison to a higher power for the permutation test in larger samples.

The second question, where we ask in what way the permutation test in significance testing of effect of the crowd on the HA qualitatively differs to the results that are found in existing literature, which uses classical testing, will be answered using data of the Premier League, La Liga and Bundesliga. We will use the permutation and t-test on this data. We found a significant effect for the HA, in the period played with a crowd for all 3 leagues and a significant crowd-effect for 2 out 3 leagues.

We want to know where this effect might come from, so we research this for individual teams of the Premier League as well. For all teams, we test the effect of the crowd on the HA individually. For this individual assessment we expect the permutation test to have the most difference with the t-test, due to small sample sizes. We use a Bonferroni-Holm procedure to correct for the multiple hypothesis testing. We find very strong evidence that there is a HA in the Premier League for games with a crowd and some evidence, which is not as strong, that the crowd is the driving factor

in this HA.

The third question, where we ask how the crowd-effect on the HA differs for different subsets of games, is answered by creating multiple subsamples of teams in the Premier League. These are respectively the top, middle and weakest teams. The permutation test is used to determine the significance of the crowd-effect. We find a clear difference between the top- and middle teams and the weakest teams, as the first two show a significant effect, whereas the weakest teams don't even show a crowd-effect somewhat close to a significant value.

2 Home Advantage

2.1 The Model

We start with defining how we set up our model for the outcomes of games and how the HA is included in our model. We make use of the goal difference between home- and away teams, further referred to as respectively GHT and GAT. Our most important requirement for the model is that the HA has a positive influence on the goal difference. In mathematical notation this yields $E[GHT - GAT|HA] > E[GHT - GAT|No\ HA]$.

Example 2.1. Poisson model

A possible, but not required, distribution for the goal difference is the Poisson distribution[1]. We use $\lambda_{1;T_1,T_2}$ and $\lambda_{2;T_1,T_2}$ to represent the scoring potential of the home- and away team respectively. We define GHT_{T_1,T_2}, the goals of home team T_1 against away team T_2 and GAT_{T_1,T_2} corresponds to the goals of away team T_2 against home team T_1. The distribution for this is as follows:

$$GHT_{T_1,T_2} \sim Poisson(\lambda_{1;T_1,T_2}), \tag{1}$$

$$GAT_{T_1,T_2} \sim Poisson(\lambda_{2;T_1,T_2}) \tag{2}$$

with $\lambda_{1;T_1,T_2}$ and $\lambda_{2;T_1,T_2}$ given as:

$$log(\lambda_{1;T_1,T_2}) = \mu_{T_1,T_2} + \gamma_{\text{HA}} + Attack_{T_1} - Defense_{T_2}, \tag{3}$$

$$log(\lambda_{2;T_1,T_2}) = \mu_{T_2,T_1} + Attack_{T_2} - Defense_{T_1}, \tag{4}$$

where μ_{T_1,T_2} is the intercept that can be used as a base rate for goals for a specific match-up and the $Attack$ and $Defense$ are respectively the attacking and defensive strengths of the teams. For ease we assume these distributions to be independent. This model is specifically used for the simulation study and explanation purposes and is not required for the testing of the HA and crowd-effect.

We are not interested in the exact size of the HA but merely its presence and therefore consider γ_{HA} to be either zero or non-zero.

[1]The model that is used is introduced in Dixon and Coles (1997). This model is able to simulate the amount of goals scored for home- and away teams, taking attacking strengths, defensive strengths, a possible base level of goals and a intercept for the HA into account.

2.2 The HA

We are not interested in the exact modelling of the outcomes, but we are interested in the HA. We consider the HA to be built from two-effects. One part of the effect is gained from the crowd in the stadiums, referred to as the crowd-effect, and the other part are the extra effects. This consists of advantage due to less travel times, which results in more rest, familiarity with the pitch and other extra factors. These extra effects are not the focus of our research so we will not dig too deep into this. We decompose the HA as follows:

$$\gamma_{\text{HA}} = E[GHT - GAT|HA] > E[GHT - GAT|\text{No HA}] = \gamma_{\text{Crowd}} + \gamma_{\text{Extra}} \tag{5}$$

For now, we want to test whether this HA is present in the leagues we use our research on. We only expect there to be an advantage when playing at home, so we consider one-sided testing. The according null- and alternative hypothesis that will be used are given by:

$$H_0 : \gamma_{HA} = 0, \tag{6}$$

$$H_a : \gamma_{HA} > 0. \tag{7}$$

We test this for two adjacent periods. The first period considers games played with a crowd and the second period considers the games played without a crowd.

However, it is not possible to observe the HA directly. We only observe the goals scored and the according outcomes in real life. We have assumed that the HA has a positive influence on the goal difference between home- and away teams. Therefore we may use the expected average goal difference between home- and away games as a measurement tool for the significance of the HA. This is further explained in Section 3.3.

The according hypotheses with this new statistic will be given:

$$H_0 : E[GHT_{T_1,T_2} - GAT_{T_1,T_2}] = E[GAT_{T_2,T_1} - GHT_{T_2,T_1}], \tag{8}$$

$$H_a : E[GHT_{T_1,T_2} - GAT_{T_1,T_2}] > E[GAT_{T_2,T_1} - GHT_{T_2,T_1}]. \tag{9}$$

The goal difference is a sufficient statistic for the points won in matches, an increase in expected goal difference leads inherently to an increase in expected points won. Teams are ultimately most interested in the amount of points won so we will use the expected points won in home- and away matches as a measurement for the HA too. The goal difference does provide us with more information than the amount of points won.

The outcomes will be used as supporting evidence for the presence of the crowd-effect. In Section 4 is explained how the testing of the crowd-effect is done and its according assumptions.

3 Permutation test

To test both the HA and the crowd-effect we will use the permutation test. The idea behind this test and how this test works is explained in this section.

3.1 Exchangeability

There is one crucial assumption that has to be made for the permutation test to work and that is the assumption of exchangeability. This is that, for a given set of variables, the joint probability distribution of the samples does not change if the groups are rearranged.

In our case this means that the goal difference of home- and away matches, GD, needs to be exchangeable. In mathematical notation this yields the following: for any permutation $P_k \in P$, where P is given as all possible permutations, $GD =^d P_j GD$, where the symbol $=^d$ denotes equality of distributions and GD is the goal difference. This means that the goals can be seen as exchangeable if their joint distribution is invariant with respect to permutation(Winkler et al., 2014).

To show that this holds in our case we consider a sample of only 2 games, particularly two matches played between the same teams alternating between playing at home. We will make use of our distribution for GHT and GAT mentioned in (1) and (2). Both T_1 and T_2 can have the values {1,2}. Under the assumption that there is no HA we can say the following:

$$E[(GHT_{1,2} - GAT_{1,2}) - (GAT_{2,1} - GHT_{2,1})|No\ HA] = 0. \tag{10}$$

Thus if there is no HA both matches we expect both matches to have the same outcome. This makes it possible to swap both matches without changing the distribution of the goal difference. Therefore we can assume, and is also required for the permutation test, that under H_0 we can swap home- and away matches freely.

However, every match has its own specific characteristics. A game between high-level teams could largely differ to a game played between low-level teams. Therefore we require that the swapped games are always between the same teams. In this way it is possible to limit the influence of the

match specific characteristics as much as possible.

Example 3.1. Poisson model with two matches

If we would swap the match Arsenal(A)-Aston Villa(V) with Tottenham(T)-Arsenal this could lead to incredibly biased results. Using (1) and (2), and keeping in mind that Tottenham and Aston Villa have different attacking and defensive strengths, we get the following:

$$E[GHT_{A,V} - GAT_{A,V}|\text{No HA}] - E[GHT_{T,A} - GAT_{T,A}|\text{No HA}] \quad (11)$$

$$=_1 (E[GHT_{A,V}|\text{No HA}] - E[GAT_{A,V}|\text{No HA}]) - (E[GAT_{T,A}|\text{No HA}] - E[GHT_{T,A}|\text{No HA}]) \quad (12)$$

$$=_2 (Attack_T - Attack_V) + (Defense_T - Defense_V) + (\mu_{A,V} - \mu_{A,T}) + (\mu_{V,A} - \mu_{T,A}) \quad (13)$$

$$\neq 0, \quad (14)$$

where $=_1$ holds due to the assumption of independence of the goals scored by the home- and away team and $=_2$ is obtained by using E[X]=λ if X $\sim Poisson(\lambda)$ and crossing out double terms.
Even in a case without a HA, you would still expect a difference in outcome between the home- and away game and therefore make it a biased estimator.

However, if Arsenal would play, for example, Aston Villa in both the home- and away game, all the terms would be equal to zero and therefore make it an unbiased estimator.

We have to assume one more thing: The swapped games are independent. If the games would be dependent of one another, a win in the first game might influence the outcome in the second game, which would give biased results. These will be further described in Section 3.2 and expanded for the use of multiple matches.

For the use of points, the distribution is slightly different. Seen from a home team's perspective, the division of points is over the set {0,1,3}. If the HA increases the probability mass moves from left to right in the set. In mathematical notation this yields $E[Points_{HT}|\text{HA}] \geq E[Points_{HT}|\text{No HA}]$. However, under the null hypothesis we assume the distributions of the home- and away games to be the same and therefore make it possible to swap outcomes as well.

3.2 Assumptions

In the last section we considered swapping two games. However, with only two games we are not able to test anything. Therefore, we must consider multiple matches. Unfortunately not all teams are exactly the same and show a perfect indication whether the HA is present. We consider an environment with multiple teams, also known as a competition in the football world. It would be the easiest to determine the HA if we considered a large sample of games for just two teams, but this would mean we needed games from years and years back and this is not representative for the situation now.

With the use of match results from multiple teams across a competition some other issues do come our way. There is one assumption that has to be made and that is the interdependence of the matches. Even though we only swap the matches of the same team it could be possible that, for example in a title race or a relegation battle, the games of your opponents influence the way you play. We assume this interdependence of matches to be negligible. This assumption ensures that we can swap the home- and away team for all games separately. This leads to an average HA on a competition level. A different outcome in one game would alter the outcome of a different game, so swapping two outcomes would lead to a new sample. This assumption is not entirely perfect, as some games might be correlated, but due to the large amount of games played within one competition we consider this reasonable.

Due to the randomization of the game schedule we may assume that differences in strength of teams throughout the season, due to injuries or form, do not lead to biased results.

For classical testing the assumptions are different and require normality. We will check this, with the help of the Shapiro-Wilk test, introduced by Shapiro and Wilk (1965). However, due to possible small sample sizes it can be possible the test does not reject even when the samples are in fact not distributed normally.

3.3 Test statistic

The test statistic that we will be using is given by $T = (\overline{GHT}_{T_1,T_2} - \overline{GAT}_{T_1,T_2})$ - $(\overline{GAT}_{T_2,T_1} - \overline{GHT}_{T_2,T_1})$. This is the average observed difference in means between the goal difference at home

and the goal difference away. We have already shown it is possible to swap the home- and away under the null hypothesis in (10). Under the alternative we have the following:

$$E[(GHT_{1,2} - GAT_{1,2}) - (GAT_{2,1} - GHT_{2,1})|\text{HA}] \tag{15}$$

$$=_1 E[(GHT_{1,2} - GAT_{1,2})|\text{HA}] - E[(GAT_{2,1} - GHT_{2,1})|\text{HA}] \tag{16}$$

$$>_2 E[(GHT_{1,2} - GAT_{1,2})|\text{No HA}] - E[(GAT_{2,1} - GHT_{2,1})|\text{HA}] \tag{17}$$

$$>_3 E[(GHT_{1,2} - GAT_{1,2})|\text{No HA}] - E[(GAT_{2,1} - GHT_{2,1})|\text{No HA}] \tag{18}$$

$$= 0, \tag{19}$$

where $=_1$ holds due to the arithmetic rules of expectation and $>_2$ and $>_3$ hold due to our definition of the HA.

A test-statistic that is greater than zero can therefore be seen as an indication of the presence of a HA. Now that we have ensured we can swap the home- and away team for all matches within the competition, we start with the permutations. This is permuting the home results with the away results. This is done for all permutations. This leads to a new group composition with all new test statistics. The distribution of these recomputed statistics is used to compute a distribution for the entirety of the group. The proportion of recomputed statistics exceeding the observed test statistic is the p-value of the permutation test. In mathematical notation this yields, for real observed test statistic T, the following distribution for all possible values t.

$$J(t) = \frac{1}{K} \sum_k I(T_k \geq t), \tag{20}$$

where K is the number of permutations, T_k is the test statistic for a given permutation and t are all possible values. I is an indicator function that is equal to 1 if it is true and 0 otherwise. For the 'real' observed test statistic T this yields a corresponding p-value. All p-values are discrete and will never be zero, as only $0 * \frac{1}{K} = 0$. For $0 < \alpha < 1$, the permutation test rejects for α if T is greater than the the $1 - \alpha$ quantile of $J(t)$.

3.3.1 Adjusted rejection probability

However, in the case of discrete data, such as goal difference or amount of points, it is common to obtain duplicate values of the test statistic, e.g. a permutation only swaps one game which has

the same outcome. This might lead to trouble as the permutation test will be conservative. It will not reject as much as it would need to. Therefore we consider an approach introduced in Hoeffding (1952), which compensates for this low rejection rate. When T is equal to the $1-\alpha$ quantile of $J(t)$, we consider a certain rejection probability. This is given by b:

$$b = \frac{K * \alpha - S^+}{S^-} \tag{21}$$

Where S^+ and S^- are given by,

$$S^+ := \#\{k \in K : T(k) > T\}, \tag{22}$$

$$S^- := \#\{k \in K : T(k) = T\} \tag{23}$$

The downfall of this approach is that it makes the test randomized in the boundary case, instead of exact, but it does make the results more accurate.

The number of permutations is in practice not always equal to the maximum amount of permutations, which is 2^N, as this would lead to an enormous amount of processing time, but the permutations are randomly selected. A frequently used amount of permutations is 1000. This will lead to an approximate distribution. This is also the amount that is used in this paper.

3.4 Simulation Study

To answer our first research question, where we ask if the permutation test outperforms classical tests that correspond to relevant settings for the HA, in terms of power and size, we consider a simulation study. We use the model mentioned in Section 2. We consider multiple simulation setups with different values for sample size, attacking strengths, defensive strengths, amount of teams and size of the HA. On top of the comparison of the permutation test and the t-test, our aim is to further analyze the possible benefits of the use of goals and a level playing field, in terms of power. For our tests, we use the paired permutation test and the paired sample t-test. The paired sample t-test has an extra assumption that needs to be made. This is the assumption of normality. For large sample sizes we expect this to hold, but for smaller sample sizes this might not, which would make the test invalid. Therefore, we expect the biggest difference between these two tests for small sample sizes.

Now we get to the simulation conditions. We are only interested in the significance of the HA and how many times it will be significant out of all simulations. We use a significance level of $\alpha = 0.05$. For all the different set-ups of simulations that we have we consider 10000 simulations. On top of this, we use 1000 permutations for the permutation test, for all 10000 simulations. We set the HA as equal for all teams and let it increase from 0 to 1 with increments of 0.05. Thus, this leaves us with 21 times 10000 simulations for every set-up.
For all set-ups we are interested in the power that corresponds to our tests. We define the estimator for the power under the alternative hypothesis as follows:

$$Power \approx \#Rejection/\#Simulations, \tag{24}$$

where the $\#Rejection$ is equal to the amount of times the test rejects for α for the simulations. For example, in a case of 2000 rejections, $Power \approx 2000/10000$=0.2. We use the power under the alternative hypothesis as we modify the magnitude of the HA ourselves and therefore know it is present. Next we are interested in the precision of the power. We know that $\#Rejection \sim Binomial(\#Simulations, p)$, with p equal to the rejection probability. The normal approximation of the binomial distribution ensures that the variance interval of $\#Rejection$, i.e. the power, is equal to $\frac{p(1-p)}{\#Simulations}$. This ensures that the variance interval is around 0.01 for 10000 simulations. So for example, for a power of 0.2, this leads to a 95% confidence interval of the power between 0.19 and 0.21.

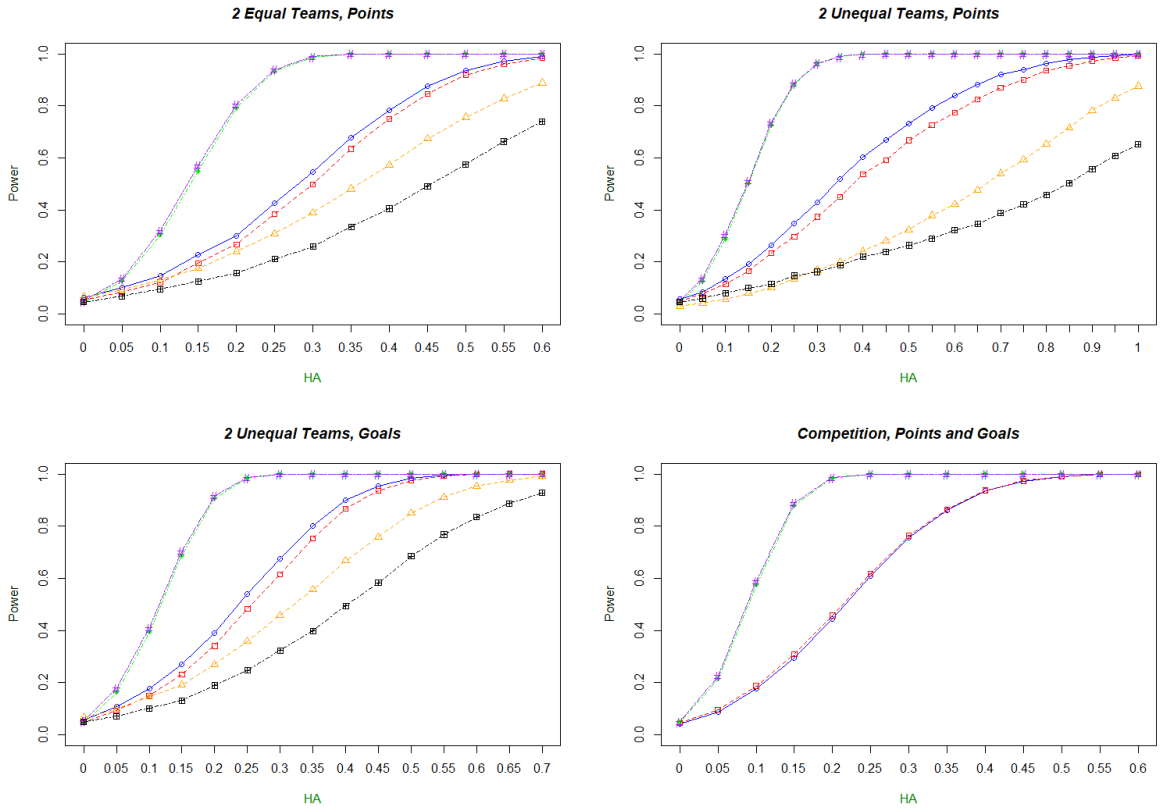
We expect and want a size of around 0.05 for a setting without a HA. This is due to the significance level. We test the HA over the whole sample, so not individually. The value of the strengths given below are equal to the attacking strengths as well as the defensive strengths. The samples are defined below:

- 2 teams
  - Strength: $\{1, 1\}$
    - Sample Size: 10, 20, 100
  - Strength: {1, 0.5},
    - Sample Size: 10, 20, 100
- 10 teams
  - Strength: {1, 0.78, 0.56, 0.33, 0.11, -0.11, -0.33, -0.56, -0.78, -1}

∗ Sample Size: 90

For all samples we investigate the influence of the HA on the the average goal difference and for the average amount of points. The power that we found for the aforementioned levels of HA are shown in the graphs below.

Figure 1: Power vs HA



Notes: Each graph represents the outcomes of the aforementioned samples. The top left, top right and bottom left graph all show the outcomes for a set-up with two teams, differentiating between the use of goals and points and the strengths of the teams. This differentiation is shown in the title of the graphs. Please note the difference in scale for the HA. The orange line represents the outcomes for a t-test in a 10-game sample. The black line represents the outcomes for a permutation test in the 10 game sample. The blue and red are respectively the outcomes for a t- and permutation test in the 20 game sample. The green and purple line are respectively the t- and permutation test for the 100 game sample. In the bottom-right graph this is slightly different: The blue and green line are respectively the t- and permutation test on the outcomes of the game, whereas the green and purple line correspond to the t- and permutation test for the goals scored in the game

We start with the size and power analysis of the two teams that are equally strong. All tests show a size of around 0.05, except for the t-test in the 10- and 20 game sample. This shows a size

of around 0.07 and 0.065, which is an overestimation of the size for this particular test. This is significantly different from 0.05.
However, for the 20 game sample, we see a difference in power between the two tests. The t-test shows a significant higher power than the permutation test, with a maximum significant difference of 0.05. For the 10 game sample this difference in power is even greater, up to 0.1 for some values of the HA. This is most probably due to the fact that the outcomes can only take 3 values and therefore lead to many duplicate test-statistics in smaller samples.

For the 100 game sample we see a slightly larger power for the permutation test, but no distortion in the size. We expect the normal approximation to be more precise here. Therefore, the assumptions of both tests hold and the tests will both be valid. For the model setup described in Section 2 the average Premier League HA is close to 0.3. For a HA of 0.3, the power is almost equal to 1, which means we can easily detect it.

Next up is the size and power analysis of the points won in an unequal sample. The size of the t-test is only 0.03 for the 10-game sample. The power of the permutation test in this sample exceeds the power of the t-test until a HA of 0.3. The differences between the t-test and permutation test in the 20- and 100 game sample are similar to the sample described before. The difference between equal and unequal teams is that in a sample with unequal teams, the power is 0.2 lower for the same levels of the HA . This is an indication that there might be some room for improvement in testing the HA. This is done by creating subsets of teams with the same strength and testing the HA within these subsets. This will improve performance in terms of power and size.

Our first improvement is the use of goals, shown in the bottom-left graph. The power has increased significantly, in comparison to the use of points. The power and size of the t- and permutation test are very similar in comparison to the use of goals. However, the difference in power between the two tests has declined and the size of the t-test for the 10 game sample is 0.07, in comparison to 0.03.

In the final graph we see the competition setup. It shows a very similar form as the 100- game samples mentioned before. Yet again, the amount of goals scored provides us with a higher power than the amount of points. There is a small difference between the power of the permutation test and the power of the t-test, where the power of the permutation test is around 0.015 higher. Even

though this is a small difference, we can still conclude that this is significant due to the aforementioned precision of the power.

From all these simulation set-ups there are a few things we can conclude. The HA is far better detectable in a level playing field. We use this in our advantage by creating subsets with equally strong teams. Regarding the use of points, as the sample size got smaller we observed a larger difference in power between the two tests, where the t-test showed a higher power. With the use of goals this difference declined but was still present. The t-test does show a distortion in the size for 10- and 20 game samples. The permutation test shows it is reliable in all settings, which even showed a higher power than the t-test for larger samples. In the composed subsets, which will have smaller sample sizes, the t-test will be invalid, whereas the assumptions for the permutation test will still hold.

The use of the goal difference shows a significant increase in power for all values of HA in comparison to the use of points. For our next section, we definitely take goals into account for our testing of the HA.

4 Crowd-effect

Now that we have shown how we will test the HA and how the permutation test performs in testing the HA, we are interested where this is HA exactly comes from. As mentioned before, we consider it to be build up out of two effects, the crowd-effect and extra-effects. Here we are mostly interested in the crowd-effect on the HA. We assume that the crowd-effect on the HA is the only difference between the period played with a crowd and without a crowd and therefore that the extra-effects remained equal in the two periods..

$$\gamma_{\text{Extra;Crowd}} = \gamma_{\text{Extra;Empty}}, \tag{25}$$

which ensures, due to the definition in (6) that testing the difference in HA between the two time periods is the same as testing the crowd-effect on the HA. This leads to new hypotheses. These are given by:

$$H_0 = \gamma_{\text{HA;Crowd}} = \gamma_{\text{HA;Empty}}, \tag{26}$$

$$H_a = \gamma_{\text{HA;Crowd}} > \gamma_{\text{HA;Empty}}, \tag{27}$$

where the null hypothesis implies that there is no effect of the crowd on the HA.

4.1 Assumptions

This is once again tested with the use of points won and goal difference for all matches. For this we use the permutation test. However, the exact implementation of the permutation test will be different. Under this new null hypothesis we need to make new assumptions. The assumption of exchangeability is altered to outcomes played within different time periods. Under the new null hypothesis, which states that the HA is the same in both periods, we can swap matches played with a crowd with games played without a crowd, due to the crowd-effect being zero. For this to hold we do have to assume that there haven't been changes in other possible factors that might influence the HA.

The way that the matches are swapped, are different in this case. For the testing of the presence of a HA within each period it is possible to swap the outcomes for each home- and away match between two teams, i.e. the home- and away team are paired. In the case of two different seasons, this is slightly different. We swap matches played between two particular teams with a crowd and matches between the same teams played without a crowd. This sometimes requires us to remove games, due to an unequal amount of games played with a crowd in comparison to without a crowd.

5 Club specific HA

We are interested in individual testing of the HA too. This will be done with the hypotheses described before, but now on a club level. We consider all clubs individually with their corresponding home- and away games. This will only be done for the clubs in the Premier League. Which clubs we exactly use is described in Section 6.1.

For this individual testing, some other issues arise. For the testing of the HA within each period, we run into unequal sample sizes. So, the amount of games played at home are not the same as the amount of games played away. This is only the case for some teams. Our solution to this problem is removing the games played that are played furthest from the separation time. For the sample that is played with a crowd this would mean the games that are played earliest in terms of date and for games without a crowd this would mean games that are played latest in terms of date. After removing these games we are able to swap the remaining matches.

There are some assumptions we do have to make for this to hold. These are the same assumptions as mentioned before, but slightly altered to an individual perspective. Firstly, the HA for an individual team is constant within each period. On top of this we need to assume that the HA of the opposing teams is on average equal. Due to the randomization of the schedule and every team playing (almost) every other team this would be the league's average HA.

5.1 Club specific crowd-effect

The same problem occurs for the individual team testing of the crowd effect as for the testing of the full competition. The amount of games played with a crowd and without a crowd are different. This is therefore resolved in a similar fashion as for the comparison between the HA for the full competition. This will make use of the assumptions mentioned for the competition setup, but altered to an individual perspective.

5.2 Multiple hypothesis testing

We are interested in two things, the significance of the HA and crowd-effect for the entire competition and for the teams in particular. Therefore we have a joint null hypothesis, for all teams together, and simultaneous null hypotheses, which are for all teams separately. Suppose that the joint null hypothesis is rejected, which would be a large indicator for the presence of an HA or crowd effect. In order to find which teams show a presence of HA, we would like to know which individual hypotheses can be rejected. However, testing a finite number of individual hypotheses simultaneously leads to a problem. This is called a multiple testing or simultaneous inference problem. This is described in Chapter 9 of Lehmann and Romano (2006). If all individual hypotheses are tested at significance level α, then the probability of a false rejection of at least one of these hypotheses increases very quickly for the amount of hypotheses. For our case, where the number of teams and thus the number of hypotheses is 15 and α is equal to 0.05, the probability of at least one false rejection when all individual rejections are actually true is equal to $1-(1-0.05)^{15} \approx 0.54$. This is called the family wise error rate(FWER). There are methods to control for the FWER.

The method that we will be using is a stepdown procedure called the Bonferroni-Holm method, introduced in Holm (1979). This method ensures that FWER $\leq \alpha$. There are two reasons we have

chosen this method. It is uniformly more powerful than the Bonferroni correction and it does not require assumptions of independence of the individual tests. Other methods do require this, but due to the teams playing against one another whilst being tested individually, we don't expect this to hold.

Sometimes it is not possible to reject any of the hypotheses separately, but we still might suspect that one of the hypotheses need to be rejected. Then we can use the Poisson distribution. This test relies on the following: If we find substantially more p-values smaller than the original significance level, this suggest that some tests might be significant. Based on the Poisson distribution with $mean = Teams * \alpha$, it is very unlikely that we see, for example, six initial rejections. Even though we don't know which hypothesis to reject in particular, this would still be evidence of a presence of the HA or crowd effect. This method is not perfect but does provide us with extra evidence.

6 Application

In this section we discuss the results we have found and whether or not these results can be deemed valid. First off is the significance testing of the HA within each period and the crowd-effect. We start off with the three full competitions, the Premier League, the La Liga and the Bundesliga. Then we discuss the results of the Premier League teams individually and finally the created subsamples.

6.1 Descriptive statistics

All three datasets consist of all games played in the seasons 2018/2019, 2019/2020 and 2020/2021. However, not all games are included. Only the teams that have featured in all three seasons will be kept in the dataset. All matches of the other teams are removed. This is done to create a more balanced dataset, where the teams are more equally strong and there are less team-dependent effects throughout the years. The teams that are included now form the core of the leagues. The simulation study showed that this leads to a better detection of the HA. This also makes it possible to test the same teams individually as within the full competition.

Included key variables are the home team, the away team, the goals scored by the home team, the goals scored by the away team and the overall result. As our variable of interest is the crowd, we have separated the games played with a crowd from the games without a crowd. The description of the key variables and the corresponding average points, winning percentage and percent of points

won in home- and away games can be found in a table in the Appendix.

To gain a better insight in team-specific effects within the competitions, we chose to investigate the Premier League in particular as the difference between the two periods in the average goals scored and the average points won is most evident here. The descriptive statistics of each team can be found in the Appendix.

6.1.1 Subsamples

We compute three different subsamples. These are based on the average amount of points won by a team in the games played without a crowd. This is not a perfect measure of equal strength, but it comes closest to it. The three different subsamples are the top five teams, the middle five teams and the bottom five teams. These are as follows:

- Top teams: {Chelsea, Liverpool, Manchester United, Manchester City, Tottenham}
- Middle teams: {Arsenal, Everton, Leicester, West Ham, Wolves}
- Weakest teams: {Brighton, Burnley, Crystal Palace, Newcastle, Southampton}

For the subsamples of games we only consider games played against teams within the subsets of teams.

6.2 Competition

6.2.1 HA

The results of the significance testing of the HA in the two time periods, in the form of p-values, can be found in Table 1. These are, as mentioned before the outcomes for the paired permutation test and the paired sample t-test. Due to the assumptions we have made, we may see the results in testing the difference of the outcomes between home- and away games, which is essentially what we do, as the results of testing the HA. We expect the assumptions mentioned for the competition setup to hold for these subsamples.

There is a significant HA in all leagues, for goals as well as points in the period where there still was a crowd present. However for the period where the crowd was absent, there is only a significant HA for the goal difference and points won in La Liga. Even for this league the p-value is larger, and therefore shows a weaker indication of a presence of the HA for the games without a crowd. In

Table 1: Testing of the HA

Premier League	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	0.0032	0.458	Goals	0.003	0.264
	Permutation 1	0.004	0.505		0.001	0.266
	Permutation 2	0.003	0.457		0.001	0.245
La Liga	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	<0.001	0.015	Goals	0.003	0.264
	Permutation 1	<0.001	0.016		0.001	0.266
	Permutation 2	<0.001	0.012		0.001	0.245
Bundesliga	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	0.0244	0.296	Goals	0.0135	0.212
	Permutation 1	0.025	0.338		0.008	0.223
	Permutation 2	0.02	0.27		0.007	0.201

Notes: Due to discreteness issues, which results in many duplicate values, we consider two outcomes for the permutation tests. We consider a value in between these two values as the true p-value. In this case, it does not change the significance for any of the tests.

other words, the effect of the HA is seen to be less significant. This is already a big indication that crowd influences the HA.

6.2.2 Crowd-effect

The results of the significance testing of the crowd-effect, which due to our assumptions is the same as testing the difference in HA between the two periods, are shown in Table 2. These are, as mentioned before the outcomes for the paired permutation test and the paired two sample t-test.

For both the Premier League and La Liga the crowd-effect, tested on the basis of the amount of points won, is found to be significant. For both on a 10 % significance level for the permutation test. The t-test is even significant for a 5% significance level for the Premier League. For the La Liga the crowd-effect, tested on the basis of the goal difference in home- and away games is found to be significant. Only for the Bundesliga, the results are not significant, but they do show a much lower HA in the matches without a crowd.

Table 2: Testing the crowd-effect

Premier League	Test	Points	Goals
	t-test	0.050	0.1303393
	Permutation	0.056	0.135
La Liga	Test	Points	Goals
	t-test	0.073	0.064
	Permutation	0.079	0.067
Bundesliga	Test	Points	Goals
	t-test	0.207	0.198
	Permutation	0.221	0.204

Notes: The permutation test now already uses the duplicate values and transforms it into one p-value.

6.2.3 Validity and conclusions

Now we need to know whether these results are valid. The assumptions of the permutation test are not violated. The Shapiro-Wilk test is rejected for every single sample, for goals as well as points. Therefore the assumption of normality, which is needed in the two-sample t-test, does not hold.

The results of the permutation test are valid, whereas the results for the t-test are not valid. The difference between the two tests is not very big, even though we found a difference in significance level for the testing of the crowd-effect in the Premier League. This small difference might be due to the relatively large sample sizes.

On a competition scale there is most definitely a HA present in games played with a crowd. This is mostly due to the crowd-effect, as the testing of the HA is no longer significant in the games played without a crowd for the premier League as well as the Bundesliga, in terms of goals and points. Additionally, the difference between the two periods is deemed significant on a 10% level for both the Premier League and La Liga. These two combined make a very strong case for the presence of a crowd-effect.

6.3 Team-specific

6.3.1 HA

Now we get to the individual approach for all teams. The way that these are tested is explained in the previous sections. All results can be found in the Appendix, but due to the large amount of teams, we did not want to show them here. An overview: In the games with a crowd 7 teams showed to have a significant HA for the amount of points scored. In the games with a crowd 8 teams showed to have a significant HA for the level of goal differences. Not a single team that had a significant HA for the games played with a crowd, had a significant HA for the games played without a crowd.

We are interested in two things, whether there is a global HA present, and which teams cause this HA. Because we have multiple hypotheses we need to adjust the significance level with the corresponding tests. We do this with the Bonferroni-Holm method mentioned in Section 5.2. If one of the hypotheses is significant after this correction, the Bonferroni-Holm method ensures that we can reject the global hypothesis, under a FWER $\leq \alpha$.

For the games played with a crowd, only one team shows a significant difference in goals between home -and away games for α equal to 0.05 for the outcomes of the t-test. This team is Arsenal. This only shows a significant difference for α equal to 0.10. Everton shows a significant difference for α equal to 0.10 for both tests. With the use of goals we see somewhat the same pattern. Now Arsenal shows a significant HA for α equal to 0.05 for both tests, and no other tests are significant for α equal to 0.10

This provides us with strong evidence that some teams have a HA and that there is an overall HA present in the Premier League when there is a crowd. For the games without a crowd, even without the adjustment of significance levels, there is not even one test significant for α equal to 0.05. An indication that the crowd-effect might be the driving factor in the HA in the Premier League.

6.3.2 Crowd-effect

Next, we are interested in the crowd-effect. This showed to be significant for 6 teams in terms of points and for 5 teams in terms of goal difference. Out of the 7 teams that showed a significant

HA in the first period, 5 teams showed a significant crowd-effect measured in points. For the use of goals this was 5 out of 8.

Yet again we are interested in the global crowd-effect and the individual crowd-effect. We make use of the Bonferroni-Holm method again.

After adjusting the significance levels, we were not able to reject any of the hypotheses separately. However, we can also test for at least one null hypothesis to be rejected, without differentiating between the teams. This is done with the use of a Poisson distribution, this approach is mentioned in Section 5.2.

For the outcomes of the t-test, with the use of goals, we can say that at least one of them is significant for α equal to 0.10. This can not be said for the permutation test. With the use of goals we can say that at least one of them is significant for α equal to 0.05 for both the t-test and the permutation test.

6.3.3 Validity and conclusions

The Shapiro-Wilk was not rejected for most teams, but this could also be due to the small sample sizes. For small sample sizes, the power of this tests is very low. However, we do know that the assumption of normality seems to be violated for small sample sizes. Therefore we have checked the distributional graphs of the goals and the points. In this we can see that the points are most definitely not normal distributed and we are not very sure about the distribution of the goals. However, we don't see any indicators that the assumptions of the permutation test do not hold. In the results we also see a clear difference between the two.

We found strong evidence that there is a HA present in the period played where there was a crowd allowed, in contrast to the absence of evidence of a HA in the period where there was no crowd allowed. On top of this we found evidence that at least one of the alternative hypotheses is true for the crowd-effect. This is not very strong evidence but more an indication that there is a crowd-effect present for some teams.

6.4 Subsamples

6.4.1 HA

The outcomes of the testing of the HA within the subsamples are given in Table 3. These are generated in the same way as for the competition setup. Because there is no overlap between the teams in the subsets, these 3 samples can be seen as independent.

Table 3: Testing the HA

Top Teams	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	0.048	0.774	Goals	0.024	0.762
	Permutation 1	0.068	0.848		0.030	0.793
	Permutation 2	0.031	0.696		0.018	0.729
Middle Teams	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	0.024	0.748	Goals	0.014	0.689
	Permutation 1	0.038	0.822		0.020	0.731
	Permutation 2	0.014	0.676		0.011	0.641
Weakest Teams	Test	Crowd	No crowd		Crowd	No crowd
Points	t-test	0.789	0.821	Goals	0.698	0.766
	Permutation 1	0.844	0.881		0.733	0.796
	Permutation 2	0.730	0.760		0.645	0.709

Notes: Due to discreteness issues, which results in many duplicate values, we consider two outcomes for the permutation tests. We consider a value in between these two values as the true p-value.

We see the same pattern as for the competition setup, with a significant HA for games played with a crowd and no real indication of a HA for games without a crowd. This is for goals as well as points. However, this only seems to be the case for the top teams and the middle teams. For the weakest teams, the HA for both games with a crowd and games without a crowd vanishes. Possible explanations for this are the smaller crowd these teams generally have, or a change of playing style when playing home or away, with according differences in the score. This is a good indication that the crowd size might also have an influence on the HA, but this is not discussed into further detail.

6.4.2 Crowd-effect

The outcomes of the testing of the crowd-effect within the subsamples are given in Table 5. These are generated in the same way as for the competition setup.

Table 4: Testing of the crowd-effect

Top Teams	Test	Points	Goals
	t-test	0.050	0.038
	Permutation	0.069	0.043
Middle Teams	Test	Points	Goals
	t-test	0.039	0.040
	Permutation	0.054	0.048
Weakest Teams	Test	Points	Goals
	t-test	0.391	0.388
	Permutation	0.453	0.426

Notes: The permutation now already uses the duplicate values and transforms it into one p-value.

We see similarities between the top- and middle teams sample, which are both significant for α equal to 0.05 for the t-test, with the use of points as well as goals and significant for respectively α equal to 0.10 and 0.05 for the permutation test, with the use of points and goals. The difference between the permutation test and the t-test is again greater for the points won than for the goals scored. This is strong evidence of the presence of a crowd-effect for the "stronger" teams in the Premier League.

6.4.3 Validity and conclusions

The Shapiro-Wilk test showed similar outcomes as with the competition setup. It is rejected almost every time. There is no indication that the assumptions for the permutation test do not hold. This makes it possible for us to say that the permutation test is valid and that the t-test, with according normality assumption, is not safe to use. In two cases it has led to a difference in significance for α equal to 0.05, where the t-test rejected and the permutation test did not.

Furthermore, the HA and crowd-effect seemed to vanish for the weakest teams in the sample. This is an indication that the crowd size might also be of influence on the crowd-effect and therefore

the HA. The other subsamples showed extra evidence of the HA for games played with a crowd, as well as the crowd-effect.

7 Conclusion

In this report we have considered multiple uses of data and according test methods to answer our research questions. With the use of the simulation study we can answer the first sub-question: *Does the permutation test outperform classical tests, which assume normality, that correspond to relevant settings for the HA, in terms of power and accuracy?*

We start with the size of the tests. For smaller sample sizes (20), the size of the t-test is distorted. This is due to the violation of the normality assumption. The size of the permutation test is good for all sample sizes. This makes the permutation test more reliable than the t-test. However, for these small sample sizes, the power of the t-test substantially exceeds the power of the permutation test. As the sample sizes get larger, this difference declines and eventually the power of the permutation test exceeds the power of the t-test. This difference is small, yet significant.

An extra finding in this simulation study, is the increase in power in a level playing field of teams in contrast to a uneven playing field of teams. A level playing field, lead to an increase of power of 0.2. Another improvement which showed was the use of goals instead of points. This led to a doubling of power for the same size of HA.

The use of the permutation test and the t-test for the team-individual and competition wide testing of the HA and crowd-effect allow us to answer the second sub-question: *In what way does the use of the permutation test in significance testing of effect of crowd size on the home advantage qualitatively differ to the results that are found when using methods of existing literature, which use classical testing?*

In our analysis of the assumptions of the permutation test we have seen that there is no real evidence that these do not hold. Therefore, we may say that these results are reliable. There is a difference for the t-test. The normality assumption is violated for almost all samples and makes the test invalid. In multiple cases this even lead to a significant value for α equal to 0.05 or 0.10 for the t-test, and an insignificant value at this level for the permutation test. This was expected from our answers in the simulation study, but nevertheless it does show the same pattern for the empirical research.

As we found the permutation test to be more reliable, we can only conclude things from the outcomes of this test. On a competition scale we can most definitely see a HA present in games

played with a crowd. The crowd-effect seems to be the driving factor as the HA disappears for games played without a crowd. The Premier League and the Bundesliga, show a significant crowd-effect for α equal to 0.10. This is already strong evidence that there is a crowd-effect present in these leagues.

For the team-individual outcomes we saw the same pattern: Strong evidence that there is a HA present in the first period and an absence of evidence of a HA in the second period. On top of this we found evidence that at least one of the alternative hypotheses is true for the crowd-effect. These two combined do show an indication but cannot provide us with strong evidence there is a crowd-effect present for some teams.

With the creation of subsets of teams and the application of the permutation test on these subsets we can answer the third sub-question: *How does the effect of crowd size on home advantage differ for different subsets of games?*

We considered three subsets, the top teams, the middle teams and the weakest teams of the Premier League. The top- and middle teams showed a similar significance level for the HA in both periods and the crowd-effect. The HA in the first period and the crowd-effect were found to be significant for α equal to 0.05, in contrast to the insignificance of the HA in the second period. This is strong evidence that for stronger teams in the Premier League, there is a crowd-effect present and the crowd-effect is the driving factor for the HA. The HA in both periods and the crowd-effect were insignificant for the weakest teams. This might be an indication that crowd size also has an influence on the crowd-effect.

The sub-questions are used to answer the main question of our report: *In what way can the permutation test be an addition to the existing literature on the testing of the effect of crowd size on the HA?*

We have found that it is easier to detect the HA with the use of goals in comparison to the use of points. Additionally, a level playing field of teams provides an increase of power in contrast to an unequal playing field. However, this does lead to smaller samples, as not all teams have the same strength. In this setting, the power of the t-test substantially exceeds the power of the permutation test. However, the size of the t-test can be distorted for these sample sizes, whereas the size of the permutation test is good for all sample sizes. This makes the permutation test a more reliable, yet less powerful, test for the crowd-effect. This would also make it possible to obtain reliable results

in the research of the crowd-effect in specific subsets, such as Champions League games.

8 Discussion

Our paper has some limitations and thus suggestions for further research. This is mostly about the assumptions that have been made and the model that is used.

In our goal model, we expect the goals to be independently drawn from each other. However, this does not always hold. Even though we have accounted for some dependency by including the teams attacking and defensive strengths for both parameters of the goal distributions, Groll et al. (2015) have pointed out that when fitting this kind of model data the estimates of the attack and defense abilities of two opposing teams are negatively correlated. Groll et al. (2018) provides a solution to this; a bivariate Poisson distribution which takes this negative correlation into account. The exact modelling of the games was not the main goal of our research but this might give even more accurate results in the testing of the HA and crowd-effect.

Another thing that could be improved is the use of correction methods for multiple hypotheses testing. The correlation between the individual team hypotheses, could make the Bonfferoni-Holm method conservative. At first, there is not a clear solution to this. However, in the case of the creation of subsamples, it could be possible to make these subsamples independent of each other. This would make it possible to use more powerful correction methods that control for the False Discovery Rate(FDR) instead of the FWER. These tests tend to be less conservative, but do require the independence assumption. This could be used in further research.

Another interesting addition to this paper, is the inclusion of attendance size, attendance density and possible other factors, such as game intensity, for the estimation of the crowd-effect. This could give a better insight where this effect comes from and which parts of the effect have a significant influence. The size of the crowd-effect could be researched as well, instead of just the significance, which is done in this paper. This is not the main use of the permutation test, which focuses on significance but could be an interesting addition.

References

A. Bryson, P. Dolton, J. J. Reade, D. Schreyer, and C. Singleton. Causal effects of an absent crowd on performances and refereeing decisions during covid-19. *Economics Letters*, 198:109664, 2021.

K. S. Courneya and A. V. Carron. The home advantage in sport competitions: a literature review. *Journal of Sport & Exercise Psychology*, 14(1), 1992.

C. Cueva. Animal spirits in the beautiful game. testing social pressure in professional football during the covid-19 lockdown. 2020.

C. Deutscher and D. Winkelmann. Bookmakers' mispricing of the disappeared home advantage in the german bundesliga after the covid-19 break. *arXiv preprint arXiv:2008.05417*, 2020.

A. Dilger and L. Vischer. No home bias in ghost games. 2020.

M. Ferraresi, G. Gucciardi, et al. *Team performance and audience: experimental evidence from the football sector*. Società italiana di economia pubblica, 2020.

K. Fischer and J. Haucap. Does crowd support drive the home advantage in professional soccer? evidence from german ghost games during the covid-19 pandemic. 2020.

M. A. Gómez, R. Pollard, and J.-C. Luis-Pascual. Comparison of the home advantage in nine different professional team sports in spain. *Perceptual and motor skills*, 113(1):150–156, 2011.

A. Groll, G. Schauberger, and G. Tutz. Prediction of major international soccer tournaments based on team-specific regularized poisson regression: An application to the fifa world cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2):97–115, 2015.

A. Groll, T. Kneib, A. Mayr, and G. Schauberger. On the dependency of soccer scores–a sparse bivariate poisson model for the uefa european football championship 2016. *Journal of Quantitative Analysis in Sports*, 14(2):65–79, 2018.

Y. Hill and N. W. Van Yperen. Losing the home field advantage when playing behind closed doors during covid-19: Change or chance? *Frontiers in psychology*, 12, 2021.

W. Hoeffding. The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, pages 169–192, 1952.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

J. P. Jamieson. The home field advantage in athletics: A meta-analysis. *Journal of Applied Social Psychology*, 40(7):1819–1848, 2010.

Á. Jiménez Sánchez and J. M. Lavín. Home advantage in european soccer without crowd. *Soccer & Society*, 22(1-2):152–165, 2021.

R. Johnston. On referee bias, crowd size, and home advantage in the english soccer premiership. *Journal of Sports Sciences*, 26(6):563–568, 2008.

M. B. Jones. The home advantage in individual sports: An augmented review. *Psychology of Sport and Exercise*, 14(3):397–404, 2013.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

D. McCarrick, M. Bilalic, N. Neave, and S. Wolfson. Home advantage during the covid-19 pandemic in european football. 2020.

N. Neave and S. Wolfson. Testosterone, territoriality, and the 'home advantage'. *Physiology & behavior*, 78(2):269–275, 2003.

A. M. Nevill and R. L. Holder. Home advantage in sport. *Sports Medicine*, 28(4):221–236, 1999.

R. Pollard. Worldwide regional variations in home advantage in association football. *Journal of sports sciences*, 24(3):231–240, 2006.

R. Pollard and M. A. Gómez. Comparison of home advantage in men's and women's football leagues in europe. *European journal of sport science*, 14(sup1):S77–S83, 2014.

R. Pollard and G. Pollard. Home advantage in soccer: A review of its existence and causes. 2005.

R. Pollard, J. Prieto, and M.-Á. Gómez. Global differences in home advantage by country, sport and sex. *International Journal of Performance Analysis in Sport*, 17(4):586–599, 2017.

J. J. Reade, D. Schreyer, and C. Singleton. Echoes: what happens when football is played behind closed doors? *Available at SSRN 3630130*, 2020.

V. Scoppa. Social pressure in the stadiums: Do agents change behavior without crowd support? *Journal of economic psychology*, 82:102344, 2021.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

F. Sors, M. Grassi, T. Agostini, and M. Murgia. The sound of silence in association football: Home advantage and referee bias decrease in matches played without spectators. *European journal of sport science*, pages 1–9, 2020.

M. Tilp and S. Thaller. Covid-19 has turned home-advantage into home-disadvantage in the german soccer bundesliga. *Frontiers in sports and active living*, 2:165, 2020.

A. Tversky and D. Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

9 Appendix

9.1 Descriptive Statistics

Table 5: Overview Leagues

	Games Played		Average Goals		Average Points	
League	Before	After	Before	After	Before	After
Premier League	373	233	H: 1.51 A: 1.25	H: 1.41 A: 1.33	H: 1.57 A: 1.20	H: 1.38 A: 1.37
La Liga	409	257	H: 1.48 A: 1.07	H: 1.33 A: 1.11	H: 1.67 A: 1.04	H: 1.52 A: 1.18
Bundesliga	313	214	H: 1.76 A: 1.49	H: 1.59 A: 1.48	H: 1.58 A: 1.19	H: 1.43 A: 1.29

Notes: This table shows the games played, the average goals scored at home(H) and away(A) and average points won at home and away for all 3 leagues. For the average goals scored and the average points won this is thus an average over the entire league.

	Games Played		Goal Difference		Average Points		Percentage of points	
Teams	Crowd	No crowd	Crowd	No crowd	Crowd	No crowd	Crowd	No crowd
Arsenal	H: 26 A: 23	H: 15 A: 17	H: 19 A: -13	H: 0 A: 0	H: 1.88 A: 0.91	H: 1.27 A: 1.35	H: 67.4% A: 32.6%	H: 48.3% A: 51.7%
Brighton	H: 23 A: 25	H: 17 A: 15	H: -4 A: -20	H: -12 A: -2	H: 1.22 A: 0.68	H: 0.82 A: 1.2	H: 64.1% A: 35.9%	H: 40.7% A: 59.3%
Burnley	H: 26 A: 24	H: 14 A: 18	H: -16 A: -25	H: -8 A: -12	H: 1.15 A: 0.67	H: 0.79 A: 1.28	H: 63.4% A: 36.6%	H: 38.1% A: 61.9%
Chelsea	H: 26 A: 25	H: 14 A: 16	H: 19 A: -2	H: 12 A: 5	H: 1.81 A: 1.40	H: 1.86 A: 1.625	H: 56.4% A: 43.6%	H: 53.3% A: 46.7%
Crystal Palace	H: 24 A: 25	H: 17 A: 15	H: -14 A: -4	H: -21 A: -20	H: 0.83 A: 1.40	H: 0.65 A: 0.80	H: 37.3% A: 62.7%	H: 44.7% A: 55.3%
Everton	H: 25 A: 26	H: 16 A: 14	H: 10 A: -12	H: -2 A: -1	H: 1.80 A: 0.85	H: 1.31 A: 1.50	H: 68.0% A: 32.0%	H: 46.7% A: 53.3%
Leicester	H: 25 A: 25	H: 15 A: 15	H: 4 A: 10	H: 5 A: 0	H: 1.40 A: 1.40	H: 1.73 A: 1.40	H: 50.0% A: 50.0%	H: 55.3% A: 44.7%
Liverpool	H: 25 A: 23	H: 15 A: 17	H: 47 A: 25	H: 9 A: 13	H: 2.84 A: 2.35	H: 1.60 A: 1.71	H: 54.7% A: 45.3%	H: 48.4% A: 51.6%
Man United	H: 26 A: 24	H: 14 A: 17	H: 11 A: -8	H: 8 A: 21	H: 1.77 A: 1.21	H: 1.64 A: 2.29	H: 59.4% A: 40.6%	H: 41.7% A: 58.3%
Man City	H: 24 A: 25	H: 16 A: 15	H: 39 A: 24	H: 35 A: 18	H: 2.46 A: 1.92	H: 2.5 A: 2.133	H: 56.1% A: 43.9%	H: 54.0% A: 46.0%
Newcastle	H: 25 A: 26	H: 16 A: 15	H: -9 A: -24	H: -11 A: -12	H: 1.12 A: 0.85	H: 1.00 A: 1.07	H: 57.0% A: 43.0%	H: 48.4% A: 51.6%
Southampton	H: 25 A: 26	H: 16 A: 14	H: -21 A: -22	H: -4 A: -16	H: 0.88 A: 0.96	H: 1.19 A: 0.79	H: 47.7% A: 52.3%	H: 60.2% A: 39.8%
Tottenham	H: 23 A: 26	H: 18 A: 15	H: 13 A: 0	H: 12 A: 4	H: 1.65 A: 1.31	H: 1.83 A: 1.2	H: 55.8% A: 44.2%	H: 60.4% A: 39.6%
West Ham	H: 25 A: 25	H: 15 A: 15	H: -6 A: -22	H: -2 A: -4	H: 1.2 A: 0.84	H: 1.27 A: 1.07	H: 58.8% A: 41.2%	H: 54.3% A: 45.7%
Wolves	H: 25 A: 25	H: 15 A: 15	H: 4 A: -3	H: -2 A: -13	H: 1.48 A: 1.32	H: 1.27 A: 0.87	H: 52.9% A: 47.1%	H: 59.4% A: 40.6%

Arsenal	Test	Before	After	Difference
Outcome	t-test	0.003	0.568	0.080
	Permutation	0.005	0.613	0.093
Goals	t-test	0.002	0.500	0.070
	Permutation	0.003	0.545	0.080

Brighton	Test	Before	After	Difference
Outcome	t-test	0.054	0.815	0.029
	Permutation	0.070	0.855	0.037
Goals	t-test	0.053	0.839	0.023
	Permutation	0.066	0.864	0.029

Burnley	Test	Before	After	Difference
Outcome	t-test	0.084	0.853	0.041
	Permutation	0.102	0.885	0.050
Goals	t-test	0.216	0.448	0.400
	Permutation	0.237	0.484	0.433

Chelsea	Test	Before	After	Difference
Outcome	t-test	0.139	0.322	0.548
	Permutation	0.161	0.369	0.590
Goals	t-test	0.073	0.198	0.493
	Permutation	0.084	0.227	0.524

C. Palace	Test	Before	After	Difference
Outcome	t-test	0.953	0.651	0.831
	Permutation	0.940	0.708	0.855
Goals	t-test	0.841	0.479	0.855
	Permutation	0.865	0.444	0.870

Everton	Test	Before	After	Difference
Outcome	t-test	0.004	0.649	0.037
	Permutation	0.006	0.700	0.046
Goals	t-test	0.037	0.542	0.100
	Permutation	0.043	0.595	0.115

Leicester	Test	Before	After	Difference
Outcome	t-test	0.5	0.251	0.700
	Permutation	0.541	0.290	0.729
Goals	t-test	0.655	0.312	0.752
	Permutation	0.674	0.347	0.769

Liverpool	Test	Before	After	Difference
Outcome	t-test	0.021	0.587	0.184
	Permutation	0.042	0.640	0.206
Goals	t-test	0.011	0.581	0.197
	Permutation	0.016	0.608	0.230

Man. U	Test	Before	After	Difference
Outcome	t-test	0.064	0.940	0.022
	Permutation	0.078	0.956	0.029
Goals	t-test	0.050	0.798	0.049
	Permutation	0.060	0.807	0.057

Man. City	Test	Before	After	Difference
Outcome	t-test	0.069	0.201	0.337
	Permutation	0.088	0.265	0.371
Goals	t-test	0.121	0.103	0.603
	Permutation	0.135	0.118	0.626

Newcastle	Test	Before	After	Difference
Outcome	t-test	0.221	0.557	0.352
	Permutation	0.254	0.617	0.383
Goals	t-test	0.096	0.433	0.316
	Permutation	0.113	0.471	0.342

Southampton	Test	Before	After	Differer
Outcome	t-test	0.592	0.185	0.762
	Permutation	0.636	0.227	0.793
Goals	t-test	0.496	0.117	0.790
	Permutation	0.523	0.154	0.802

Tottenham	Test	Before	After	Difference
Outcome	t-test	0.203	0.088	0.522
	Permutation	0.232	0.108	0.558
Goals	t-test	0.125	0.271	0.322
	Permutation	0.143	0.303	0.345

West Ham	Test	Before	After	Differenc
Outcome	t-test	0.167	0.339	0.379
	Permutation	0.193	0.392	0.418
Goals	t-test	0.110	0.407	0.249
	Permutation	0.125	0.454	0.271

Wolves	Test	Before	After	Difference
Outcome	t-test	0.325	0.197	0.651
	Permutation	0.366	0.238	0.681
Goals	t-test	0.237	0.135	0.741
	Permutation	0.268	0.158	0.765