Statistics Coursework

AIM:

After careful consideration over the possible investigations that can be carried out with the data at hand, I have made a decision on the investigation that I am going to carry out.

I want to investigate the times in ten minute periods at which goals are scored; this will be done for all four of the league results that I have. I want to do this investigation so that I can find out in what ten minutes of the game, the most goals are scored.

METHOD:

The way in which I can obtain the data that I want to analyse, has to be unbiased. The reason is so that the investigation becomes fair and as correct as it can possibly be. I already have me resource for data, this resource contains the football results from four of England's largest football leagues and contains the results of them for six consecutive weeks. I am going to use this resource to carry out my investigation; these results are totally factual and have no biased approach to my investigation whatsoever.

From observation of the resource that I have, I see that the Premiership league has less games played on a weekly basis than the other leagues do, this is due to less teams being in the Premiership than in the other leagues and divisions. Due to this, I will have to pick as many teams as possible per week from the Premiership for analysis, and then use that same number to pick from the other leagues. For example, if the Premiership only has nine results and the other leagues have eleven, then to make the investigation fair, I will pick the maximum number of results from the Premiership, that would be nine, and then I would pick nine from the other leagues and divisions. The main thing to do at this stage is to choose the nine results from the other leagues totally at random so that the selection is not biased in any way. To make things as simple as possible, I will use a method of sampling called clustering. I will pick a random starting point on the other leagues and then pick the nine results that come after that starting point. Of course this is just an example and therefore with the real investigation, the number may be different, so until then this is the method that I will use for the investigation.

I will carry out this method for all of the leagues and divisions then take the goals that were scored and at the times that they were scored, then put them in a table which will be split up into ten minute periods. I will then analyse the results that the table shows and hopefully draw up a conclusion on the time at which the most goals are scored.

SUB INVESTIGATIONS

During the investigation I will probably make new discoveries and findings. After this occasional new finding, I may add a not describing how I am going to add a new sub investigation to the main project. The sub investigation(s) will be added so that the answer to it or the answers to them can be used to help me find the answer to the main investigation; it is a support method to help with the investigation.

There are many possible sub investigations that can be pointed out already at this stage. The reason why I have not stated them in simple. The reason is because I do not know whether the sub investigation would be relevant in aiding the correct conclusion to my main investigation. Due to this, I cannot state the sub investigation now, however, later on during the investigation; I will be able to get a clearer image of the sub investigations in mind and whether or not they will actually be relevant for answering the main investigation clearly.

PILOT STUDY

The Pilot Study is just like a preliminary experiment, this will be in aid for the collection of data that will take place during the main investigation. What I mean by this is, the pilot study will help me to understand how to efficiently collect the correct data and how to do it. The study will also include analysis of the data which helps me to prepare for the real data collection and analysis.

Through the pilot study, I can get a good insight on the actual investigation. The results from the pilot study, after careful analysis, will be able to give me a basis for my prediction on the final conclusion to the main investigation.

The pilot study that I am going to carry out will be of the same sort that has been planned for the main investigation. I will take results from the Premiership League and the Nationwide Division One from the fifth week of fixtures and results. I will take four completely random results from each of the selected league tables and then analyse them on the same basis/question as the main investigation.

I picked the two leagues in a sort of clustered sample, however you could just say I picked them for quick data collection rather than having to go from one league to another, these leagues are next to each other in the resource book making the data collection that bit easier for me. Since this is just a pilot study, I only wish to get the gist of the main investigation, and that is why I have not spent so long on the selection. Otherwise, I would have probably taken the data at hand into careful consideration and then picked the right data for the job. I would have also picked more data, as I can predict that four from each of the two leagues, would not have given me any sufficient foundation to have based a conclusion on, and therefore this pilot study will only help to make a **prediction** for the main investigation. When the time comes for me to carry out the real investigation, then I will surely use a wider variety of results, and would include a larger number of results in my investigation. This will be an advantage since I will be able to draw up a correct conclusion.

PILOT STUDY RESULTS

The previous table and its results show the goals scored at the ten minute periods within the ninety minutes of the football match. Each period of ten minutes has a number of goals that were scored in that period, the number of goals is shown for the Premiership, and the Nationwide Division One, and it also shows the overall number of goals that were scored in the match at the certain periods of time. The overall number comes from the combination of the Premiership goals and the Nationwide Division One goals put together to give an overall total for the number of goals scored in that period of time.

The reason for creating the table in the manner that I did was to see if there is any difference between the two leagues and the times that the goals were scored. The times at which the goals were scored could show a difference or a relationship, and this is what I want to find out to help me with the main investigation. I would also want to check how many goals overall were scored and how many goals were the difference at each ten minute period. For example, I would want to see if at any one ten minute period, the number of goals scored in each league is different. This would show me what league provides more goals in its games. I will be looking for any relationships between the two leagues and for any major differences between the two leagues and then make any sub investigations for them.

However, looking back at the table of results, I can already see that there are not enough results to see a certain conclusion which could come from the analysis of the results. Due to this reason, I will be thinking of creating another table exactly the same as this one, however, I will take as many results as I can from the two leagues, but most probably from two weeks rather than one. I will select two weeks at random, and the reason for having two weeks to take results from would be simply because I know that more results are needed for me to draw up a satisfactory conclusion on in my pilot study. From this conclusion, I will be able to make a solid prediction on the main investigation.

I think that this new table of results and new number of results will help me to see any relationships of trends etc. This will help me draw up a suitable prediction for the main investigation.

PILOT STUDY ANALYSIS

From the table of results from my pilot study, I have stated previously, I can tell that there is not a sufficient number of results to decipher any solid trends or relationships. However, from the table I can see one ten minute period that brings the most number of goals in the match.

Within the last ten minute period of the game, I can see from the overall results that the most number of goals are scored then. This I have taken from the table of results that I drew up for my pilot study. Since this is the overall number of goals which are scored, it means that the goals come from both of the leagues that I have taken results from, which are of course the Premiership and Division One.

The goal results put together for week Five out of the eight in total fixture results that I have analysed, together combined, I have found that the number of goals that come in the last ten minute period overall is the most in whole pilot analysis. Since the most number of goals are scored in the last ten minutes, I would now have to analyse what league brought the most goals out of the two leagues in the study and whether the rule of goals coming the most in the last ten minutes applies for the Premiership only or Division One, or even for both.

On analysis via the table of results that I drawn up, the results show that the Nationwide Division One only follows the rule that was set which made out that the most number of goals coming in a match would happen in the last ten minutes of the game. This can be seen because three goals are scored in the last ten minute period overall out of the matches analysed in Division One. From this, I can see that the largest amount is scored then, the reason being that the largest number of goals scored on the tally chart shows that three goals is the largest number whereas the other goals scored number is either no goals scored, one goal scored, or two goals scored, meaning that the three goals scored is the highest number of goals scored overall, this shows that the rule applies for this league. This shows clearly that since there is only one ten minute period which shows three goals being scored overall, that that particular ten minute period was the period that has the most goals scored. This rule that has been set, only accounts for the Division One. I have seen this from analysing the results which show that three goals come altogether in that last ten minute period.

However, in the Premiership, the most number of goals scored at any one ten minute is not within the last ten minute period like Division One or the set rule that I have found. The Premiership has the period within 11-20 minutes at which the most goals are scored. Due to this I can see that the Premiership brings its largest number of goals in the 11-20 minute period rather than the last ten minute period as in Division One. As I stated before, this pilot study is not sufficient enough for me to base anything on, and so I will just keep its analysis and results in mind, but I will not draw any firm prediction from this. I will hopefully find what I am looking for in my second pilot study.

SECOND PILOT STUDY

After my first Pilot Study I can see that there is not really a solid enough foundation for me to base any firm prediction on. Due to this I will need to repeat my pilot study; however I am going to change the length of the study and by this I mean that more results are going to be used and probably over two weeks worth of results rather than just one. The reason for this is to get more results for me to analyse and then hopefully see any trends and relationships; I can also then gather a sound prediction for the main investigation from more results because I can hopefully see at which time more goals are scored from this new data collection method. Basically, with this larger collection of data collection I will be able to find out what I am looking for, this would be by being able to see more goals and the times that they are scored, then I can hopefully also see at what times the goals are most likely to be scored, which is my main investigation, and so I will be able to draw up a firm prediction from this.

To get a larger number of results I have taken results from two weeks, these weeks have been picked at random and so week three and week five have been selected. This is a pilot study and the analysis is what really matters at this stage, and that is why I am not taking much consideration over the selection of data and the way that I am going to collect the data or choose it. On looking at the football resource I have, I can see that the Premiership fixtures are less than the Division One fixtures. Due to this I will have to select the largest number of Premiership results available and then that number will be the same number of results that will be taken from Division One. The results taken from Division One will be clustered and will have a totally random starting point from which the clustering will begin. Coming back to the selection process, I basically mean, that the number of Division One results taken will be the same as the maximum number of Premiership results that can be taken. The reason is of course because there are less Premiership results than the Division One results, so the same numbers are needed for a fair table of results.

This method of data collection will mean a totally random starting point for the Division One collection of data. I will be collecting the data in the way of clustered sampling since I want results as fast as I can so that I can obtain my prediction quickly. Hopefully these results will broaden the view that I will receive from the results via the same analysis and method. The results will show me the right sort of results from a correct analysis which I will be able to create a suitable prediction from for my main investigation.

The results will be taken from week 3 which is DECEMBER 21/22 2002 and from week 5 which is DECEMBER 28/29 2002.

SECOND PILOT STUDY ANALYSIS

On analysing the new pilot study, I can see much more clear trends and I can already say that after a thorough analysis on my new pilot study, I will be able to make a solid prediction for my main investigation.

From my previous pilot study, I found out that the most number of goals were scored in the last ten minutes of the games. This rule applied for Division One only, which I also found out from my previous pilot study. The rule did not apply for the Premiership since in between 11-20 minutes the most number of goals were scored. Since I did not get a real conclusion from my previous pilot study, I have broadened the pilot study and therefore obtained new results.

My new results have shown me results similar to the first set of results. The overall number f goals scored were greatest in the last ten minutes of the game, but once again, the Division One goals accounted for the majority of the goals this making the rule only apply for Division One once again and not for the Premiership as well.

The Premiership shows that the most number of its goals were scored in the 71-80 minute period of the game. This is different to the previous set of results which shows that the majority of goals come from the 11-20 minute period of the game. The rule of most Division One goals coming in the last ten minutes however has stayed the same between both of the pilot studies that I carried out. From my pilot study I can say that the Division One goals will probably come in the most abundance within the last ten minutes of the game. From the table of results I can see one more trend. Overall, more goals are scored after half time than there are goals scored before the break. This means that the defense are either more vulnerable after the break or that the attack becomes more efficient after half time.

This therefore brings me to my first sub-investigation. I want to see if goals are scored more in the first half or the second half. This investigation will be very simple as it is just a matter of looking at the table of results and then drawing up a conclusion from the time at which the goals were scored.

Going back to the pilot study, I have found that the goals are scored more readily in the last ten minutes of the game in Division One. However, I have yet to find a real trend in the Premiership results. However since I need to find this out for the main investigation, I will leave it till then. Right now, my prediction is that the most number of goals scored **overall** are scored between 81-90 minutes of the game. I predict that the Premiership will have most of its goals scored between the 71-80 minute period of the game. For Division One, I believe it will stick to the rule applied previously of having the most goals scored within the last ten minutes of the game.

My second sub investigation will be to find out what league provides the most goals. At the moment I can see that the Premiership doesn't have many goals scored overall compared to Division 1 and so I rule the Premiership out of being the one that provides the most goals. Looking at the table of results, I can see that the lower league provides more goals, so I predict that the most goals will come from the lowest league, I will find out after I have completely my main method of data collection and my data analysis.

PREDICTION

From my pilot studies, I can now base a firm prediction for Division One. I predict that the most number of goals scored will be within the last ten minutes of the games. I cannot base any prediction for the Premiership on such a solid foundation as I have for Division One due to insufficient consistency in my pilot studies. However, from what I can see, I would say that the most number of goals will come within the 71-80 minute period. The other leagues have not bee analysed and therefore I cannot make any predictions for the.

However, as I can see from my pilot study, more goals are scored overall in Division One than in the Premiership. This is helpful in my sub investigation prediction. I predict that the lower you go in football standard, the more goals will be scored. This means that I predict that more goals will be scored overall in the Conference league than all of the other leagues. This is my prediction for one of my sub investigations. I predict this because of the table of results that I have drawn up for my pilot study. From personal football knowledge, I would say that the lower leagues have weaker defenses and so the natural strikers find scoring much easier. Due to this, I would say that the lower in standards you go in the football leagues, the more goals will be scored. This is exactly why I believe that the number of goals scored will be greater in the lower leagues than it would be in the higher classed leagues.

I can also make predictions for my other sub-investigation. I predict that more goals will be scored in the second half than the first half. This is due to the results of the pilot study. It shows in the overall number of goals scored, that the Premiership results and Division One results put together have more goals scored overall in the second half of their matches. This can be easily seen in the table of results. Therefore I predict that more goals will be scored in the second half than the first.

NB:

[illegible]

DATA COLLECTION METHOD (MAIN INVESTIGATION)

Since I want to obtain a larger amount of results for this main investigation, I will be using the results of all of the six weeks. However, I will have to vary the amount of results taken each week according to the number of the Premiership results that will be taken. I will be taking results from the Premiership, Division One, Division Two, Division Three, and the Nationwide Conference. The results will hopefully show me a trend which I will be able to point out in my conclusion for this investigation. I have already stated my prediction previously and I will now see if they are true or not.

The first week, DECEMBER 2nd 2002, shows that there were nine fixtures in the Premiership. Due to there being nine premiership fixtures that week, I will have to take nine from each other league to analyse and put down into my table of results. The second week, DECEMBER 16th 2002, shows that there were also nine fixtures; this means that I will take nine from the other leagues as well to analyse. Week 3, DECEMBER 21/22 2002, is the only exception, this week has got the least amount of fixtures in the Conference league, this means that the same amount of the Conference league fixtures have to be taken from the other leagues as well. Therefore, since the Conference ha a total of eight fixtures that week, I will take eight results from the other leagues as well. Week 4, DECEMBER 26th 2002, has ten results meaning ten will also be taken from the other leagues each. Week 5, DECEMBER 28/29 2002, also has ten results which means that ten results will have to be taken from each league. The sixth and final week, JANUARY 18/19 2003, has ten results in all and so ten results from each league will also be taken for analysis.

Each league will have a totally random starting point, which I will chose, and from there, the relevant number of results will be taken down for analysis. I will list the weeks in order and the leagues in order, then I will show the random starting point and from where to where the results were taken. For example, in the Nationwide Division One, I would pick a random starting point, e.g. the starting point is Brighton vs. Burnley, and then the end point would be Wolves vs. Bradford. The results will be taken consecutively after the start point in a cluster. I will be going down the resource book; this will mean that the results will be taken in order for consecutive results.

I will take the results and then use the times that are given of the goals scored and then add them to my table of results. The table of results will be the same as they were when I had carried out my pilot study. From this I mean that the table used in my pilot study will be the same as the one that I am going to use for the main investigation. The only difference will be that I will add more rows for the addition of the new leagues that will be added into this part of the investigation. I will also make one slight adjustment, although I will keep the 41-50 minute period of time, I will make two extra columns which ultimately split that ten minute period into two five minute periods of the 41-45 minutes of the game and the 46-50 minutes of the game. This is so that I can see clearly overall in what half most of the goals on average are scored within.

NB: POSTONED MATCHES WILL NOT BE ACCOUNTED FOR

DATA PRESENTATION

I have carefully checked my results and have now presented them in different ways. I have used pie charts drawn by hand and made on the computer, and I have used bar charts, line graphs and tables to present my data, these have been done by using the computer. I have used different methods to show the same data because I want to get different views of my data so that I can see every aspect of the data and not miss out any points in the data.

The pie charts were used to show a clear difference in the data, however, the only disadvantage would be close data. If two or more sets of data are closely matched, then it is hard to tell which is greater. However, the pie charts do clearly show the data and show clear differences when there is a supposed to be clear differences in the data. A disadvantage that comes when hand drawing the pie charts would be the amount of time that has to be taken to calculate the angles of each data and then drawing and colouring in the charts. Otherwise, pie charts are good for presenting data. One last advantage for the pie chart is that if there is only two sets of data, then a clear difference can be seen much more easily between the two values, whereas with more sets, it becomes harder to see which set of data is the greatest and so forth.

Bar charts and line graphs are more or less the same type of things; I prefer the line graphs over the bar charts. They both clearly show the differences in data and one can easily see what data is greater, the numbers on the axis help as well to determine the data value. The difference is that the line graphs are make it easier to determine the greatness of each set of data. For example, on a bar chart if there are two sets of data next to each other on the chart, and if they were both more or less evenly matched, then it would be very hard to tell the difference in value between the two. The line graphs have lines to show even the slightest changes in value. This means that the problem is eliminated, the line graphs are very good ways of presenting the data.

I have used the chosen methods to present my data solely because they are clear and easy to understand. This is why I have used the certain methods to present the data that I have. From these presentation methods, I will be able to easily analyse the results and then hopefully draw up a conclusion.

DATA ANALYSIS

After looking at my different presentations of the data that I have created, I can see that there are clear answers to my investigations. The data is shown in very clear ways, I can easily interpret what each graph, chart or table shows me and this is what is going to help me to draw up a conclusion.

The sub investigations were to find out in which half the most amount of goals were scored, whether they were scored mostly in the first half or mostly in the second half. The second sub investigation was to find out which league brings the most goals. By this I mean, I want to find out the league in which the most goals are scored, this will be seen through the presented data in all its forms. Hopefully by analysing the data correctly I will be able to figure out the answers to my investigations, including of course, my main investigation which is of course to find out in which ten minute period in all games, the most number of goals are scored (in a football match based on the data).

After looking at my data and carefully analysing it, I can see the answers to my investigations. For my sub investigation which asks to find out which half the most amount of goals were scored in, I can clearly see that there is a trend which shows in each league, as well as overall, that the most amount of goals are scored in the second half rather than the first. This is easily seen through the pie charts that have been made on computer.

The second sub investigation asks what league brings the most goals. By looking at the graphs, I can see that the league that brings the most amount of goals was the Conference League. This can be seen very clearly from my line graphs and so the sub investigation has an answer to it.

The main investigation also gets its answer from the charts that I have made. The investigation asked in which ten minute period the most number of goals were scored in. I have found due to my graphs that the most number of goals are scored within the last ten minutes of the game. This is shown, like I have already stated, in my charts that I have created for the data and it is clearly shown within them.

CONCLUSION

After analysing my data, I have finally come to the conclusions to my investigations. For the first sub investigation, which was to find out in which half of the game the most goals were scored, my analysis shows that the most number of goals are scored in the second half of the game. This follows my prediction, thanks to my pilot study; I had made the correct prediction for the investigation. The prediction was correct and this is shows from the conclusion which shows that the majority of goals were scored in the second half rather than the first half. Fortunately, all of the leagues show that most of their goals were scored in the second half, this means that there are no anomalies and all the leagues follow the same rule.

The second sub investigation was to find out which league brought the most goals overall. I had predicted that the further down in footballing standards one went, the more goals there were likely to be. Once again, this prediction had come from the pilot studies that I had carried out and once again the prediction was correct. From analysing my presented data, I can see that the most amounts of goals that were scored came from the Conference League, which is the lowest in football standards out of the leagues chosen. This goes to show that my prediction was correct. From the graphs and the table, we can see that the number of goals increase as the standards drop, however, at Division Three, there is a slight anomaly. The number of goals that were scored in Division Three, were fewer than Division Two, this means that the rule of more goals being scored as standards dropped, was not being followed by this league. Fortunately, this was just a lone anomalous point, and because all of the other leagues followed the rule, it showed that the rule was in place and that the conclusion had been found. Therefore, after analysing the anomaly to find that there was no problem to the rest of the data, I can say that my prediction was correct that the greatest number of goals would be scored in the Conference League due to the rule.

The main investigation was to find out which ten minute period of a football match the most number of goals were to be scored. After setting my prediction solely on the outcome of my pilot studies, I had said that the most number of goals would come in the last ten minutes of the game. From looking at the table and the graphs and charts, I can immediately see that the most number of goals came in the last ten minutes of the game. This is supported very well because each league follows this rule, so there are no anomalous points and the conclusion is easily discovered.

These are my conclusions to my investigations.

EVALUATION

My investigation was very successful and I have taken the right approach towards it, the statistical approach. I have seen my faults in certain areas and then I went on to correct them for the main investigation, these mistakes can be seen in my appendix. My draft investigation was based on the same questions; however, I had planned to carry it out in a different manner.

The investigation itself was a good one. It really did test my statistical knowledge and made me use that knowledge to get answers for the investigations. I had to use tables and then draw up graphs and charts from the data that was within them. I also had to use a no biased way to gather my data (data sampling). I had to think what type of graphs and charts were going to be used in accordance to the data at hand, and I think that overall I have used my statistical knowledge well to plan the investigation and carry it out in the right manner.

I do think that a few extra graphs and charts could be added for each set of data; however, I have only done this for the main investigation not the sub investigations. The reason for this is simple; I want to spend more time on the main investigation than I do on the others. Sometimes, adding too much to the sub investigations makes it harder to analyse the data and it is a very long process to create and analyse them.

I have used the certain graphs and charts that I have due to their relevance with the data. I thought that the graphs and charts would show the relevant data clearly and in the most efficient way, I had many other types of presentation methods which I could use, however I only used the ones that I had because they are more efficient in data presentation. I have hand drawn a set of pie charts so that I can test my own understanding of how to draw up charts with set data, this has proved successful as I see that I have managed to produce sufficient pie charts to go with the data.

I have come to my conclusions for each of the investigations due to careful planning of my pilot studies and then careful analysis of them. I have taken relevant predictions from the results of the pilot studies and I have also taken the same layout of the pilot studies for my main investigation. These have worked well to show me to my conclusions. Due to my pilot studies and planning I have seen that the whole investigation became very easy. This is how I came to my conclusions so easily.

Overall, I think that the whole investigation went very well and that I have come to the correct conclusions, therefore I would put this whole investigation down as a success. I thoroughly enjoyed this investigation and have increased my knowledge of statistics from it.