Maths Statistics Investigation

Maths Statistics Investigation

To Find the Quality of Players Playing Rugby for Malvern College 1 st XV

Aim

I will attempt during the course of this investigation to rate all of the players who played for Malvern 1st XV in the first 6 matches of the 2000 season. In order to do this I must collect the data, analyse each piece of data collected, and then devise some sort of system using a formula that will rate an individual player over the course of the season. This formula must not only calculate the player's point score for each match according to a points system which I shall devise, but also average the points score over the season taking into account time missed through injury. I will also calculate the average standard of player using the mean, and also find the standard deviation to show the average deviation from the mean.

Plan

In order to devise my scoring system I must first decide on the categories that I shall use for my scoring system. I have decided that there will be two criteria for assessment of a player. These criteria will be:

*Time spent on the pitch + time out due to injury, divided by number of matches played to find the average time spent on the pitch.

*Coach's assessment percentage score for each match, taking into account position and what is required. I will add up the score for each match and divide by the number of matches to find the mean coaches score.

These criteria should be adequate to assess a player's skill in a set position. I will then use the following formula:

Mean number of minutes on the pitch per match* Mean coaches score

Using this formula, I hope to be able to find two things, firstly, by finding the number of 1st team players in each house and grouping their points scores and then adding up the points scores of each house, I should be able to predict a winner of the house competition. Secondly, I should be able to find the mean and standard deviation of each house, thereby showing the average ability and identifying the range of abilities within each house. Finally, by plotting a graph of Coach's percentage score (y axis) against Total time spent on pitch (x axis), I will be able to draw a scatter graph, and be able to rate any player in a similar way to that in which the Coach rates any individual by using the formula y = mx +c.

Aim 1

After adding up the points scores of each of the houses, I have found that the house most likely to win the house competition is number one. However, the reason for the high total is the number of good players in the house. There are 8 players who have all played first team rugby during the first 6 games of the 2000 season compared to the next largest grouping of players in one house, number two, which only has 5. For this reason I predict that number one will win the house competition, not only because of the quality of the players in the house, but also because of the number of players capable of playing first team rugby. However, my prediction can't be considered completely accurate as in order to make an accurate prediction, I would have to look closely at the 2nd and 3rd XV as well as simply the 1st XV. This would involve a considerable extension of my investigation.

Aim 2

Although I have predicted that number one will win the house competition, I would place number two in close contention. Although number one have the greatest number of first team players and the highest points score, number two, although fielding 3 fewer first team players, has a higher mean points score and a lower standard deviation. This shows me that number two have a fewer number of quality players, but the players they have got are of a similar high standard. I can therefore assume that number two do not have the same strength in depth that number one have, but the first team players available are of a generally higher standard. Number one, in fact, has a lower mean average than SH, number 5 and number 7 as well as number two and a relatively high standard deviation. This shows that although number one has a large number of first team players, they have some very good players, towards the top end of the scale as well as some poorer players who are at the lower end of the scale. This variety of abilities shows me that although number one are likely to win the competition, this is by no means assured as there are comparative weaknesses in some areas. Number 7 could also mount a fairly serious challenge as despite the range of abilities shown by the high standard deviation, they have the highest mean average showing the quality of some of the players available to them.

Aim 3

The first two aims have been relatively straightforward to accomplish, the next objective should be slightly more complex. I have plotted a scatter graph using the data and plotting Total time on the field against the Coaches percentage score. I have then taken the average scores by drawing a straight line through the points attempting to keep a similar number of points on either side of the line whilst maintaining the general trend of the results and allowing for anomalous results. The result is that I now have a straight-line graph which can be analysed using the equation y=mx +c, where m = the gradient and c = the y intercept, where the line crosses the y-axis. In order to calculate the gradient, I must divide the change in y by the change in x. The change in y is 18 and the change in x is 450. Therefore by dividing 18 by 450, I will obtain a value for m:

m = 18/450

m = 0.04 and c = 55

Therefore, $y = 0.04x + 55$

I can now use this formula to calculate the values of y given the values of x. For example if I take a value of x such as 320, I should be able to calculate the values for y in the following manner:

$y = 0.04*320 + 55$

$y = 67.8$

When I look at my graph, I find that this value is correct. In order to ensure that my formula is correct I will try it with one more example of a value taken from the graph so that I can test by reading off from the graph whether or not it is correct. I will take a value of x of 450 this time:

$y = 0.04*450 + 55$

$y = 73$

Once more this value is correct according to my graph. This means that using the formula, $y = 0.04x + 55$, I can calculate any player's percentage score according to the Coach given the number of minutes he has spent on the pitch. For example, given that a player has spent 550 minutes on the pitch, what is the Coach's percentage score?

$y = 0.04*550 + 55$

$y = 77$

I can therefore state that had a player spent 550 minutes on the pitch, his Coach's percentage score would be 77%. Similarly given that a player's Coach's percentage score is 87% I can calculate how long he would have spent on the pitch:

$87 = 0.04*y + 55$

$87 - 55 = 0.04*y$

$(87 - 55) / 0.04 = y$

$800 = y$

I can therefore state that he would have spent 800 minutes on the pitch if he had had a Coach's percentage score of 87%. I can therefore say that I can calculate with accuracy the Coach's percentage score or the total time spent on the pitch by any player. However, this system clearly has its drawbacks, as it would be impossible for a player to have spent longer than 450 minutes on the pitch. However, I could apply this scoring system to any 6 matches throughout the course of the season, and provided players play at a constant level, I could calculate the Coach's percentage score given the time spent on the field. However, it is also clear that as the player's standard of play varies depending on the match in question, and sometimes players have reasons for missing matches and therefore the time on the field, although I have taken injury into account, may not be entirely 100% reliable. Therefore, although this system provides a method of gauging the standard of individual players in the Malvern College 1st

XV, it is by no means an accurate measurement of the quality of players due to other factors. However, it is an accurate guide that could be followed by a new coach to pick the best XV players in the school.

Extension of My Investigation

I decided that to extend my investigation further, I would take one of the categories investigated, namely the time spent of the field and using the data that I had already obtained, take a different line on the analysis of the data. I therefore took the results for the Total time spent on the field taking account of injuries and made a graph of the data. I took the time on the field at intervals of fifty minutes, and counted the number of players in each 50 minute block, e.g. 0-49, 50-99 etc. I found that the graph that resulted showed a bi-modal distribution, showing that the majority of the players that had played had either spent a very short period of time on the field (between 0 and 99 minutes) or else they had played almost every game (between 300 and 349 minutes.) This shows that the coach used a small nucleus of players who were obviously his first choice squad, only bringing in replacements when injuries occurred to his first choice team. This nucleus appears not to change during the early part of the season. This assumption was backed up when I made a cumulative frequency diagram and calculated the lower and upper quartiles and the interquartile range. The lower quartile was very low, only 65 showing that many players spent a relatively short time on the field. The upper quartile, at 340, is approximately where it would be expected to be as the concentration of players who have spent very little time on the field is balanced out by the concentration of players that have spent large amounts of time on the field, again emphasising the bi-modal distribution of the graph. The interquartile range (the difference between the upper and lower quartiles) is again fairly large due to the extremely large concentration of players that have spent a very small time on the field. The median value is low for the same reason (200 minutes.)

I then decided to look at the mean and standard deviation of the data. The mean is higher than the median as it is not so affected by the early concentration of values and it gives a better reflection of the total time spent on the field. It is still affected by the weighting of numbers at both ends of the spectrum, and in this sense it does give a fair reflection of the average time spent on the field. The standard deviation in this case shows me the large deviation in the values for time spent on the field, and gives the impression of a bi-modal distribution.