Gerald's Column
by Gerald Fitton

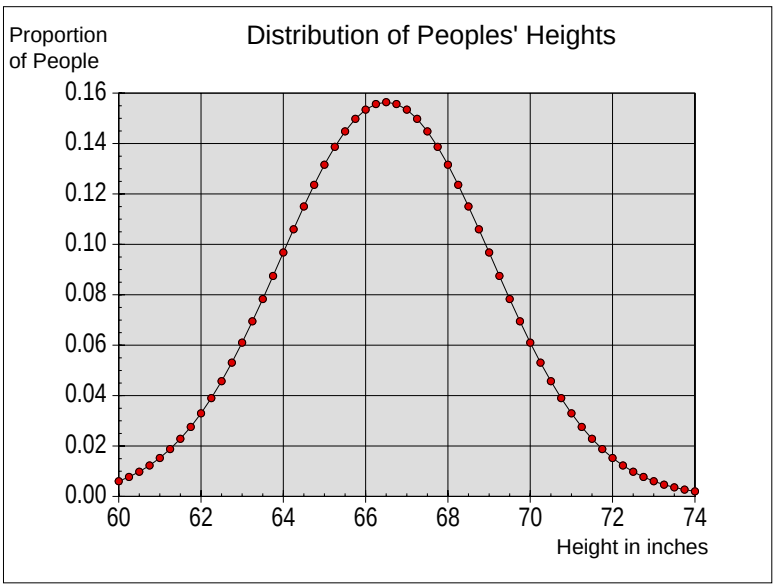
Let's go back to my column in the April 2002 edition of Archive. I wrote, "The different ways of 'building up' a PDF for a real or hypothetical Statistic can be divided into two classes. One class of methods is Measurement and the other class is Theoretical." I went on to write, "what I have called the 'Theoretical' method of creating a Probability Distribution Function (PDF) relies on symmetry". We looked at a flippin' coin, the symmetry of which we misjudged; we looked at the random arrival time of buses that never appeared. In both of these cases we guessed at a PDF, used it to do what we thought were useful calculations—and then we discovered that the sums were totally irrelevant.

This month we shall forget about building up PDFs theoretically and consider an alternative method, measurement.

Descriptive Statistics

The statistic I have chosen for my first example of a measured distribution is the height of what I shall describe very loosely as Western Adults. I do know there is a much more detailed definition of a 'Western Adult' but I don't think understanding that definition will help you to understand what my subject, 'Descriptive Statistics', is all about.

Have a look at the graph below. I have created it using PipeDream. Descriptive Statistics is about trying to describe a distribution such as this one.



Lots and lots of people had their height measured (in good old fashioned inches) about 100 years or so ago. Some were short and some were tall. Some people were shorter than 60 inches and some taller than 74 inches; I haven't included them.

The interpretation of the graph is this. The horizontal (x) axis is the height of people. If we divide all the people into small groups according to their height then the y value (plotted along the vertical axis) is proportional to the number of people in each of group.

The Middle

There is no doubt that one of the most important questions to answer about this distribution of heights is “Where is the middle?” With the distribution shown in the screenshot there is no doubt where the middle is. It is at 66.5 inches. With a fair degree of confidence we can answer: “The average height of Western Adults (100 years ago) was 66.5 inches”.

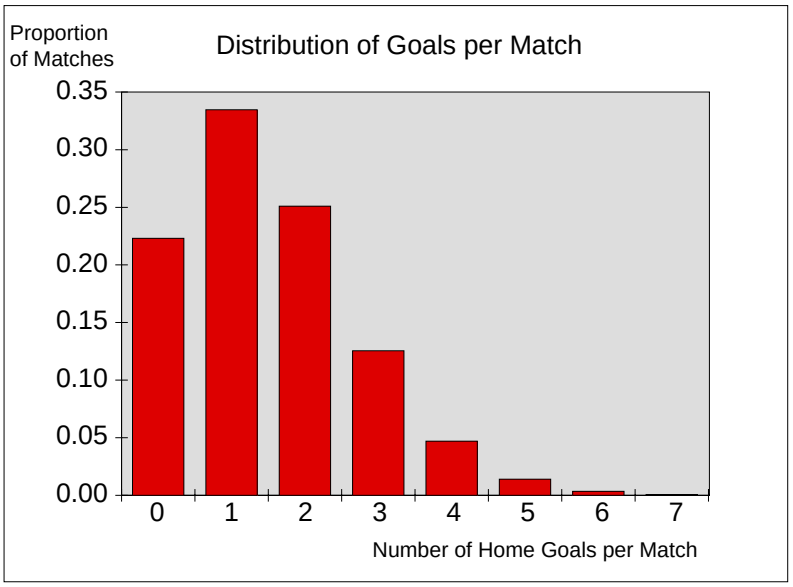
This fact alone is a great simplification of the massive amount of data. Many thousands of height measurements have been condensed down to a single meaningful number.

As a by-the-way I shall try to convince you that this one number, 66.5 inches, is useful. In a similar recent survey the average height of Western Adults was found to be close to 67.5 inches! Consideration of this one parameter has led us to the conclusion that, during the last 100 years, as a race, we have grown taller.

Skewed Distributions

The distribution of heights is symmetrical. For symmetrical distributions there is no doubt about where the middle is. If you are of average height, or very close to it, then about half the population will be taller than you and half shorter. Not only that but the average height is the most popular height!, By this I mean that a group containing all those people whose height is very close to the average height (say within a quarter of an inch) will have a greater proportion of the population than any similarly defined group of a different height.

A distribution is skewed if there is something which constrains the statistic from taking values outside a certain limit. In this, my second example, the statistic is the number of goals scored by the Home Team in a football match. The constraint is that the number of goals scored can not be negative!



My question is this: “Where is the middle of this skewed distribution?”.

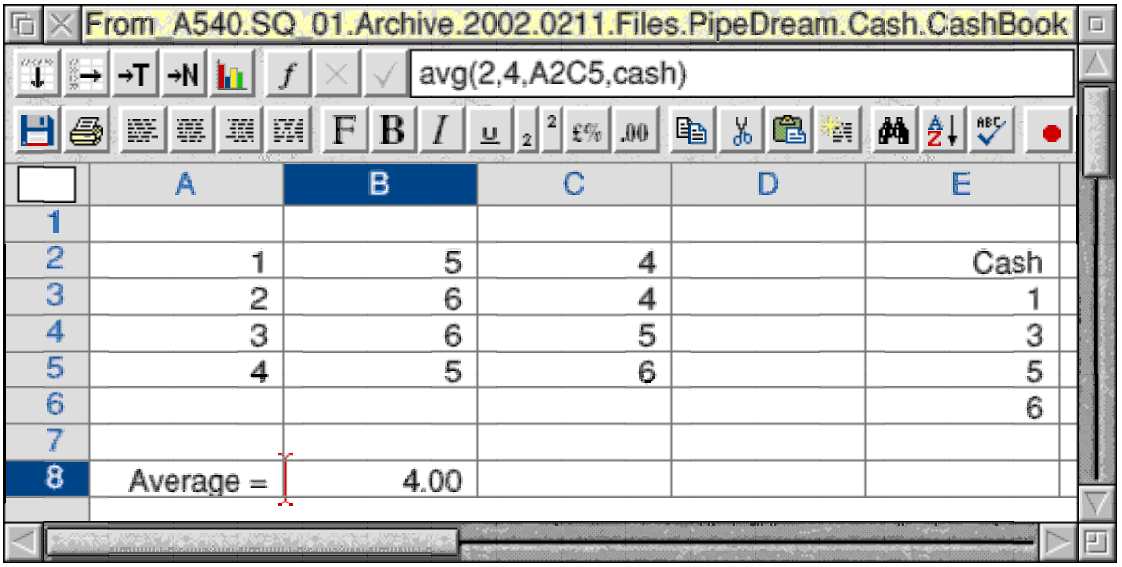
We shall return to this question later; first a digression—with a spreadsheet.

The Arithmetic Mean

One measure of the middle is called the Arithmetic Mean.

Sometimes a different mean is more appropriate, Geometric and Logarithmic means come to mind. However, when anyone refers to just 'Mean' they mean the Arithmetic Mean.

In nearly every spreadsheet package this arithmetic mean is returned by a function called Average. In PipeDream and Fireworkz it has the syntax avg(list). A 'list' can be a list of number such as avg(2,4,6,8) which will return the value 5. Alternatively a list can be a range of spreadsheet cells such as A2C5 or even an array (many numbers all within the same cell) or a range or array referred to by a Name. A list can contain a mixture of single numbers, cell references, arrays and Names.



	A	B	C	D	E
1					
2	1	5	4		Cash
3	2	6	4		1
4	3	6	5		3
5	4	5	6		5
6					6
7					
8	Average =	4.00			

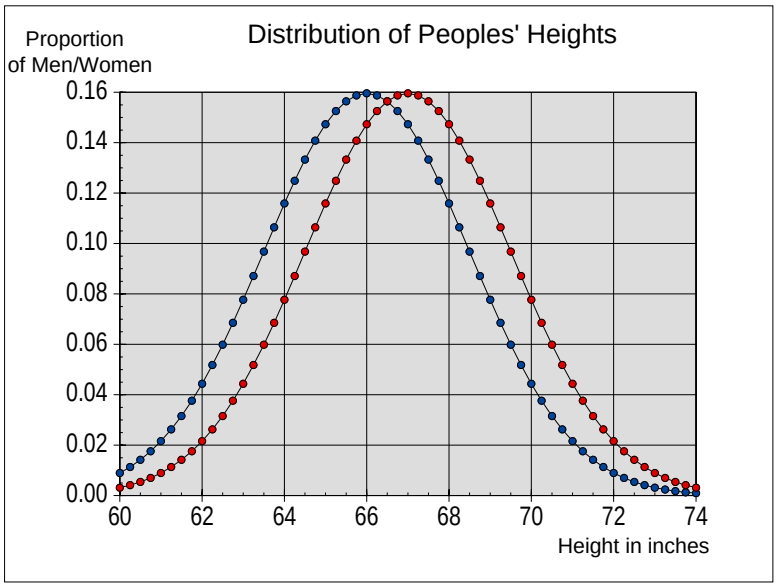
For example avg(2,4,A2C5,cash) where 'cash' is the Name of an array or block of cells, will return the arithmetic mean of all the values listed. In the screenshot above the average returned is the average of eighteen values: two are listed as 2 and 4, twelve are in the range A2C5 the last four are in the defined Name 'cash' (which refers to the block E3E6).

Men and Women

I would like to expand on my first example in order pursue the usefulness of having a number which describes the middle of a distribution.

Most of us know that men and women have different bodies. What you might not know is the information conveyed by the next graph. It is a graph which uses the data from the one hundred year old survey; the data has been split into two groups so there are two lines.

I have deliberately not labelled the two lines. You will see that the middle of the 'shorter' group of people is 66 inches and that of the 'taller' group is 67 inches. The more observant of you will have noticed the change of label for the y axes.



What was the ‘attribute’ which was used to divide ‘all people’ into the two groups? Yes! You’ve guessed it. One group is Men and the other is Women. The middle value for the height of women is 1 inch less than that for men. If you are at all interested in people’s heights (for example you might be in the clothing business) then this is a difference between the shape of men and women which you will find useful.

Although this original information is 100 years old, this 1 inch difference is still true today.

Football Scores

Have a look at the table below. It is a summary of the 380 games played in the Premier Division during the season 2001/2002 not by Win, Draw or Lose but listed by the number of goals scored by (F = for) the Home team and against (A) it.

Football Match Results for the Premier Division 2001/2002
by: Goals Scored For and Against the Home Team

	A	0	1	2	3	4	5	6	Tot	Poisson
F										
0		34	29	20	7	5	1	1	97	88
1		36	49	19	16	2	0	0	122	129
2		32	33	15	7	2	0	0	89	94
3		13	15	10	2	3	2	0	45	46
4		8	5	0	3	1	0	0	17	17
5		4	3	0	0	0	0	0	7	5
6		0	1	1	0	0	0	0	2	1
7		0	1	0	0	0	0	0	1	0
Total		127	136	65	35	13	3	1	380	380
Poisson		118	138	81	31	9	2	0	380	

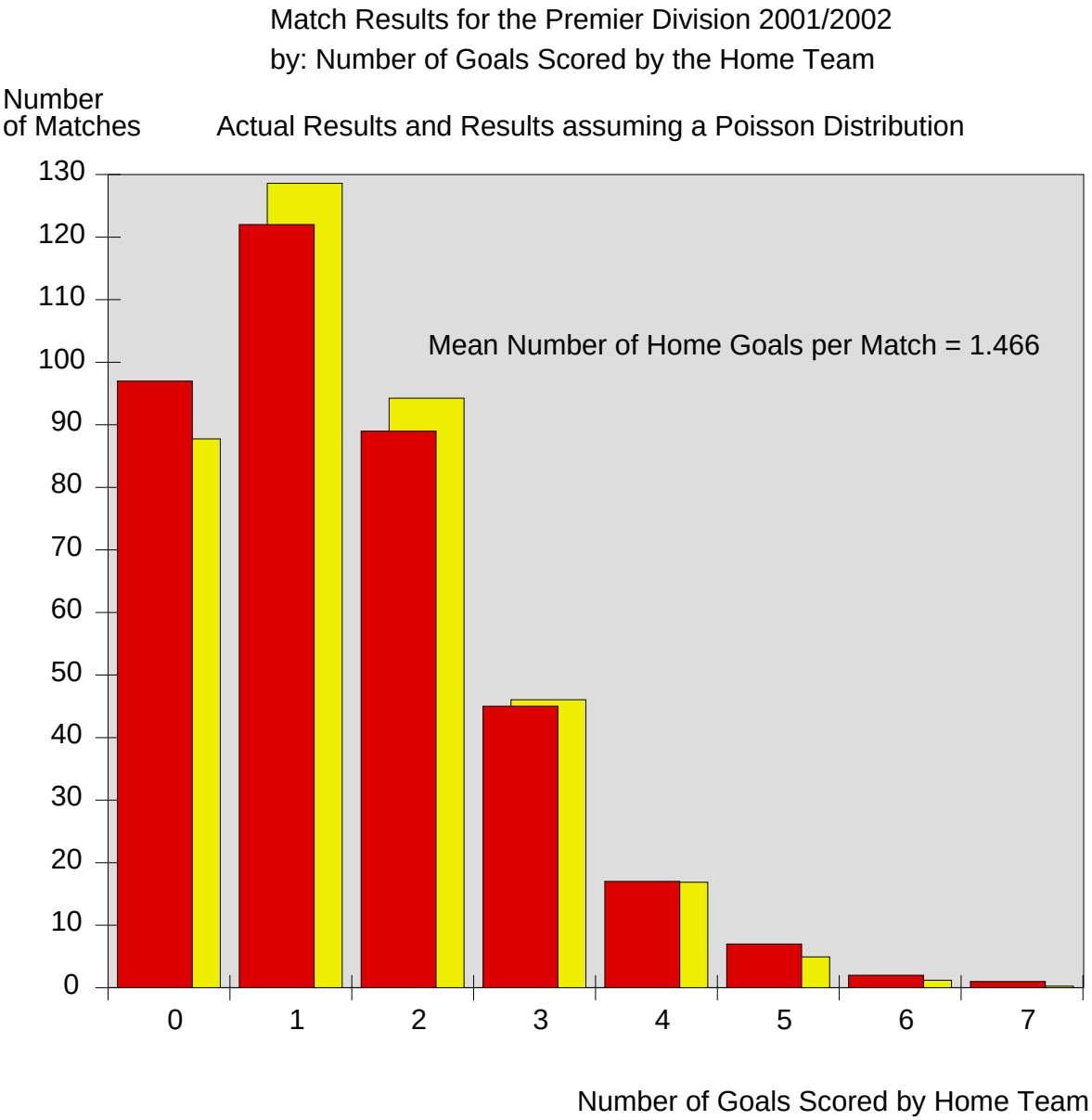
I created this table using Fireworkz. This is the way you read it.

First look at the column headed “Tot” to the right of the table. The first entry, 97, shows that, of the 380 games played during the season, in 97 of them the Home team scored zero goals. In 122 matches the Home team scored only one goal.

For every match during which one of the contestants is at Home there is another team playing Away. Of course there are 380 games in which one of the contestants plays Away. The row labelled “Total” near the bottom shows the distribution of Away goals. During 127 of the 380 matches the Away team scored zero goals against the Home team.

Home Goals

On the Fireworkz chart, two bars are shown for each ‘Home goals’ value. The bar on the left of each pair is the value corresponding to the actual number of matches resulting in the Home goals shown on the x axis. For example the first bar has a height of 97 because 97 of the 380 matches resulted in zero goals for the Home team.



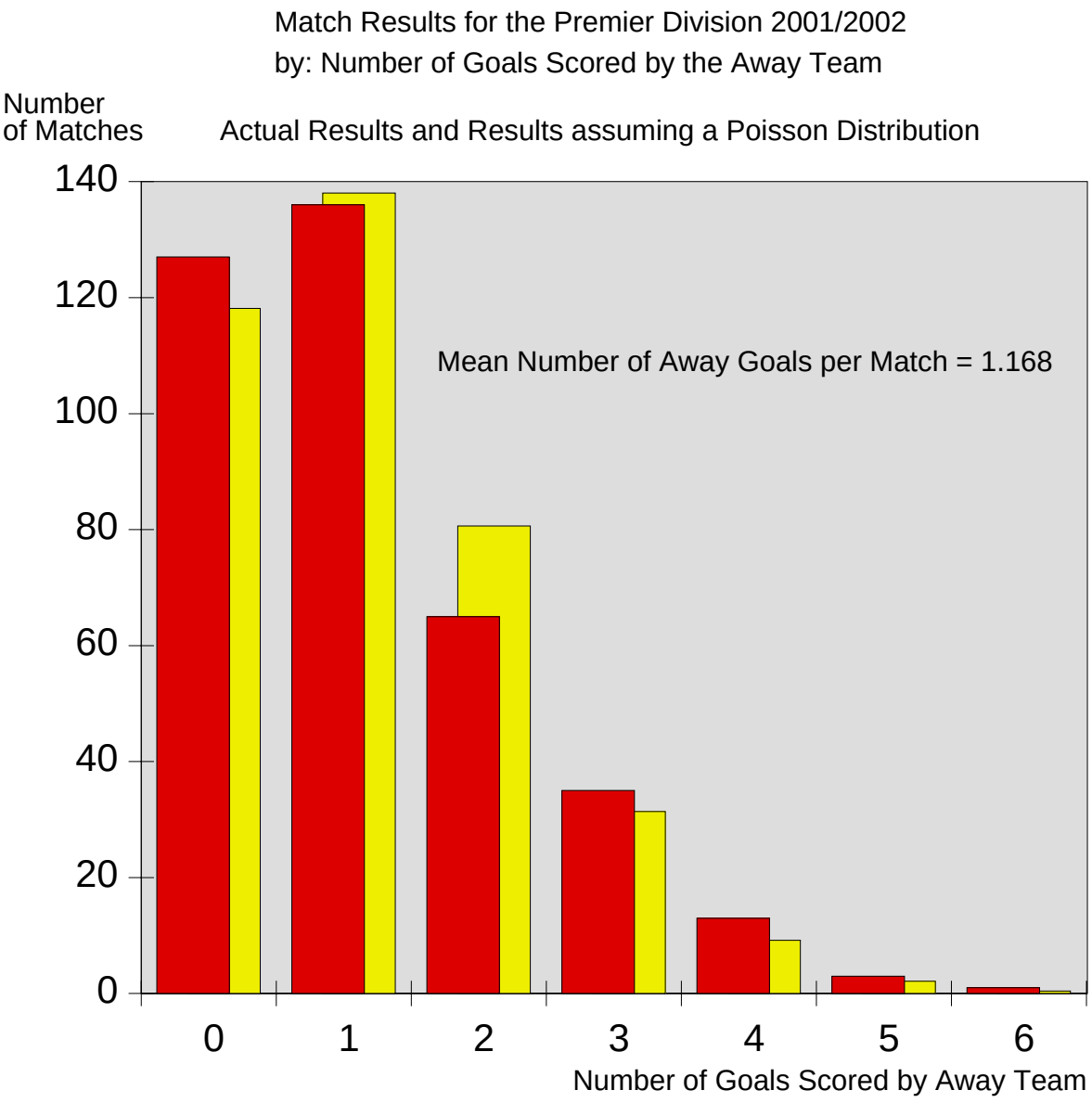
Back to the earlier question. "Where is the middle of the distribution?"
The Arithmetic Mean (which we met earlier) is 1.466 Goals per match.
The most popular score is 1 Goal per Match with 0 goals being more likely than 2 goals!
There were 380 matches played. Half this is 190 matches. Matches in which the Home team scored either 0 or 1 goals together total 97 + 122 = 219 matches. This is more than 50% of 380 Matches.

Certainly it would seem that the arithmetic mean (1.466), the most popular value (1 goal) and the 50% marker all give different values for the position of the middle.

This does not imply that a middle doesn't exist. It does imply that you have to choose which 'middle' value (not limited to the three I have described) is best suited to your particular (analytical) purpose.

Away Goals

All followers of football will know that playing at Home is advantageous.



How much is the ‘home advantage’ worth?

One very useful measure might be the proportion of home to away wins. In the example the number of home wins is 165 and the number of away wins is 114. I assure you that I do have an analytical method which starts with assumptions about goal distributions and predicts the number of home wins, draws, etc. It works out quite well over a season but not for an individual game. I don’t want to describe that method now.

Another useful measure is the mean number of goals per match scored at home and away. These values are 1.466 at home and 1.168 when away. One interpretation of this is that the ‘home advantage’ is worth abut 0.3 of a goal!

I remember doing sums similar to this a long time ago back in the 1950s. In those days the home and away average goals per match were 1.5 and 0.9 respectively. It would seem to me that over the decades the ‘home advantage’ has been considerably reduced principally because away teams are doing much better than they used. However, I have not done similar sums for the lower divisions and I do have some recollection that the ‘home advantage’ effect was stronger for the lower divisions than the top division.

Time for a Pause

Although there is a lot more which could be said about ‘middles’ of distributions, particularly skewed distributions, it is probably best to pause at this point whilst you have a little think about how to represent the middle of a skewed distribution.

Under what circumstances is the arithmetic mean the ‘best’ choice. Or should we choose the 50% marker? Or is it important to know which value is most popular? Drop me a line or email if you have any ideas. Certainly drop me a line if you have an example of a skewed distribution you would like to share with other readers.

Files Available

All the files referred to in this article are available. One of these is the full ‘game by game’ scores of all the matches played in the Premier Division for the 2001/2002 season. This ‘scores’ spreadsheet is in Fireworkz format. If you want it and can’t get a copy from the usual sources (web site, CD) then drop me a line and I’ll send it to you on a floppy disc.