In this practical example, we will explore several scenarios where understanding how a data set is distributed is truly beneficial. We will examine different data samples which follow a normal, a Student's T, and a Poisson distribution. Furthermore, we will analyze instances of exponential and binomial data to help us appreciate the elegant statistics these distributions possess.

Imagine We are working as a head project manager for one of the most renowned companies in the world of video games, EA Games. Wer various responsibilities include:

supervising the development and release of the 2018 edition of the soccer game titled "FIFA 19".

Above all else, We need to ensure the game is well-rounded and provides a genuinely enjoyable experience for all the customers. The game has a professional competitive scene so it needs to be balanced. By balanced, we mean that no team or individual player should invariably be a preferred option regardless of the opposition. Therefore, we expect to have an equal number of good layers and poor players in the game.

We provided We with access to a data set containing the stats for each individual player in "FIFA 19".To begin with, examine the overall column. It represents the quality of a player in their natural position on a scale from one to 100. This value is a sort of weighted average of the many dividual stats each player has. As We probably know, the importance of attributes varies for different positions in the field. For instance, acceleration and top speed are more important for a winger than tackling. However, the inverse is true for centre-backs. Thus, we alter the weight for each stat

based on the position of the player. Therefore, we do not have a single formula which calculates the overall evaluation.

To get an idea of how well distributed the overall values are, we can construct a histogram and set the bin size to one. We do so by selecting the overall column, clicking on insert, then inserting the statistics chart and selecting histogram. To adjust the size of the bins, Right-click on the x-axis of the graph and press the format axis before setting bandwidth to one. The graph is bell-shaped

and resembles a normal distribution.

* But wait, aren't we dealing with discrete values?
* How can this be a normal distribution?

Although, although that may be true continuous variables can take discreet values but not vice versa. Furthermore, since we are dealing with rounded averages we are inclined to believe that the overall value is not entirely discreet but rather an approximation. Let's take a closer look at the graph. Now we can notice its thin tails which suggest a smaller number of outliers. This reflects real life quite accurately since very few professional players are exceptionally good or bad at every single aspect of the sport. Besides even the least skilled professional soccer players are far superior to the average person. That explains why the lowest overall values start from around 50 rather than zero. The stats should reflect the performance of players in the real world. As normal distribution is the most frequently observed in nature, it is only logical that the data resembles this distribution. Moreover, the bell-shaped graph with thin tails further supports this idea. Since one of the main characteristics of a normal distribution is symmetry, the overall values are symmetrically distributed. Thus, we can safely consider the game balanced and acceptable for competitive play.

It is also worth noting that players within a single team or division share similar stats. This skews the data a certain way and explains why we cannot expect the values to follow a normal distribution.

Now if we wish to further test the balance of the overall stats, we can examine a small sample of random players. For instance, we can construct a histogram of the first 30 players in the data set

based on their ID number. Since our data is limited we need to adjust the size of the bins, otherwise, it is possible for each value to occur only once or twice. That would result in many bins of one or two and make the histogram redundant. If we adjust the bin size to three, we will see that the graph slightly resembles a normal distribution. However, we will also notice the fatter tails.

Since the number of observations is limited we can safely consider this sample follows a Student's t-distribution. Recall that the Student's t-distribution is also symmetric. So we are confident that even the small sample we are examining confirms our goal of a balanced game.

Before we move on to other aspects of the development of the game, let's explore how a single stat is distributed among the players in the game. Take the shot power column for example. If we construct a histogram and set the bin size to one, we will see a distribution with two peaks. It resembles two graphs placed side by side. A way to interpret this is having two distinct groups of players, one with a mean of around 21 and another one with a mean of around 65.

The reason behind this is the presence of goalkeepers in the game. The stats important for them are completely different from the stats essential for outfield players. Thus, it only makes sense

that they will have distinctly lower values for many of the non-goalkeeper-specific stats. If we examine a goalkeeping trait like GK diving, we will be able to see the division into types more clearly. We have two completely different clusters. The low value represents how outfield players

would perform in goal... And the higher one represents the actual goalkeeper’s performance.

If we only examine the goalies we will see the values are normally distributed once again so the game is indeed balanced.

Another aspect which meets the game more enjoyable is creating a sense of realism. For instance, Weng professional soccer players outnumber veterans. ***First***, a significant number of promising Weng players suffer bad injuries, which significantly slow down their progress or even halt it altogether. ***Second,*** some are forced to retire while others simply decide to quit after spending too much time off the field. Last but not least, Weng players who are not given the opportunity to play often decide to go to university instead of pursuing a career in soccer. All of these factors lead to attrition which results in having fewer players above the age of 35 than players below the age of 20. To make sure the game captures this aspect of the sport check out the age column. Once again, we can construct a histogram and set the bin size to one. We already demonstrated how to do this for the overall column, so just follow the same steps. By setting the bin width to one every age gets represented by a separate bar on the graph. Age is a discreet variable representing the age of each player. In addition, age has a minimum value of 16 since the game only consists of first-team players who have signed a professional contract. Thus, We can consider 16 as the starting point for any player who can sign a professional contract. We may view it as sort of an origin for a Poisson distribution. Then each bar in the graph would showcase the likelihood of a certain player within the data to be a specific age. Since a Poisson distribution is skewed, the Wenger players outnumber the older ones. As we mentioned before, that is also true in real life.

Therefore, this creates an additional layer of realism to the game and should make it more enjoyable for the customers. Do We remember that as a head project manager, apart from the development of the game, We also need to supervise the official release? One of its most important aspects is social media marketing. Now, imagine Wer main competitor is trying to expand their customer base by uploading free video previews of their new games each Monday prior to their launch. A month ago, We assigned one of the interns to keep track of the progress of their views. We can find the recorded viewership values in the Daily Views Excel file accompanying this lecture. Before we proceed with the analysis, I recommend that We download and open the file. Okay, the Excel file contains a single sheet titled Views, which comprises two columns. The first one indicates the number of days post-release when the value was recorded. The second one shows the number of views since the last check. To get a better understanding of the data We would wanna see how viewership changes over time. In order to do so, We decide to graph the data set. The easiest way to do this is by marking columns A and B and clicking on insert. The next step is going to charts and selecting a scatter plot. Since most of the views occur within the first few days the graph starts off at a very high point and drops down rather quickly.

We can see that daily views start at around 100,000 but fall to about 20,000 within a week. Once the new video is released and promoted viewership drops to around 10,000 per day and steadily decreases as it loses relevancy. By the time a second video has been released around the 14th day, the video gets barely a few thousand views per day. This kind of behaviour resembles an exponential distribution. To check how accurate our assumption is, we can select the chart elements button and select a trend line. If we do not specify the type of relationship we expect, Excel is going to assume a linear one and create a straight trend line. Since this distribution resembles an exponential one we pick an exponential trend line instead. The curve of the trend line fits the data points accurately. If we assume that the views in fact follow such a distribution, then the trend line would represent the PDF for a view occurring on a specific day.

To test whether views really follow an exponential distribution, we should look at the CDF graph as well. We can graph the relationship between the first and third columns. Since total views represent the cumulative number of views up to a given period in time, it shows the aggregated number of views the video got.

Let's create another scatter plot following the same steps as last time. We can notice that the curve goes up at a decreasing rate before eventually plateauing. This also matches our expectation of the CDF of an exponential distribution. Now that we know the viewership fluctuates each day we can state that each video loses relevancy rather quickly. This means that such a campaign is only beneficial in the short term. Therefore, We advise the marketing team to release similar videos only during the last month before launching the game. That way, all the videos will generate enough attention to make the game feel immense without losing customer interest in the process.

In addition to competitor analysis, We need to conduct some customer analysis as well. We certainly care about which of our clients can afford to spend more on in-game purchases, so We send out a survey. One of the survey questions is whether the customer is a premium member of the official fan club of any team in the game. Since these fans are more devoted and financially capable, We want to find out if there is any other feature We could use to target this group. We decide to examine a small sample of the data which contains the age of the customer according to their EA sports account and whether they are a premium member. This data is stored in the Customer's Membership Excel file accompanying this lecture. After opening the file, we see two columns, one with numeric values and the other one with ones and zeros. The first column represents the age of the customer. The second one shows whether they are a member or not.

If the customer is also a member of the fan club we put one in the second column. Alternatively, if they are not, we write down a zero. Now, if we construct the scatter plot we are going to see that most people under the age of 34 don't have a membership. Whilst most people over the age of 34 do. Of course, there are exceptions to this rule, which is normal when we are dealing with real-world data. That being said, the data looks like it follows a logistic distribution since the likelihood of having a membership sharply rises after nearing a specific value. In this case, we can think about 34 as the location of the distribution. This leads us to believe that 34 is the approximate age

at which customers have already reached financial stability and can afford higher membership fees. This insight suggests we should target customers above the age of 34 since they're more likely to spend more. One way to use this information is to release more expensive legend FIFA ultimate team cards for players who have retired in the past 20 years. Fantastic work.

In this lecture, We were able to see numerous examples where knowing how to deal with distributions is truly beneficial.

We developed an understanding of the practical aspect of probability, and discovered why knowing how the data is distributed can help us make correct business decisions.