Transcript for Session 009

Listen to the podcast session, see resources & links:

http://chandoo.org/session9/

Transcript:

Hello everyone. Welcome back to chandoo.org podcast session 9. This show is all about making you awesome in data analysis, charting, dashboards and VBA using Microsoft Excel.

Before we begin our show, I just want to take a minute and thank you so much for taking time to learn from me and through this podcast. The fact that you're taking the time to learn, sharpen your skills and understand how to analyze data better, shows that you are really an awesome person. Thank you so much for joining us.

Before we move into the show, I just want to share a quick announcement with you. Those of you who are listening to this podcast might find this announcement a little meaningless in the future. Nevertheless, I must share it. As a family, we have a tradition of going on a long trip somewhere in the month of May or June - it could be a road trip, a vacation to different country or just visiting some relatives or friends who live in a different city. This year as well, we are going on a fairly long road trip and we'll be away for about ten days starting day after tomorrow. I'm really excited as this is the first time that we going on a long road trip since the time we bought a new car in February this year. So, my wife Jo, and my kids Nishant and Nakshitra (who are about 4 years old) are really excited and are looking forward to the next ten days visiting some places that mean a lot to us, watching new sights, eating different things and probably having a good time. I'm saying probably because this is also the first time that we are going on a long trip with the kids in India. We have done that when we were in the US last year, but that's a whole another thing because the US has a freeway system and everything is really smooth, clean and easy. Whereas in India, even though the highways are pretty good, I feel they are quite adventurous as well.

Anyway, I don't want to share too many personal details here, but this is the time of year when we go on a trip and I'm really excited and hoping to share some thoughts with you after the road trip in a different podcast episode as well.

Let's move on to our episode for the day. This is something that I have kind of observed for a while and I always teach my students in a live setting when this happens. This is the topic of averages. Averages are a very popular and famous way to analyze data. Whenever an Analyst is asked this question, "hey can you analyze this data and tell me what's in this data" - the first instinct for us as Analysts would be to calculate the average of the data. It doesn't matter whether the data set has 3 data points or 3 million data points; our instinct is always to calculate the average for it. That's why I chose this as a topic. I want to bust some of the myths that surround averages and I want to share some information about how to analyze data better.

Averages are a good start but there are better ways to analyze data. So how do we go from average to 'above average'! That's why I titled this episode as 'averages are mean'. What I really want to convey by mean here is that averages are stingy as they don't reveal much. When you look at a person and you say that he/she is really mean, we convey that the particular person is probably possessive or not able to give out much. They take a lot but are not able to give out much. That's why I say that averages are mean! I know it gets a little bit tricky when we have to use the word 'mean' in several different ways. The word mean can convey that somebody is stingy, greedy or something like that. It can also convey the word average. Average is also called mean. It can also convey a definition as such. So, I don't want to play with this word too much because it can get tricky in a podcast setting. It's different in a live class as you can see my face and decipher what meaning I am trying to convey.

Anyway, coming back to the topic of our podcast - why do we say that averages are not a good way to analyze data. Before we understand why averages are a poor way to analyze data, let's first spend a minute or two to understand what average really is. I'm sure that by now all the people listening to this podcast would have calculated an average of some data points at least once in their life. This is something that we learn as kids in school as part of basic mathematics. As you progress into higher studies, you also learn a lot about averages in courses on statistics, statistical analysis, theory of probability, distributions etc.

The concept of averages is something that is fairly ingrained in us - we all know what it is and we know what to do when we're asked to calculate the average for a bunch of values. For the sake of repetition and setting common ground, let's take a very simple data set and calculate the average. I'm sure that at this point you're saying, "Oh, Chandoo, come on, we know how to calculate average!" Let's keep it really simple and straight-forward.

Imagine that you're looking at the sales data of five employees in your company. You're their boss and you're asking what their average sale amount is. These five people are John, Maria, Matt, Jason and Cheryl. Their names don't matter. What matters are their numbers! For the sake of simplicity, let's imagine that John has sales of $100, Maria has sales of $120, Matt has sales of $150, Jason has sales of $200 and Cheryl has sales of $130. You're looking at these numbers and asking what the average sales of your employees are. What would you do? You'd calculate the total of the 5 numbers and divide it by 5.

So, the definition of average is the sum of values divided by the count of values. In this case, the average turns out to be $140. As an Analyst, what do I get by looking at this average? In a report, presentation, dashboard or anywhere else in an analysis context, you're saying that the average of the values is $140. But what do we really mean by it? Well, we obviously mean that if we took the sum of the values and divide it by the count of the values, I get $140. This is the standard mathematical definition of average.

Let's come to the business side of it. As a decision maker, what does it mean to me? For example, I am a department head who manages these five people and I want to understand what it means to me when someone tells me that the average of my employees is $140. We might be able to go into the semantics and split hairs! But, purely from a managerial or business point of view, it seems to me that the average is pretty much meaningless. I know that's probably blasphemy for some people to think of it like that. But, to me, in this scenario where we have 5 people and we are calculating their average sales, the number 140 does not convey much.

I mean it does convey a few things. For example, given the fact that we have five employees and someone tells me that the average is $140, I automatically get to know that we made total sales amounting to $700 ($140*5). I get to derive some information from the average, so it's not purely meaningless. It has some significance. But, all of that is just stating the same fact in another way.

Back in the days when I was doing my MBA, this is what we used to call meaningless CP. Let me tell you a small interesting story. In the days when I was doing my MBA, whatever grade I got was completely dependent on the exams that I took. For example, if I got 89% in Mathematics in High School, what it really meant is that I scored 89% by answering several questions or problems in class tests, unit tests or final examinations. In High School or undergraduate courses, you are given a grade based on tests or laboratory experiments. But, in MBA, for most students in India and around the world, some percentage of the grade is allocated to class participation. When class participation is a criterion, they want us to actively listen to the lecture that is being delivered, ask some interesting questions or participate in a lively discussion in the class so that they can give us some extra points. Just like anything else in life, when someone puts weightage or a reward on class participation, it's a purely qualitative thing. There is no way to say that person X's argument was way better than person Y's argument. A lot of other things come into the picture. It's completely qualitative.

If class participation is given weightage and I don't speak up in the class, I would be considered as somebody who had no class participation at all and hence my grade would go down by 15% or whatever the weightage may be. So, I'm compelled to speak in the class even though I don't have anything to contribute. Here's something that really happened in my class - somebody would say something like the average sales for ABC company is $3 million for the last 4 years. The professor would be teaching something and this person would raise her hand and she would say this. Another person who's also compelled to say something to get the grade point, would raise their hand and they'd say that means that the total sales are $12 million. Do you see what they've done? They've taken the same numbers - $3 million and 4 years - and multiplied them to give another fact. In reality, both of these are pretty much meaningless. Maybe the first one was okay, but the second one is simply a re-hash of it.

This is what averages are also doing in the business sense for us. They don't present a valuable insight in many contexts. There are some places where averages work beautifully. But in many contexts that I've seen, they present very little information. That's why I wanted to spend a podcast episode busting some of the myths around average and suggest some better ways to analyze data. Now that we understand averages and you've got a peek into the miserable part of my MBA life in the form of forced class participation, let me tell you a few other pitfalls of averages.

This is the time for a memorable quotation. I don't remember where I heard it first, but it's a really clever way to understand why averages suck. -- "One in every two averages is useless!" -- I don't remember, but it's something that I probably heard in a witty presentation or something like that and it stuck with me. That kind of signifies how useless averages can be for many situations.

Again, I want to be clear that there are some places where averages can be really useful, but for most business scenarios, I find that average alone is just like garbage. You want to add detail to average so that it becomes meaningful.

Moving on, as you can see even in our example of 5 sales people where we said that the average is

$140, averages tend to hide a lot more than they reveal. So, what did average do in this case? It took those 5 numbers - 100, 120, 150, 200 and 130 - and it kind of gobbled all of them up. If we imagine average to be a black box, it took all these 5 numbers and it spit out one number in the form of 140. What 140 really tells us is that smoothens out everything. It took 200 which was the highest number reported by Jason and it took 100 which was the lowest number reported by John, and it kind of merged everything together into one number and give it to us. That's what averages do. They reveal very little compared to what they hide.

If I have to compare averages to something in real life, I would compare averages to skirts. They reveal enough to raise your imagination, but not everything! Again, many people say the same thing about statistics also. And, average is nothing but one of the most used and most popular statistic. So we can say that averages are like skirts and move on! That's a little about the drawbacks or pitfalls of averages.

There are many scenarios where averages are flat out useless. Even in the case where we took the example of 5 people in the sales department and calculated the average, we could technically argue that average is okay in this case. It's not really brutal. But imagine the same sales department with one star employee who's a go getter and would go and score 5 or 10 times more orders than the rest of the team because she has these super awesome capabilities. So imagine that you are managing a team of 5 people and one of them is a rock star like this. Then the numbers would look something like this - 100, 120, 150, 100 and 1000. As you notice, one of the numbers is way off the charts. What would happen if you calculate the average now? You won't have any idea about the significant performance of that person. You've kind of squished everything together and when you calculate the average, it looks like your average sales is somewhere around $280. But the reality is not that. Nobody made a sale of 280. Four of your employees have made a sale of less than $200. And there's only one employee who made a sale of $1000 and because of that employee, everyone else's average is moving away from them. These kinds of things are completely hidden by the average.

So, you want to make sure that when you're using averages, you're using it for data that has some sort of closeness to it. If the data is far apart, then average won't work very well.

This is where it is important for us to understand a few more concepts. You know the concept of average. That's good. But, as an Analyst, it's important that you also learn some other statistical or data analysis concepts.

The very first concept has to be standard deviation. If you've never calculated standard deviation by hand, then this definitely sounds like a technical term. It's as good as any of the jargon terms that we hear in any popular media or research journals. But, I assume that most of you have calculated standard deviation at least once in your life. I've calculated standard deviation as part of the mathematics classes' back in school, college and even in statistical courses in my Engineering degree as well.

But what I fail to understand is what standard deviation means. I know what standard deviation calculation is and I can calculate it. I mean if you give me a paper and a bunch of numbers right now and tell me to calculate the standard deviation, I won't be able to calculate it accurately but I kind of know my way around it in terms of the calculation. But, as a businessman or as a manager or as an Analyst, I know it now but I used to have a very poor idea of what standard deviation really means. So let's understand it from a business point of view. Let's take the example of the five employees we were

talking about. Earlier we said that one of these employees is a rock star and she scores much higher sales than everybody else. So because of that number our average is really skewed. It moves away from everything else. In this case we have an average of $280 but four employees are less than $200 and one employee is more than $100. When you add everything together it makes the total average to be $280. When I look at the number $280, I have no idea how these numbers deviated from the average. That's what standard deviation tries to explain.

The calculation part of standard deviation is something that you can understand, so let me explain the meaning part of it. As an example, let's say that the standard deviation in this case is 100. You have 5 employees, the average sale is 280 and the standard deviation is 100. Looking at these two, you can easily see that the individual five numbers deviate from the average of 280 by an amount 100 in various directions. This is a very loose way of explaining it but it helps you visualize it.

If you have another five set of employees and you calculate their average which also turns out to be 280, but their standard deviation is 10. So you have two data sets - one data set has an average of 280 and a standard deviation of 100, and another data set has an average of 280 and a standard deviation of 10. Looking at these two sets of statistics, you can immediately tell that the second data set (with an average of 280 and standard deviation of 10) is far closer to each other. Those numbers are very close to the average of 280. So you can be confident that there are no extreme values in the second data set. Whereas the first data set indicates a lot of jumpiness; things are all over the place in the first data set. That's why the standard deviation is more. That's what standard deviation signifies from a business perspective.

I may have botched up the definition of standard deviation from a statistical point of view, but from a business user, Manager or Analyst point of view this is how I try to understand standard deviation. When the standard deviation is way too high it means that I can't trust the average anymore. When the standard deviation is close to zero it means that I can rely on the average to sensibly understand the data. That's the very first concept - the standard deviation.

By now you already know it, but it’s important for me to say this. The standard deviation, average and a few other concepts that we are going to talk about can be easily calculated in Excel using simple formulas.

The formula for average is:
=average()

The formula for standard deviation is:
=stddev()

They have a bunch of other formulas for standard deviation. They have fixed some numerical and floating point errors and other things in newer versions of Excel, but the ‘stddev’ formula works pretty much fine.

You may be curious to learn a little more about standard deviation, so I'm going to link to an article. This is an excellent article written by my friend and colleague, Mike Alexander, who you may remember from one of the earlier podcast episodes where we talked about BI for masses. I will link to Mike's article

about standard deviation and how to understand it, how to understand it for a sample versus population and all those statistical things. I will link to it in the show notes. You can get that from http://www.chandoo.org/session9/. Standard deviation is the first and very important concept. However, if you add these statistics to a report - for example average 280 and standard deviation 100, it kind of boggles, confuses and throws people off-guard, especially if you take this to a board room meeting. The chances are high that people will keep asking you what the standard deviation means. Not many people know how to interpret standard deviation from a business point of view. Again your industry or bosses may differ, but from my personal experience, I have very rarely come across bosses who know what standard deviation is.

As an Analyst, it's important for you to know what these things are and how to present them. If you take the same case of an average of 280 and a standard deviation of 100, as an Analyst, I won't put these two numbers in the dashboard. That's because from these two numbers I can see that the values are all over the place and so there's no point in calculating an average. Instead I'll do something else which I will reveal in the latter part of this podcast. So just stay tuned.

The second technique is median. Median is a very powerful way to analyze data. It is similar to average but it's a lot better for many situations. Median is the middle point of the data when you sort it in either ascending or descending order. If I arrange the five employees by the total sales that they have made and I pick the middle value, it is the median. In our original data we have these five numbers - 100, 120, 150, 200 and 130. If I have to re-arrange them then 130 would be the middle value. It's the third value and that's what I would pick as the median. You might be wondering what happens if there are two values in the middle, i.e. If we have six employees then two of them are technically in the middle. In this case, the median would be the average of those two values. You'd just take the two numbers, sum them up and divide by 2.

When I look at the median sale of 130, it's 10 less than the average sale of 140. I can immediately understand a couple of things when I look at the number 130. I know that half of our employees have made sales of 130 or less. So three employees have made a sale of 130 or less and three employees have made sales of 130 or more. So the median gives me a far better picture of the spread of the data, i.e. how the data is and where the middle point is. Again you can't tell a lot with median alone but it is a fairly good indication of how the data is. The median is also a very good way to explain the data when you have a lot of numbers. If you just have 5 numbers in business scenarios, I wouldn't even bother with the average or median. I would just go and present all the 5 numbers because we can squish those 5 numbers into a slide, chart or report. Whereas if you are analyzing the number of times people have rented a particular genre of movies at Netflix - for example you are the head of the Netflix comedy movies and they probably have thousands of movies in the comedy genre - you're looking at these movie listings and you want to know how many times people are renting, downloading or streaming the movies. If you calculate an average it'll give you a pretty much meaningless number because there will always be movies that have been streamed or rented zero times or 1 time and there will always be movies that are probably streamed millions of times. A really funny movie like 'There's something about Mary' would have been streamed millions of times whereas a poor funny movie which is pathetic, stupid and meaningless would probably have been rented zero times.

We have zero at the bottom and a million at the top, so that would tell me that on an average a comedy movie is streamed 500 times a week. But that's just meaningless and an incorrect way to look at this

data. When you have so many numbers you cannot put all of them in a slide as it'll run into 340 slides! The median would be perfect in such cases. You'd just sort the list in ascending or descending order and pick up the middle value. The middle would be a mediocre comedy movie and you would get a sense of how many times that is rented. And that'll give you a fairly good indication of how many times you are renting comedy movies. So that's the median.

You see median applied quite often in many situations, both in business and economic analysis. A very common example is housing prices - if you look at any house price portals, or if you read in economic newspapers or websites, you always see the statistic – 'median home price is $120,000'. That gives you a sense that half the houses are below $120,000 and the other half are more than $120,000. And that gives you a picture of how much a typical middle class family's house is worth.

But as I said the median is just one data point. In the case of Netflix or house price sales we took thousands or millions of data points and we kind of squished them into one number. Whereas there could be slightly more numbers that explain the pattern of the data. That's where the next concept comes into the picture which is called quartiles. Median explains where the middle of the data is; you can also think of median as the 50th percent value of the data. Again I am using very loose terminology here because if I go with strict statistics, I have to say that median is the 50th percentile of the data. However, from a business point of view, not many of us use the word percentile very often whereas we are familiar with percent. So let's just go with that. Median is the mid-point or the 50th percentile of the data.

Quartiles explain two more numbers. If I say that the median house price is $120,000 and the 25th percentile or the 1st quartile is $75,000, by that I mean that half of the houses are less than $120,000 and the first 25% or the first 1/4th of the houses are less than $75,000. If the entire data is chopped into four equal blocks, you get a sense of where each block is ending. That's how you would interpret the statistic of the 1st quartile being $75,000, i.e. one-fourth of the houses are less than $75,000 and the rest of the three-fourths of the houses are more than $75,000.

Then comes the second quartile which is nothing but the median.

And lastly comes the third quartile which is the 3/4th percentile or the 75th percentile. If I say that the third quartile is $300,000, I mean that the first three-fourths of the data is less than $300,000 and the last 25% of houses are more than or equal to $300,000.

The quartiles, along with the median, give you a better perspective of the data in terms of how the data is spread across. When you look at these three numbers - first quartile, median and third quartile - you can understand a lot more about the data - how the data is spread and where the middle chunk of the data is. In normal population or economic lingo, this is called middle-class, i.e. where the middle or middle-class of the data is can be easily understood when you look at this quartile distribution.

Along with quartiles, let's say you are calculating all this data for the county, town or city where Bill Gates is living. Bill Gates is a really rich person and so obviously his house is also part of this listing. Not that he's going to sell his house, but you're calculating the house prices and doing some analysis on it, so obviously you would include Bill Gates's house. Maybe he has multiple houses, so you'd put all of them in this list.

And, he's the richest man in the world, so the cost of his houses isn't going to be $300,000. Of course I'm not sure, but to me it seems possible that he has a couple of houses in this list that are worth millions of dollars. So you're doing calculations with all these numbers but you're not explaining the extreme values in the list of numbers. If there is a house worth $35 million for one of the mansions of Bill Gates and it's part of the list, that's where the fourth concept called 'outliers' comes into the picture.

Outliers are the extreme values, i.e. the values that are too high or too low. In the case of our example, a value that's too high could be the price of a house that belongs to Bill Gates and a value that's too low could be a house priced at $5 because somebody could not repay a loan taken on the house and so the property was foreclosed and the bank put it up for auction but nobody was willing to buy it, so the house was eventually listed at $5. Anybody willing to pay $5 and assume the property, pay off the taxes and any other outstanding amounts on it could purchase it. So this would be an example of the other extreme. In reality, you don't really find such drastic values. But in many cases, especially in cases of property prices, employees' salaries or any of those kinds of analysis, you are likely to have these extremes. You will have someone who is only making $15,000 per annum and you'll also always have someone who is making $15 million per annum. Those two extremes co-exist thanks to the way things are working in this world. We call these extreme values outliers. As an Analyst, it's important for you to understand what the outliers of your data are and maybe exclude them from your analysis.

There is no point in calculating the average price of homes if you have Bill Gates house and the foreclosed house in your data set. You might have to exclude them because they are the kind of values that need to be treated separately. It's the same with our sales data where we had our rock star as part of the team and she was always scoring thousands of dollars' worth of orders while everyone else was struggling with $200 or $150. From an analytical point of view, you would analyze the normal values, i.e. the values that are close to each other, and you would exclude the outliers and present them in a different way. For example, you might want to set up a training session with the rock star so that she can train others on the ways of making better sales. Or, you can sit with her and find out what is working for her and then mentor others according to those directions. So these are the kinds of treatments that you want to try to do with the data – you want to treat any numbers that are outliers separately. That's why it is important to understand the concept of outliers and identify them in your data.

Now we come to the last concept which is the distribution of values. This is very important especially in business scenarios because the kind of analysis you'd do depends on the type of distribution of the data. Again this sounds too scientific and too much like jargon, but what distribution really means is how the data is spread – is the data within a range of values, what are the minimum and maximum values, how would it look if I plotted this data etc. The spread of values is already partially explained by the quartiles, median and the outliers that we have already talked about. When I combine the quartile, median and outlier information and if I try to depict it in a simple diagram, it would give me a sense of the distribution of values.

This is where we use quite a lot of tools – we can make a chart, or a box-plot which is one of the very popular ways to explain how the values are distributed. It would be tricky to explain how a box plot looks in a podcast, so I encourage you to visit the show notes page where I will put an image of a box-plot and also a tutorial of how to create it from your data using Excel. You can visit the show notes page at http://www.chandoo.org/session9/.

In a nutshell, that's the common sense that you need to grasp as an Analyst to analyze data better. These five concepts are:

- **Standard deviation** – this explains how data is spread across from the average. Standard deviation explains how the data differs from the average and how it differs from the original data. It's something that you want to keep at the back of your mind. There's no point of including it in your presentation as it would mean nothing to a lot of business users. You should understand it and see if it explains anything to you as an Analyst, for example if the standard deviation is way too high it means that there would be no point looking at the mean, instead you should look at the distribution of the data which includes the it's spread, quartiles, median and outliers.
- **Median** – this is the middle point of the data. It's a better way to understand data as compared to average.
- **Quartiles** – these are the one-fourth and three-fourth points of the data once you sort the data.
- **Outliers** – these are the extreme values in the data set and include both extremely low values and extremely high values.
- **Distribution** – this explain how the values are distributed.

I am trying to explain these concepts using very vague and loose terminology so that you can appreciate them as a business user or a budding Analyst and understand the significance of them. When it's time to implement these concepts in your reports and analysis, it is always advisable to spend some time learning these concepts in a little more depth. For example, I've explained the concept of standard deviation in really plain English, but it's important for you to understand what standard deviation means and how to interpret it for your data set. For example if you're dealing with data in millions then having a standard deviation of 100 is really nothing, because on a scale of millions, a value of 100 is almost like 0. So it's important to understand how to interpret these things within the context of your data and business.

These are the concepts that you should be familiar with as an Analyst. Initially, I wanted to have a session in this podcast explaining how to analyze data with these concepts using Excel, but now that we have run beyond 40 minutes already, I'm thinking of doing a second episode of this podcast that I'll release soon.

Probably in the next episode or the one after that, we'll talk about Excel techniques for analyzing data using all these statistical or analytical concepts that you've learnt so far.

In a nutshell, that's why averages are a poor way to analyze data and how to move from average to other better ways to understand and analyze data.

In the next session we will talk about some of the Excel techniques that you can apply to analyze data.

Thank you so much for tuning into this podcast. Please visit http://www.chandoo.org/session9/ to access the show notes, resources, a full transcript and all the links and ideas that I have mentioned. On the show notes page you can also find a link to Mike Alexander's standard deviation explanation, an

example of box plots and how to make them using Excel, and a few other techniques and formulas that I have mentioned.

Stay tuned for the next episode where we'll talk about how to use Excel to analyze data with all these concepts.

Thank you so much for tuning into the episode. I really appreciate your time and energy.

If you have a minute, please take a minute to visit our iTunes page. You can visit http://www.chandoo.org/itunes or just open up the iTunes app on your phone, desktop or Mac and search for chandoo.org in the iTunes store and please leave us a rating and review.

Again, I really appreciate your honest reviews and ratings. Thank you so much for that. You have a fantastic day. Bye.