Transcript for Session 030

Listen to the podcast session, see resources & links:

http://chandoo.org/session30/

Transcript:

Hey folks, welcome to http://chandoo.org/ podcast. This is session 30. Thank you so much for joining me in this episode of http://chandoo.org/ podcast. Our podcast is designed to help you become awesome in data analysis, charting, dashboards, charting and VBA using Microsoft Excel.

Today's topic is quite interesting. In a world where we are more and more getting disconnected - I know it sounds very strange because we are living in a connected world - but, I mean disconnected in the sense that, suddenly, we are working with people all over the world. So, you may be living in USA, Canada or Australia and you are listening to me (Chandoo) living in India and talking in this podcast. Likewise, you might have colleagues spread across the world; you may have partners, vendors, clients and customers all over the world and everybody is disconnected from each other in the sense that they are no longer living in the same place to work.

Let's say that you are a Product Manager; the product that you are designing is not going to be manufactured on the same office floor where you live. It could be manufactured all the way in Kansas or China or India or Bangladesh or whichever country your manufacturing location is in. After the manufacturing is done, it might not be quality checked there; it could be somewhere else in the shipping facility where the quality is checked. And, finally, the sales might not even happen where you live. Your customers could be living in downtown Manhattan. So, everybody is disconnected from each other. The only way that we know that there is somebody else out there is through a communication medium like the internet or email or those kinds of things.

If you are wondering why I am getting philosophical, there is a reason for it. In such a disconnected world, one pertinent problem is fraud. Here is a recent email that I got from one of our readers; my reader received an email with an Excel worksheet that contained some data. He wouldn't tell me what exactly the data was but it was basically like a list of numbers or observations. He asked me, "Am I being lied to?" He was wondering whether the data that he was receiving on a frequent basis in this workbook was fudged, i.e. was there fraud in the data - was the data really from original observations or was somebody just cooking up these numbers in Excel? He was wondering about that and he asked me some mechanisms to detect fraud in data.

Obviously, detecting fraud or forecasting the future is not something that we can do with any amount of certainty. But, especially with data that is sourced from multiple places, there are certain mechanisms using which we can detect fraud. This is what we are going to talk about in this podcast episode. It is an overview of certain **techniques that will help us detect fraud in the data** and detect any manual interventions in the data. I'll talk about this a little later.

First, let me share a couple of quick announcements. As you may know, I am running a **new course** on http://chandoo.org/ called **50 Ways to Analyze Data** and this course is closing enrolments on Friday, 20th February. So, if you are interested in joining this program then Friday, 20th February is the last date for enrolling. I am going to re-open this again in July 2015 so in case you are listening to this podcast way in the future (after 20th February) then you can always come back to http://chandoo.org/ sometime in July and sign up for this program. It is a program designed to help you learn how to analyze data, how to answer questions like, "can you find out what is going on with this data; can you analyse this?" - I.e. the kinds of questions that our managers or clients or CEO's would like to throw at us and we would like to be able to answer them confidently through using various procedures and techniques that are available in Excel.

As the name suggests, the course contains 50 different ways of analyzing data. Please visit http://chandoo.org/ to learn more about it. Please go to http://chandoo.org/session30/ which is the link for this podcast where I will provide a link to this course. If it sounds interesting to you, go ahead and check it out.

The second announcement that I have for you is that for some of you who have been listening to the podcasts already know that I have taken on a new hobby since July 2014. I have taken to cycling as a hobby; actually not as a hobby but more as a fitness exercise and I have been cycling quite regularly. I have cycled more than 1700 kms (roughly 100 miles) from July 2014 to mid-February. It is something that I am really enjoying and this Saturday, 21st February, is the first time that I am **taking part in a 200 km BRM**. BRM is a cycling event that happens all over the world. Basically, in a BRM, you go on a cycle; it is not a race. It is just an event in which you are supposed to cover a certain amount of distance in a certain amount of time. So, in this 200 km BRM I would have about 13.5 hours to cover the distance. There is a set route and there are some time checks along the way where I will have to go and get stamped to account that I have shown up at that point. It is a self-supported tour which means that I will have to take care of my bicycle, food, water and everything else and do this as an event. It is quite interesting. They did one of these last month in Vizag where I live. But, I couldn't go for that because I was too busy preparing for the 50 Ways to Analyze Data course. Since the course is completing its launch window on Friday, I am looking forward to this event on Saturday. I will update you about how it goes in the next podcast. I couldn't contain myself; I had to share this news with you.

Let's now go back and talk about the fraud topic. Fraud is a very prevalent problem. We are experiencing more and more fraud and more creative ways of doing things that are not supposed to be done. For example, back in early 2000, we had these huge accounting frauds from Enron and WorldCom etc. They might have faded in our memory since they were in the distant past but even as recently as 2-3 years ago there have been lots of hedge fund frauds. Hedge fund managers have reported consistent returns and then it turned out that it was all fraud; they just made up those numbers. How would you actually look at a bunch of data and suspect that there might be some fraud? Let me warn you that if you are expecting to detect fraud just by solely using Excel or any type of software for that matter, I would say that it is not possible. This is because sometimes both fraud and actual data could look the same. So, there is really no way to detect fraud completely. That said, you could at least raise a red flag and do a lot of several different tests and if a certain set of data is showing red flags consistently in all these tests then that would mean that we need to go and investigate the data a lot more.

So, fraud detection is basically a heuristic technique. We are trying to replace our gut feeling or what we feel about the data with some sort of statistical and analytical procedure where we look at the numbers, do some number crunching on them and say that there is a probability that some fraud is there in this data. When you want to detect fraud in any data then the key thing to keep in mind is that you need to understand the nature of the data. If I just give you a bunch of random numbers and don't tell you what these numbers are and where they came from and I just give you a list of 2000 numbers and I ask you to detect whether these are made up numbers or real observations, you wouldn't probably be able to say one way or another. This is because you have no idea what this data is. **The first step whenever we want to detect fraud is that we must have a lot of clarity about what this data really is**. Only when we know our data better and only when we have certain expectations from the data, we can go and validate those expectations to see if there is any fraudulent behavior observed. This is the key ingredient when we want to detect fraud. **We need to know our data.**

Once we know our data, we can establish or use a certain set of techniques. I want to highlight that I am no expert in fraud detection. So, whatever I am going to talk for the next 20-25 minutes is purely based on what I have learnt about fraud detection over the years and some insights and ideas based on that. This is not the entire story about fraud detection; there are a few techniques that I have not even mentioned and I am not going to mention them in this podcast because I don't know them very well. So, if you are an expert or if you are working in an industry like insurance or financial markets or forensic departments or a place where you obviously take a look at a lot of data and there is always a high chance of fraud (especially in markets like insurance and if you are working in the claims department of an insurance company then chances are that you would be dealing with fraud data on a pretty regular basis; at least once a week or maybe even once a day) then you need to be an expert. Many insurance companies have in-house experts to detect fraud and raise red flags every now and then. If you are one of those kinds of people and you are listening to http://chandoo.org/ podcast, I encourage you to please visit http://chandoo.org/session30/ and drop your insightful comments there so that I can also learn from you. And, if you want to be a guest on the podcast, just drop me a note and I would be glad to interview you so that we can talk about this topic and our readers and listeners can learn more about it.

Moving on, for me, when it comes to fraud detection, at least **5 techniques** come to mind. These are not the only 5 techniques nor are these the techniques that you should be doing from top down. So, it doesn't mean that technique 1 is better than 2 or 3. It is just the order in which I listed them in my notebook when I was researching for this podcast.

The **first fraud detection technique** that I can use with a lot of certainty is called **Benford's Law**. This is very useful when you are analyzing numbers and you want to figure out if these numbers are made up or if they are real observations. Let's say that you are working as a Product Manager and you have manufacturing locations all over the world and one of your vendors or partners who is running a manufacturing plant has sent you the number of faults found per lot and each lot contains about 1000 units and you produce millions of units. Every day, they send you a list of faults per lot in a spreadsheet and this spreadsheet contains 100 rows or 300 rows or something like that. And, at one point, you have a suspicion that they are just giving you some random data and they are not really doing these observations themselves. So, you want to check and make sure that these numbers are not fraudulent. This is where Benford's Law can be useful. It says that the first digit of numbers - the numbers have to be in variation of magnitude, so if all the numbers are less than 10 then you are not going to find Benford's Law to be true in such a case but if the numbers are spread across a wide range like 0 to 1000 then Benford's Law says that the **distribution of the most significant digits** would be in such a way that number 1 occurs so many more times than number 2 and number 2 occurs so many more times than number 3 and so on and so forth (the most significant digit, i.e. if the number is 100 then 1 is the most significant digit and if the number is 75 then 7 is the most significant digit). So, if you plot the distribution of the first digit of the numbers then it would look like a **curve that goes from top to bottom in a curved shape**. The law states that about 30% of the time we would find 1 as the significant digit and so on and so forth. I will leave a link to the Benford's Law page on Wikipedia in the show notes so that you can go and study that.

Calculating the most significant digit in Excel is very straightforward. You could use the LEFT function, so:
=LEFT(number,1)
will give you the left-most character in the number which could be 1 or 7 or 6 or whatever else. Once you calculate all of these, you just **calculate the frequency** either using the histogram tool from your data analysis add-in in the ribbon or you can just manually calculate it by using the COUNTIFS function. Either way, we will be able to figure out how many times number 1 has occurred and how many times number 2 has occurred and so on and so forth. Then, we can plot that and if there are anomalies in it - i.e., if it so happens that every number has occurred equal number of times, for example 1 is there 10% of the time, 2 is there 10% of the time etc. - then, it means that we are looking at data that might have been fudged.

Remember, Benford's Law **does not apply to all sorts of numbers** but it is **typically observed in most naturally occurring phenomenon**. Examples of naturally occurring would be defects in manufacturing, number of bacteria observed in a petri-dish over a period of time and things like that. So, if you see some sort of aberration in the behavior compared to what is normal which is that number 1 should occur more times than number 2 then we can raise a red flag. It doesn't mean that there is fraud but it at least means that you have to go, roll up your sleeves and do a little more investigation on the data. That's Benford's Law for you.

The **second technique** is called **auto-correlation**. Typically, when you have a set of data, for example, you are looking at returns published by a certain mutual fund or hedge fund - the mutual fund house would publish an annual report that shows the per day change in the net asset value of the mutual fund or what the return on the mutual fund it - and you are wondering whether these numbers are realistic or if somebody just sat and made up these numbers. One way to figure this out would be to check whether there is any auto-correlation. You might already know what correlation is. Correlation is when you are saying that one set of data is related to another set of data. Let's say that you are analyzing share prices of Microsoft and Intel. This is an example from a book that I recently read. We would say that if the Microsoft share prices are going up which is probably not the case these days but go along with me on the example. So, if Microsoft share prices are going up this year - it doesn't always mean this because the stock market is really crazy and idiotic and so we can’t say with certainty that if Microsoft shares are going up then Microsoft itself as a company is doing well as sometimes, it could be just pure, irrational euphoria in the market that could be kicking up the prices - but, let's imagine that Microsoft shares are going up in a normal, rational environment then we could simply say that Microsoft as a company is doing well. What would happen if Microsoft is doing well? It would mean that they're able to sell more of their software products like Windows and Microsoft Office and stuff like that. If they are selling more of these products then it naturally means that there are more computers being bought which means that since computers use Intel chips, as a consequence, Intel should also be doing well. So, this is the kind of situation where we could examine some sort of correlated behavior. Not every change in Intel would be explained by change in Microsoft prices because Intel has many other things apart from chips and likewise, Microsoft also has many other things like X-Box, Windows Phone and many other things that don't rely on Intel chips to succeed. But, with a fair amount of confidence, you would say that there is a correlation. This is a typical correlation.

Auto-correlation means when data depends on itself. For example, if you are looking at a bunch of mutual fund data and you are wondering if there is any auto-correlation, it really means whether this data is really dependent on itself (let's assume we are looking at 12 months data as an example). This means that somebody might have taken the first 2-3 months of data and then used the same pattern to repeat the behavior for the rest of the months. So, they haven't really used actual data; they have just used some sort of observed pattern in the data and replicated that over time. This means that the data is dependent on itself. The future values or some of the values are derived from the other values. This is what auto-correlation means in loose terms. If there is an auto-correlation in your data then it could mean that somebody might have fudged the numbers or something spooky is going on there. Again, there could be some genuine cases where auto-correlation might be observed but if there are red flags on all these kinds of tests then you can be sure that some kind of fraudulent activity might be going and it is worth examining. Auto-correlation is one way to examine that.

How would you do auto-correlation in Excel? I've tried to research this aspect and there are many tests. Some of the names of the tests are so long and confusing that I can't even pronounce them very clearly. The easiest way to me seems if you have a data set of 1000 numbers and you want to check for auto-correlation for order of 1 - order of 1 means whether these 1000 values would depend on themselves with a distance of 1, i.e. 1 to 999 values are dependent on 2 to 1000 values - so, if you take the values and shift them one row down and you check whether these two sets have any correlation, all you have to do is take the data and create a new set of data in the adjacent column that just refers to the previous row of the original data. Now, you have two sets of data and you run the CORREL function on that and Excel will tell if there is any correlation observed. If the correlation is very high, i.e. if it is between 0.7 and 1 then you would have observed an auto-correlation of order 1 in this data. But, the challenge is that we don't really know if this is of order 1 or order 2 or order 3 or order 4 or order 5 and so we have to do these kinds of tests with multiple different orders and sometimes, it can be a very time-consuming and expensive procedure, i.e. expensive in the sense that the amount of time it takes the computer to do all these possible combinations and figure things out. This is where some statistical tests come in. Again, as I said, I am not an expert on the actual process of finding auto-correlation but I will leave a link to the auto-correlation techniques in the show notes and you can go ahead and study them. I will also provide you some links to an excellent website where I found many of these tests done on Excel data sets. You can download the Excel workbooks from that site and study them.

So far, we've talked about 2 techniques, one is Benford's Law and the other is auto-correlation. Let's go ahead and talk about the **third technique** which is **discontinuity at expected value** which is usually also called **discontinuity at zero**. Let's say that you are looking at some data about the number of faults per million or the number of times an employee has attended work or maybe you run the welfare program in a large government where you give free meals to school going kids. And, since you cannot personally go and visit all the school and verify whether the meals are served or not, you have asked all the 1000 participating schools to send you data on a daily basis, i.e. how many students have taken part in the program and had the free meal served to them. Now, schools are incentivised in such a way that if they serve free meals to more than 34 students per day then you will give them some extra funding otherwise they won't receive any funding because, by definition, schools already get some funding for providing free meals. So, if they are serving free meals beyond the threshold then you will fund the school with some money. So, in this situation, economically speaking, schools are incentivised to feed more than 34 people. Schools will send you data specifying how many people they've fed on a daily basis and you're looking at these numbers for a particular school for a given three-month period; let's say that 90 values are there and you're wondering if there is any fraud going on and whether the school is mis-reporting these numbers. Now, since the incentives are in such a way that the school will benefit if they feed more than 34 people there is a chance that if somebody is messing with this data or doing some fraud activity on the data that they will try to have more than 34 values in the data. But, anybody doing fraud on the data would be sensitive (unless they are really dumb in which they might list 35 for all the rows) to smartly and randomly juggle the numbers to make sure that there are more numbers greater than 34.

Then, what you would do at a discontinuity at zero or discontinuity at an expected value analysis is that you take the expected value which is 34 (you are expecting the schools to report values at 34; if you are analysing hedge fund or mutual fund data then you would expect the returns to be greater than 0 as only then the mutual fund is incentivised and more people will buy the fund and they will get money) and count the frequency of the values that are two buckets or bins further to it and one bin before it. In plain English, it mean that we take all the possible values for the number of students, let's say that 34 is the expected value and we bunch the students into a bucket of 4 each so that 0-4 is one bucket, 4-8 is another bucket and so on and so forth. So, we have buckets of 28-32, 32-36 and 36-40 and we take these three buckets - two of them are to the right hand side of 34 and one of them is to the left hand side of 34, i.e. greater than and less than, and we count the frequency. Then, **ideally speaking, the middle value should be around the average of the other two buckets.** Again, in certain types of data, this may not be true. But, in certain other types of data, especially where there is an incentive involved, if there is some fudging happening, you would observe that the middle value will not be at the average of the other two buckets. The middle value will be abnormal; it will be dis-proportionate. This is what we mean by discontinuity. When you plot them as a chart, there won't be a smooth transition between the buckets. There will be an abrupt change at 34. This is because when somebody is manually messing the data it will show up as this behavior on a frequency calculation and chart. It will look like discontinuous behavior at that point and hence the name. You can calculate all of this using your typical COUNTIFS and SUMIFS formulas and examine the changes by plotting them with an ideal scenario. You can do this for all the 1000 schools and any schools that are exhibiting a discontinuous pattern are the candidates for further investigation. For those schools, you could send a surprise team to verify whether they are serving lunch to those many students or maybe install some sort of mechanism to ensure that they are not committing fraud. This is about discontinuity at expected value.

The **fourth technique** is **analysis of distribution**. This is quite useful especially if you know that the numbers are supposed to follow a certain type of distribution. When it comes to distribution, there are many distributions. I am not going to talk about all the distribution techniques that are there because we will be talking until Christmas and I don't even know that much about statistical distributions! But, let me explain, at a high level, some of the common distributions that are often observed in nature. When it comes to distributions, we also need to know one thing which is what type of variables can be there. A variable is nothing but the value that you are observing. For example, if you take a dice and throw it, the value that you would see on the face of the dice would be one of the six values. It could be 1, 2, 3, 4, 5 or 6. You won't see a value like 3.75 or 4.12. You will only see discrete values. Such variables are called **discrete random variables** because the values are defined by a finite set from A to B and X to Y and you can randomly get any value. So, these are discrete random variables. Certain other variables follow a **continuous pattern**, for example, the height of people in a typical city. You won't find people at

5 feet, 6 feet and 7 feet only; you would find people of all types of heights. A person could be 5 feet 7 inches and another person could be 6 feet 3 inches. It is a continuous variable; values vary from one point to another very smoothly and all possibilities can occur within this range. There are infinite heights possible although, technically, it won't be true because you can't really measure with such accuracy but you get the point. So, at a very high level, we have discrete and continuous. Discrete are 1, 2, 3 kind of values and continuous mean all possible values between 1 and 2 or 1 and 5 etc.

Now that you understand the basic random variables or basic variable nature, let's talk a little about the **distributions** that can occur. Of course, when we talk about distributions, we also need to have good clarity about what kind of a variable it is. If it is a discrete variable, it can have certain types of distributions and if it is a continuous variable, it will have some other types of distributions. Let's make all of this more understandable in plain English. Let's say that you are looking at the number of customers arriving in a bank or a railway station. The number of people coming into that particular establishment will be discrete. You can't have 7.23 people showing up in an hour. It will be either 7 or 8 people. That's a discrete variable. But, how many people will show up in any given hour? The distribution of the number of people showing up in an hour depends a lot on the kind of environment we are studying. But, typically, the number of people showing up in a bank would follow a certain distribution pattern. For example, I might want to analyse how many people will be there in a bank at any given hour. Why are we talking about all of these distributions? This is because once we know the expected behavior of this data, we can then go ahead and see if the data set that we have on hand exhibits similar behavior or not. If not, that means that there could be some fraud. This is why we are trying to understand the distribution. There are many distributions out there. Again, since we are beyond the 30 minute mark, I don't want to talk a lot about distributions but just know the common distributions like **normal distribution** which is the bell-curve pattern that we see in many situations and then there are distributions like **binomial**, **poisson**, **exponential**, **weibull** etc. I will provide some links to the basic articles and the basic statistics behind these distributions in the show notes. You can go ahead and study them. Some of them can be highly intimidating especially if you go to the Wikipedia pages of some of these distributions. You don't understand anything when you read them because of all the Greek and Latin symbols that you see there. But, give it a try and I am sure you will appreciate how the distributions work and once you know them, you can go ahead and see if the sample data that you have on hand exhibits that kind of distribution or not, and if not, then that calls for some investigation.

The **last technique** for fraud detection, according to me (again, as I said earlier, these are not the only techniques and I don't know everything about fraud detection), is that you can use some sort of a **learning system and decision trees**. Let me tell you a story from my past. Back when I was doing my computer science engineering, we had an undergraduate project and we had to take a large data set of clinical data (data for people about blood pressure and lots of other things) and we had to predict whether they were going to get diabetes or not. It was some sort of predictive analysis. One way to do this was to take one data set for which we knew for sure that the person has diabetes or not; we took roughly 700 such records. For each of them, we knew whether they had diabetes or not and we used that data set to train a system. That data set contained about 25 different columns - height, family history, hereditary aspects and lots of different things - and we used all of that data to train our system. So, our system was learning from that data. It created what we call a decision tree. For example, if age is less than 25, do this etc. Once the decision tree is constructed and it is trained then we take the testing data set, i.e. the data set where we don't know whether somebody has diabetes or not and then we run that data set through this tree to predict whether they are going to get diabetes or not with a certain percent of confidence. We can't really do a 100% prediction so we'll say that given this person's attributes, according to this decision tree, we predict that there is a 95% chance that they will get diabetes.

You can use the same concepts and take non-fraudulent data that you know hasn't been fudged and train your decision trees and then take the other data and run it through the decision tree so that it will raise a flag for suspicion or not. Now, when it comes to how to implement this decision tree mechanism in Excel, unfortunately, there is no easy way. You could write a VBA program, implementing some of the common decision tree strategies in VBA, and then train using the training set, and run it using a testing test. Or, you could just understand the process behind it and then go ahead and implement it in some sort of a specialty system like any statistical processing systems like R, SPSS or even a programming language like DotNet or Java or PHP or even in a distributed environment like Hadoop or one of those kinds of things. That's how you can detect fraud.

I want to conclude with some additional notes, one of them being a **reminder about the success-failure rate of fraud detection**. At best, what you can do with fraud detection is that you can raise some red flags and do further investigation. Sometimes, we might hit a wrong identification and we might wrongly say that this data could be fraudulent whereas the data is really like that. It is similar to lie-detector tests or predicting the future. We can't really say for sure what's going to happen; all we can do is that we can do some analysis and throw out some numbers from it. So, fraud detection is a science where the best work that you are going to do is something marginally better than your gut feel or your guess.

That said, it is a very useful technique. Principles like Benford's Law, auto-correlation, discontinuity at zero are very easy to implement in Excel to test data and raise red flags early on. If you wait for a long time and only then detect that the supplier has been duping you all along then it is no good. So, you can use this kind of data from time to time to check and make sure that human intervention and fraud is not happening on your data.

Thank you so much for listening to this podcast. I hope you enjoyed it. I know it has been a lot of techniques and statistical words but these things are quite interesting and, from time to time, I find that learning about these things expands our horizons and we end up discovering or at least identifying new ways to analyze data. So, I hope you enjoyed this podcast. Go ahead and download the podcast episode and show notes from http://chandoo.org/session30/ and if you have any comments or suggestions or if you want to share your own experiences for fraud detection, please visit our show notes page at http://chandoo.org/session30/ and leave your comments there. Thank you so much. I'll see you again in the next episode. Bye.