DSC520 Week 1 assignment – Exercise 1.2

Team – DSC520-T301 Statistics for Data Science (2215-1)
Created by – Manish Shekhar

Task 1 – R installed

Task 2 – Installed R studio desktop

Task 3 – Created GitHub repository and completed "Hello World" tutorial.

Link to my Git Hub repo - https://github.com/datatodecision/stats_for_data_science
Git Hub repo is integrated with R studio.

Task 4 – What is the level of measurement of following variables?

1. The number of downloads of different bands' songs on iTunes.

This variable is a measure of type **discrete** because it can contain only whole numbers and it would be possible to define categorical bins. For example, number of downloads between 0 and 10, 10 and 20 etc. It is also a **ratio** because it can have 0 as a true value which would mean no downloads. We can also relate it to another value in real life i.e., number of downloads by person A is 2 times the number of downloads by person B.

2. The names of bands that were downloaded.

The variable is a dimension of type **nominal** (categorical) because it would contain definitive list of string values. Also, there is no significance of order between two values as they are names i.e., it would be rare that one song is downloaded always after another, which would then be mere coincidence.

3. The position in the iTunes download chart.

The variable is a dimension of type **ordinal** (categorical) because it talks about the position of the bands' song in the download chart. It does not tell exactly how many songs were downloaded to be at that position but talks about which songs are ranked at what spot from amount of downloads perspective.

4. The money earned by the bands from the downloads.

The variable is a measure of type **continuous** (numerical) because it can literally take any value not necessarily to be the whole number. It is also of type **ratio** because we can show proportional income between two bands i.e., 1 band may make 3 times the other during a time period.

5. The weight of the drugs bought by band with their royalties.

The variable is a measure of type **continuous** (numerical) because it can literally take any value not necessarily to be the whole number. It is also of type **ratio** because we can show proportional weight between two buys i.e., 1 artist may buy 2 times the weight of the drug than the other.

6. The type of drugs bought by the bands with their royalties.

The variable is a dimension of type **nominal** (categorical). This is probably just the names won't have any orderly significance.

7. The phone numbers that bands obtained because of their fame.

The variable is a dimension of type **nominal** (categorical) as well. These are distinct values mostly in particular format. There is no significance in summarizing the phone numbers and also, they do not have any orderly significance.

8. The gender of the people giving the band their phone numbers.

The variable is a dimension of type **binary** (categorical). In most scenarios they will have one of two values i.e., Male or Female.

9. The instruments played by the band members.

The variable is a dimension of type **nominal** (categorical). These are just the type of instruments or instrument names i.e., guitar, flute, drum, piano, violin etc with no orderly significance.

10. The time, they had played learning to play that instrument.

This one is very tricky. Time is ofcourse **continuous** as it can be represented as various fragments i.e., year, month, week, day, hour, minute, second, microsecond etc. It does not have to be whole numbers. Also, it is **ratio** because it may happen that a band does not play a particular instrument and thus time spent to learn that instrument for them will be zero. Also, it can be proportional to others i.e., time taken by band A to learn guitar could be 5 times more than that of band B.

Task 4

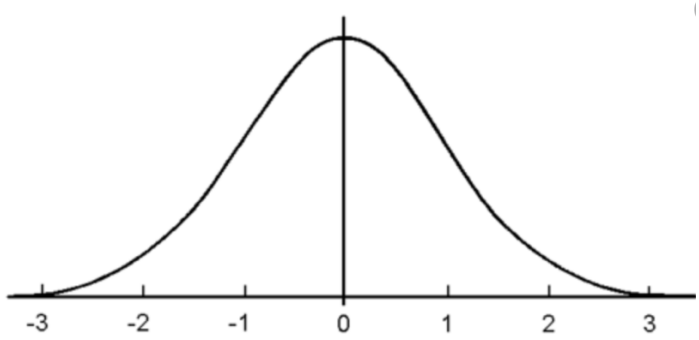
Say I own 857 CDs. My friend has written a computer program that uses a webcam to scan the shelves in my house where I keep my CDs and measure how many I have. His program says that I have 863 CDs. Define measurement error. What's the measurement error in my friend's CD counting device?

Measurement error is defined as deviation between original value and the calculated or estimated value. We generally use measurement error as a parameter to define the model efficiency. It's used a lot after building and training a model during model validation using test data. Minimal the measurement error, better the model would estimate (this comes with the trade-off of overfitting which is not needed to be discussed here). My firend's CD counting device has measurement error of 6 as it counts +6 than actual.

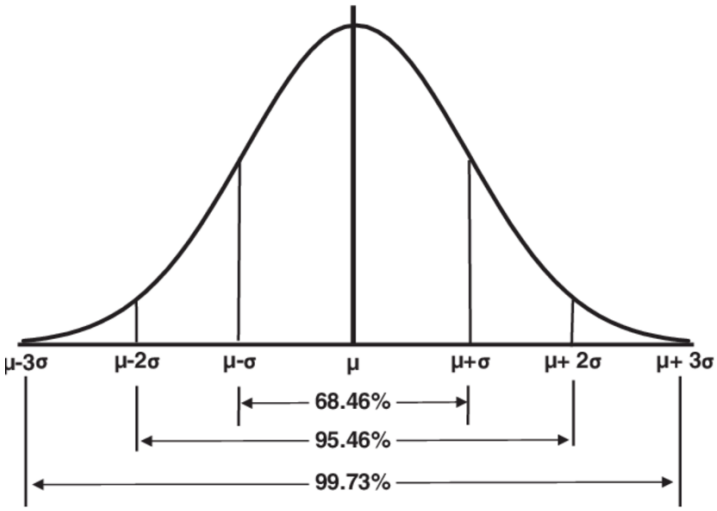
Task 5

Sketch the shape of normal distribution, a positively skewed distribution and a negatively skewed distribution.

It's phenomenal to note that for any data sample if we calculate and plot probability distribution it is always normal distribution curve (this can be achieved by calculating z-scores). In a perfect normally distributed curve, mean, median and mode are equal. Normal distribution is symmetric from the peak (mean/median/mode) which means most observations or frequency of data is clustered at the peak and data becomes less frequent when is farther away from the peak. Two main parameters of the normal distribution are mean and standard deviation. Mean represents central tendency while standard deviation represents standard deviation of each data point from mean.



Standard normal distribution is the special case of normal distribution where mean = 0 and standard deviation = 1.



Picture above shows relationship between mean and standard deviation. In any normal distribution about 68.46% of observations exist between (mean – standard deviation) and (mean + standard deviation), about 95.46% of observations exist between (mean – 2*standard deviation) and (mean + 2*standard deviation), and about 99.73% of observations exist between (mean – 3*standard deviation) and (mean + 3*standard deviation).

Positive and negative skew – when data lacks symmetry. In these cases, most frequent data points are clustered at one end of the scale. They are called positively skewed when most frequent data points are towards lower end of the scale and tail points towards the higher end. They are called negatively skewed when most frequent data points are towards higher end of the scale and tail pointing towards the lower end.