STATISTICS

DAY-1:

2 parts of statistics:

Descriptive - describe the data. what patterns, what analysis,

what details, what observations or hidden in the data.

that is descriptive analysis.(something which we can precisely)

- "Descriptive statistics are like a summary of data that helps us understand its main features. It's like taking a quick look at a picture to get an idea of what it shows, without needing to examine every tiny detail. Descriptive statistics give us a snapshot of the data, showing things like the average, the spread, and the most common values.(mean, median, mode)"

Descriptive Statistics: Descriptive statistics describe and summarize data. They help us understand what the data looks like by giving us information like the average, spread, and most common values. Descriptive statistics are like a summary or snapshot of the data.

Inferential - a small sample of data and with that we make an observation

about a big population of data.

ex. find average income of Indian families.

- "Inferential statistics help us make predictions or draw conclusions about a larger group based on a smaller sample of data. It's like using a small taste of soup to guess how the whole pot tastes. We analyze the sample to make educated guesses about the entire population."

Descriptive Statistics

TYPES OF DATA:

1) Numerical data :

Numerical data can be broadly classified into two types:

1. **Discrete Data**: This type of data consists of whole numbers or counts that are distinct and separate. Discrete data cannot take on every possible value within a range. For example, the number of students in a class (can't have 5.5 students) or the number of cars in a parking lot are discrete data.
2. **Continuous Data**: Continuous data can take on any value within a range and can be measured at any level of precision. It includes decimal numbers and is often obtained by measuring. Examples include height, weight, or temperature.

2) Categorical Data:

Categorical data refers to data that represents categories or groups. These categories are often qualitative in nature and do not have a numerical value. Examples of categorical data include:

1. **Nominal Data**: This type of data represents categories with no intrinsic order or ranking. For example, eye color (e.g., blue, brown, green) or types of fruits (e.g., apple, banana, orange).
2. **Ordinal Data**: This type of data has a specific order or ranking. However, the differences between the categories are not necessarily uniform or measurable. Examples include educational levels (e.g., high school, college, graduate school) or ratings (e.g., low, medium, high).

In statistical analysis, categorical data is often used to group data points into specific categories to understand patterns, relationships, or trends in the data.

3) Boolean – True / False.

4) Date time data type

Note:

Outliers:

An outlier is a data point that is significantly different from other data points in a dataset. It's like the "odd one out" that doesn't seem to fit with the rest of the data.

For example, consider a group of friends who earn $30,000, $35,000, $32,000, $33,000, and $1,000,000 per year. The friend who earns $1,000,000 would be considered an outlier because their income is much higher than the incomes of the other friends.

MEASURE OF CENTRAL TENDENCY:

A measure of central tendency is a single value that represents the center or middle of a data set.

The main measures of central tendency are the mean, median, and mode.

Note: Mean is much affect by the outliers, so when mean fails to give the correct results, then that time we use Median, it's not much affect by the outliers. (robust- strong, healthy)

Mean:

The mean is a measure of central tendency that is calculated by adding up all the values in a dataset and then dividing by the number of values. It is often referred to as the average.

Here's an example to illustrate how to calculate the mean:

Let's say we have the following dataset representing the scores of five students on a test: 85, 90, 92, 88, 95.

To find the mean, we add up all the scores: 85 + 90 + 92 + 88 + 95 = 450.

Next, we divide the sum by the number of scores (which is 5 in this case):

450 / 5 = 90.

So, the mean score for the test is 90.

Median:

(First we sorting the data and then find middle number.)

The median is the middle value in a list of numbers when they are ordered from smallest to largest. If there is an even number of values, the median is the average of the two middle numbers.

For example, let's say we have the following list of numbers representing the ages of a group of people: 22, 25, 30, 35, 40.

To find the median, we first order the numbers from smallest to largest: 22, 25, 30, 35, 40.

Since there is an odd number of values, the median is the middle number, which is 30.

If we had an even number of values, such as 22, 25, 30, 35, 40, 45, then the median would be the average of the two middle numbers, which are 30 and 35. So, the median would be (30 + 35) / 2 = 32.5.

Note: You can use Median in the case of both balanced and imbalanced since it is more robust.

Mode:

The mode is the value that appears most frequently in a dataset. It is possible to have more than one mode if two or more values occur with the same highest frequency.

For example, consider the following set of numbers representing the scores of students in a class: 85, 90, 92, 88, 90, 95, 90.

In this dataset, the number 90 appears three times, which is more frequently than any other number. Therefore, the mode of this dataset is 90.

Certainly! Here's an example of finding the mode with categorical data:

Let's say we have a dataset representing the colors of cars in a parking lot:

- Red
- Blue
- Red

- Green
- Blue
- Red
- Yellow

In this dataset, the color "Red" appears most frequently (3 times), making it the mode of the dataset.

MEASURE OF VARIATION / DISPERSION:

Variance measures how spread out or dispersed the values in a dataset are. It tells you the average squared difference between each data point and the mean of the data. A high variance means the data points are spread out over a wider range, while a low variance means the data points are closer to the mean.

In simpler terms, variance gives you an idea of how much the numbers in a dataset differ from the average.

1. Range, 2. Variance, 3. Standard Deviation, 4.Interquartile Range

1. Range:

Range is a measure of dispersion in a dataset, indicating the difference between the largest and smallest values. It gives you an idea of how spread out the data is.

For example, consider the following set of numbers representing the heights of students in a class: 150 cm, 160 cm, 170 cm, 175 cm, 180 cm.

To find the range, we subtract the smallest value (150 cm) from the largest value (180 cm):

Range = 180 cm - 150 cm = 30 cm.

So, the range of heights in this class is 30 cm, indicating that the heights vary by 30 cm from the shortest to the tallest student.

2. **Variance:**

Variance = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Where:

- n is the number of data points in the dataset.
- x_i represents each individual data point.
- $\bar{x}$ is the mean (average) of the dataset.

Let's calculate the variance of a simple dataset to illustrate the formula:

Dataset: 2, 4, 6, 8, 10

Calculate the mean ($\bar{x}$):

$\bar{x} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$

Subtract the mean from each data point, square the result, and sum these squared differences:

$(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2$

$= (-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2$

$=16+4+0+4+16=40$

Find the average of these squared differences to get the variance:

Variance = $\frac{40}{5} = 8$ So, the variance of the dataset is 8.

(This suggests that the variance is low, meaning that the values are clustered relatively closely around the mean.)

3. **Standard Deviation:**

S.D. = $\sqrt{variance}$

$= \sqrt{8}$

$= 2.82842$

(If we can't ignore the outliers, we can't use the range, variance, and standard deviation, at that time we use the interquartile range.)

4. **Interquartile Range:**

(https://www.scribbr.com/statistics/interquartile-range/)

(Note: 4 part - quartile, 5 part – quintile, 6 part – decile & median is used for IQR.)

- Sort the data.
- Divide the data into 4 parts.
- IQR = $Q_3 - Q_2$
- Lower = $Q_1 - 1.5 * IQR$ (Remove the outlier)
- Upper = $Q_3 - 1.5 * IQR$ (Remove the outlier)

(Not all are outliers in 1st and 4th tables. It's have some good data too.)

(IQR is also used for outlier detection and removal (just ignore not delete the outliers.)

CODING:

data.describe() → Statistical Description of the data. (data = file name)

It is show: count, mean, std., min, 25%, 50%, 75%, max. of each column.

****

DAY-2:

(**See colab notebook – Statistics Examples .ipynb**)

CORRELATIONS:

Correlation: Direction and Strength of the relationship between two numerical columns. (Correlation is a measure of how two things are related or connected. This is only for Numerical data, not categorical data.)

Note: We use scatter plot for plotting the variables (column).

Covariance: Direction of the relationship between two numerical variables.

*(**correlation** tells us about both the direction and strength of the linear relationship between two variables, while **covariance** only tells us about the direction of the relationship.)*

Difference between correlation and covariance:

Covariance is a measure that indicates the extent to which two variables change together. If the covariance is positive, it means that as one variable increases, the other variable tends to increase as well. If the covariance is negative, it means that as one variable increases, the other variable tends to decrease. A covariance of zero indicates that there is no linear relationship between the variables.

Correlation, on the other hand, is a standardized measure that indicates the strength and direction of the linear relationship between two variables. Unlike covariance, which can take on any value, **correlation is always between -1 and 1.** A correlation of 1 indicates a perfect positive linear relationship, a correlation of -1 indicates a perfect negative linear relationship, and a correlation of 0 indicates no linear relationship. (we consider 0.7 to 1.0 – strong relationship, 0.5 to 0.7 – average relationship, below 0.5 not consider it. Similarly negative side.)

In simple terms, covariance tells you whether two variables change together, while correlation tells you how strong and in what direction their relationship is.

1. **Positive Correlation:**

Positive correlation means that as one thing increases, the other thing also tends to increase. And also means that as one thing decrease, the other thing also tends to decrease.

Example:

Imagine you're tracking the number of hours spent exercising per week and the number of calories burned. If there's a positive correlation between these two variables, it means that as the number of hours spent exercising increases, the number of calories burned also increases.

So, if you plot this on a graph, you'd likely see a pattern where as the x-axis (hours of exercise) increases, the y-axis (calories burned) also increases, showing a positive relationship between the two variables.

2. **Negative Correlation:**

Negative correlation is when one thing goes up, the other tends to go down.

Example:

In general, as the price of a product goes up, the quantity demanded tends to go down. This is because people are less likely to buy a product when it is more expensive.

3. **No Correlation:**

No correlation means that there is no relationship between two things.

For example, imagine you have a group of people and you measure their shoe sizes and their heights. If there is no correlation between shoe size and height, it means that having a larger shoe size does not mean a person will be taller, and vice versa. The two variables are unrelated in this case.

****

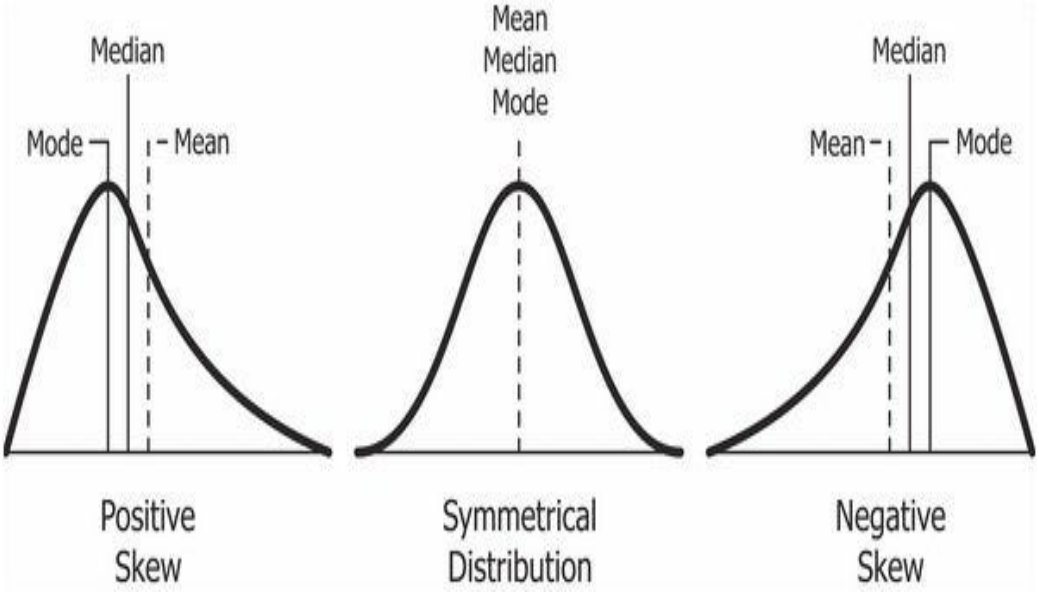
DAY-3:

DISTRIBUTION OF THE DATA:

The distribution of data refers to the way data points are spread out or arranged across different values in a dataset. It describes the pattern or shape of the data and helps us understand how values are clustered together or dispersed.

NORMAL DISTRIBUTION:

Normal Distribution is a term that is describe a distribution which when plotted gives us a shape of bell curve. It has mean of zero and standard deviation 1.



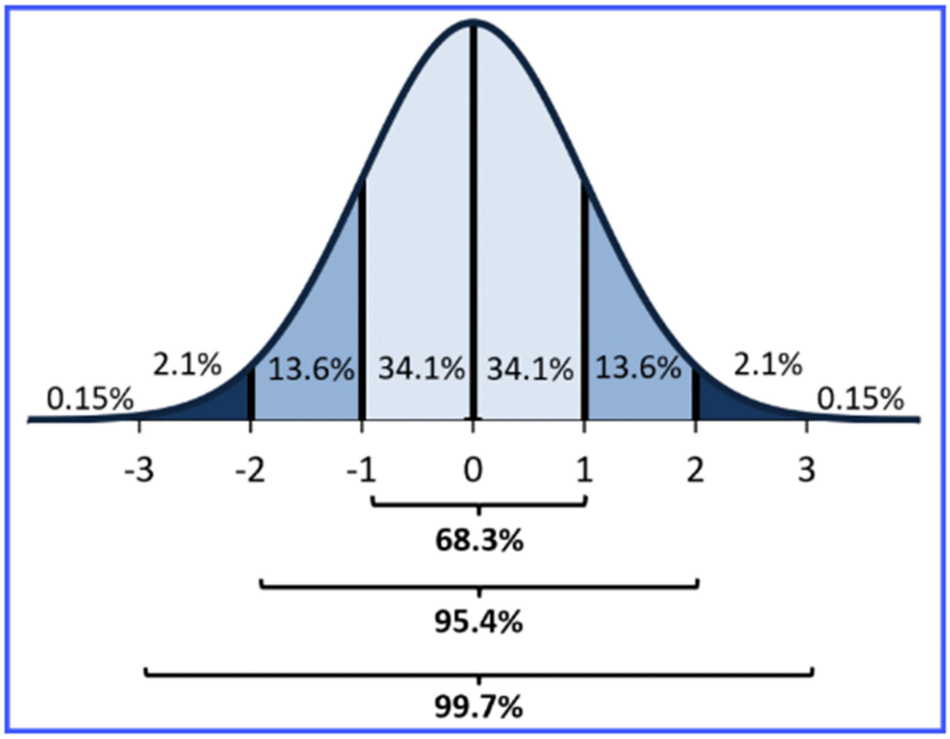
Other type of distributions is Poisson, Binomial, Bernoulli, Uniform distribution, etc.,

- We can find the distribution one by one column. (Ex. Natural things data mostly give normal no skew distribution plot. (Age, height, iq, people in the city,…)
- Use kdeplot, displot or histogram for plot the distribution of the data.
- 2nd one is balance data, we get more accurate result.
- 1st and 3rd are affected by the outliers, so we can't get accurate result. Now this time we use transformation technique to change 1st and 3rd similar to 2nd one.

- Most use transformation techniques:
    1. Log transformation
    2. Square root transformation
    3. BoxCox transformation

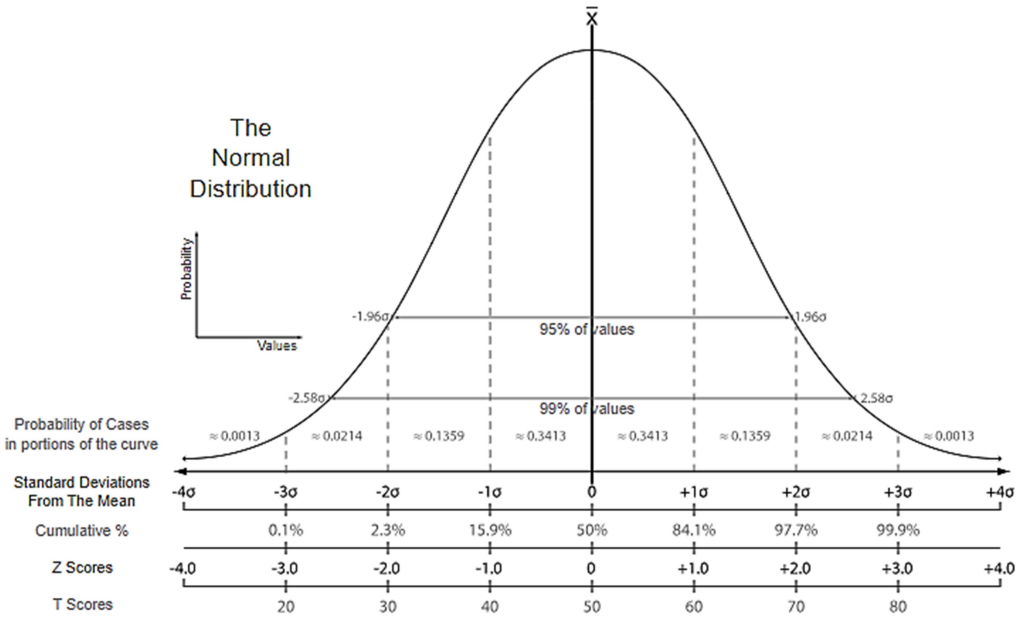
EMPIRICAL RULE:

Empirical Rule, is used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviation. (It is only used for normal plot)

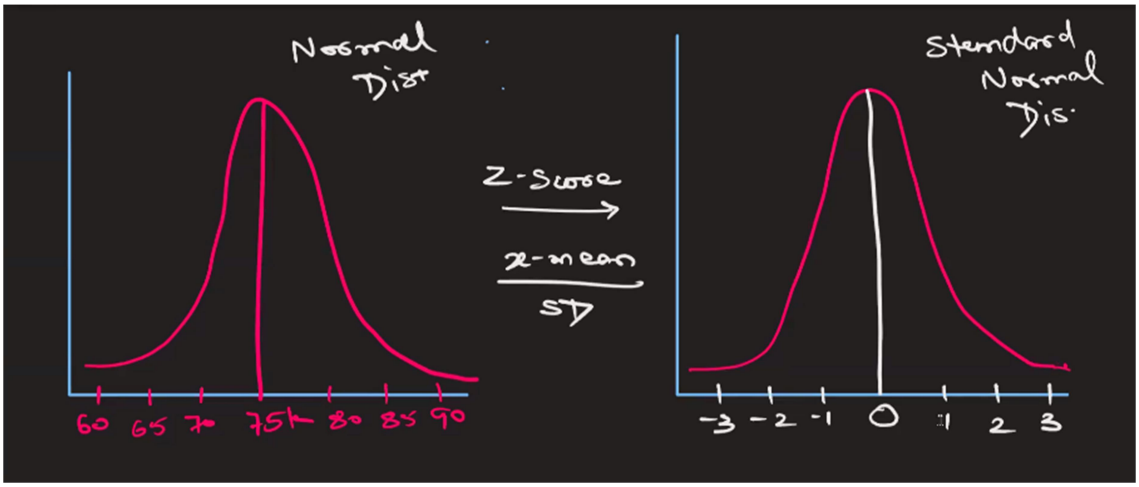


- It tells us about the amount of data that lies within a certain range of values.
- 68% of data = mean – 1*SD to mean + 1*SD
- 95% of data = mean – 2*SD to mean + 2*SD
- 99.7% of data = mean – 3*SD to mean + 3*SD
- Remaining 0.3 we consider, it to be outlier.

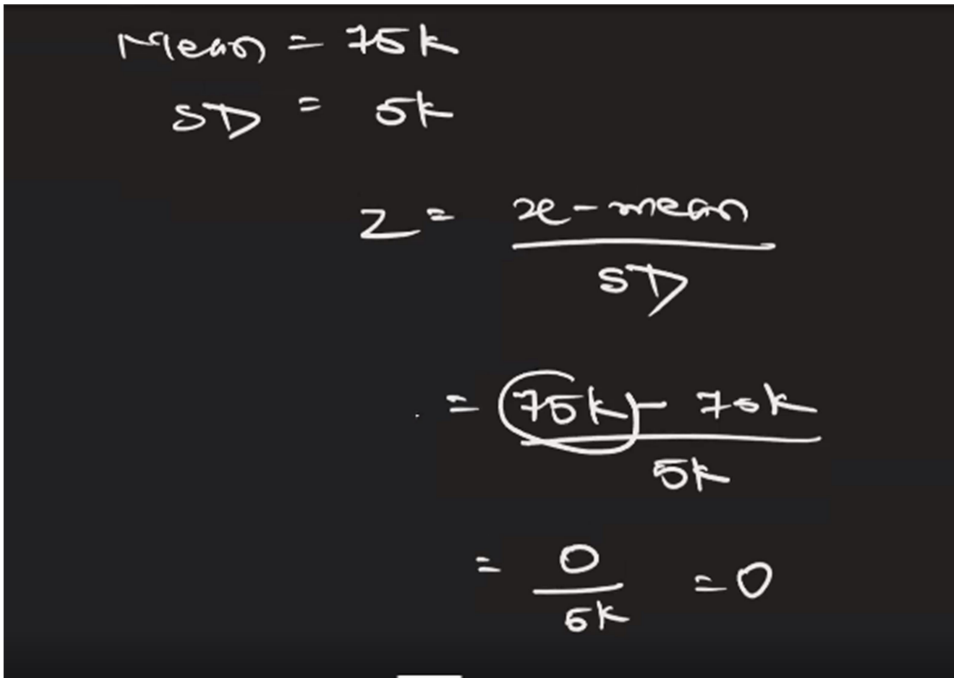
Z - SCORE or Z – TABLE or STANDARD SCORE:



- z – score also called standard score, so that distribution is called <u>standard normal distribution.</u>



- Change normal distribution to standard normal distribution



Similarly find -1, -2, -3, 1, 2, 3…

Find Percentage of earning more than 82K:

$= \frac{82k - 75k}{5k}$

$= 1.4 \rightarrow$ Z-score value

100 - 91.92 %

8.08 %

↓

Percent of employees earning more than 82k.

$= \frac{68k - 75k}{5k}$

$= -1.4$

67k

-1.6

82k

1.4

NOTE: Use Z – score table to convert standard values to percentage. It has +ve and –ve tables.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(See Drive -----> Statistics Examples)

****