WORKSHEET- VI

***Q.1: ans**. – a.*

***Q.2: ans**. – a.*

***Q.3: ans**. – a.*

***Q.4: ans**. – c.*

***Q.5: ans**. – a.*

***Q.6: ans**. – b.*

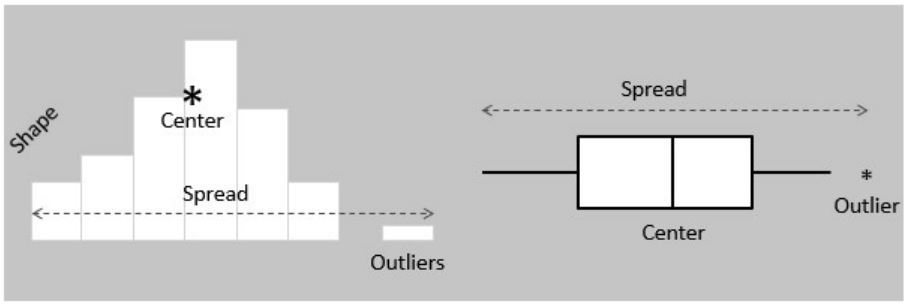
***Q.7: ans**. – c.*

***Q.8: ans**. – b.*

***Q.9: ans**. – b.*

***Q.10: ans**. - Boxplots may also depict values that are far outside of the normal range of responses (referred to as outliers). A histogram is a graphical representation of the spread of data points. Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.*

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

WORKSHEET- VI

***Q.11: ans**. - First of all, metrics which we optimise tweaking a model and performance evaluation metrics in machine learning are not typically the same. Below, we discuss metrics used to optimise Machine Learning models. For performance evaluation, initial business metrics can be used. Based on prerequisites, we need to understand what kind of problems we are trying to solve. Here is a list of some common problems in machine learning:*

- *Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.*
- *Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.*
- *Ranking. The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.*

In our case, we are solving the problem of finding mathematical metrics which will also optimize the initial business problem. Below we list basic metrics to start with. CONFUSION MATRIX

This matrix is used to evaluate the accuracy of a classifier and is presented in the table below.

Some examples
False Positive (FP) moves a trusted email to junk in an anti-spam engine.
False Negative (FN) in medical screening can incorrectly show desease absense, when it is actually positive.

ACCURACY METRIC

WORKSHEET- VI

This metric is the basis one. It indicates the number of correctly classified items compared to the total number of items.
Keep in mind that accuracy metric has some limitations: it doesn't work well with unbalanced classes that can have many items of the same class and few other classes.

RECALL/SENSITIVITY METRIC

Recall Metric shows how many True Positives the model has classified from the total number of positive values.

PRECISION METRIC

This metric represents the number of True Positives which are really positive compared to the total number of positively predicted values.

F1 SCORE

This metric is a combination of precision and recall metrics which serves as a comprise. The best F1 score equals 1, while the worst one is 0.

REGRESSION
performance metrics

MEAN ABSOLUTE ERROR (MAE)

This regression metric indicates the average sum of absolute difference between the actual and predicted value.

MEAN SQUARE ERROR (MSE)

Mean Squared Error (MSE) calculates the average sum of squared difference between the actual and predicted value for the entire data points. All related values are raised to the second power therefore all of negative values are not compensated by positives. Moreover, due to the features of this metric, the

impact of errors is higher. For example, if the error in our initial calculations is 1/2/3, MSE will equal 1/4/9 respectively. The less MSE is, the more accurate our predictions is. MSE = 0 is the optimal point in which our forecast is perfectly accurate.

MSE has some advantages over MAE:
1. MSE highlights large errors over small ones.
2. MSE is differentiable which helps find minimum and maximum values using mathematical methods more effectively.

ROOT MEAN SQUARE ERROR (RMSE)

RMSE is a square root of MSE. It is easy to interpret compared to MSE and it uses smaller absolute values which is helpful for computer calculations.

***Q.12: ans**. - Statistical significance is a measure of reliability in the result of an analysis that allows you to be confident in your decision making. Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.*

Here, a "hypothesis" is an assumption or belief about the relationship between your datasets. The result of a hypothesis test allows us to see whether this assumption holds under scrutiny or not.

A standard hypothesis test relies on two hypotheses.

- *Null hypothesis: The default assumption of a statistical test that you're attempting to disprove (e.g., an increase in cost won't affect the number of purchases).*
- *Alternative hypothesis: An alternate theory that contradicts your null hypothesis (e.g., an increase in cost will reduce the number of purchases). This is the hypothesis you hope to prove.*

The testing part of hypothesis tests allows us to determine which theory, the null or alternative, is better supported by data. There are many hypothesis testing methodologies, and one of the most common ones is the Z-test, which is what we'll discuss here.

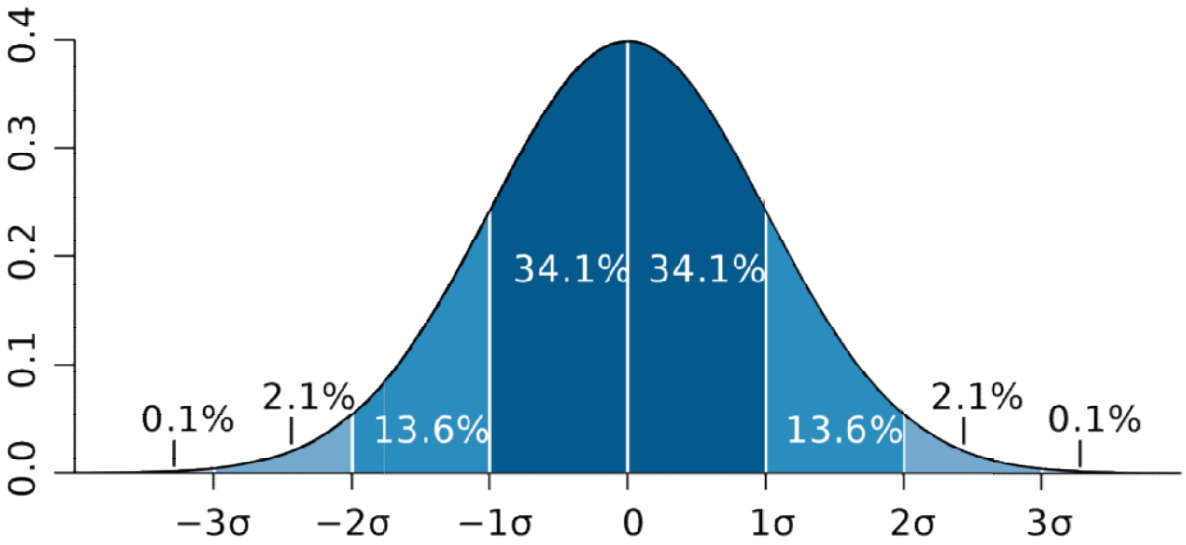
But, before we get to the Z-test, it is important for us to visit some other statistical concepts the Z-test relies on.

Normal distribution

WORKSHEET- VI

Normal distribution is used to represent how data is distributed and is primarily defined by:

- *The mean (μ): The mean represents the location of the center of your data (or the average).*
- *The standard deviation (σ): The standard deviation is a measure of the amount of variation or dispersion of a set of values and represents the spread in your data.*



A NORMAL DISTRIBUTION CURVE (IMAGE SOURCE – WIKIPEDIA)

Normal distribution is graphically depicted by what is called a "bell curve" (due to its shape). A normal distribution curve is used to assess the location of a data point in terms of the standard deviation and mean. This allows you to determine how anomalous a data point is based on how many standard deviations it is from the statistical mean. The properties of a normal distribution mean that:

- *≈68.3% of all data points are within a range of 1 standard deviation on each side of the mean.*
- *≈95.4% of all data points are within a range of 2 standard deviations on each side of the mean.*
- *≈99.7% of all data points are within a range of 3 standard deviations on each side of the mean.*

If we have a normal distribution for a data set, we can locate any data point by the number of standard deviations it is away from the mean.

For example, let's consider that the average number of downloads of a music streaming app is 1000, with a standard deviation of 100 downloads. If an app called MixTunes has 1200 downloads, we can say that it is 2 standard deviations above the mean and is in the top 2.3% of music apps.

Z-score

WORKSHEET- VI

In statistics, the distance between a data point and the mean of the data set is assessed as a Z-score. The Z-score (also known as the standard score) is the number of standard deviations by which a data point is distanced from the mean.

A Z-score is calculated by subtracting the mean of the distribution (μ) from the value of the considered data point (x) and dividing the result by the standard deviation (σ).

$Z = (x – \mu)/(\sigma)$

In the example we discussed above, MixTunes would have a Z-score of 2 since the mean is 1000 downloads and the standard deviation is 100 downloads.

Assuming a normal distribution lets us determine how meaningful the result you observe in an analysis is, the higher the magnitude of the Z-score (either positive or negative), the more unlikely the result is to happen by chance, and the more likely it is to be meaningful. To quantify just how meaningful the result of your analysis is, we use one more concept.

In studies where a sample of an overall population is considered (like surveys and polls), the Z-value formula is slightly changed to account for the fact that each sample can vary from the overall population, and thus have a standard deviation from the overall distribution of all samples.

$Z = (x – \mu) / (\sigma / \sqrt{n})$

Here, √n is the square root of the sample size.

P-values

The final concept we need to use the Z-test is that of P-values. A P-value is the probability of finding results at least as extreme as those measured when the null hypothesis is true.

For example, say we're measuring the average height of individuals in the US states of California and New York. We can start off with a null hypothesis that the average height of individuals in California is not higher than the average height of individuals in New York.

We then perform a study and find the average height of Californians to be higher by 1.4 inches, with a P-value of 0.48. This implies that in a world where the null hypothesis—the average height of Californians is not higher than the average height of New Yorkers—is true, there's a 48% chance we would measure heights at least 1.4 inches higher in California.

So, if heights in California are not actually higher, we'd still measure them as higher by at least 1.4 inches about half the time due to random noise in the data. Subsequently, the lower the P-value, the more meaningful the result because it is less likely to be caused by noise or random chance.

Whether or not the result of a study or analysis can be called statistically significant depends on the "significance level" of that test study, which is established before the study begins. The significance value, denoted by the greek letter alpha (α), is nothing but the maximum P-value we can accept to

consider the study statistically significant. In other words, it's the probability of rejecting the null hypothesis when it's true–or, simply put, the probability of making a wrong decision.

This significance value varies by situation and field of study, but the most commonly used value is 0.05, corresponding to a 5% chance of the results occurring randomly.

A Z-score can be converted to a P-value (and vice versa) using a programming language like R, or by simpler methods like an Excel formula, an online tool, a graphing calculator, or even a simple number table called the Z-score table.

Z-testing

For a Z-test, the normal distribution curve is used as an approximation for the distribution of the test statistic. To carry out a Z-test, find a Z-score for your test or study and convert it to a P-value. If your P-value is lower than the significance level, you can conclude that your observation is statistically significant.

Let's take a look at an example.

Imagine we work in the admissions department at University A, located in City X. In order to show that we're a great university, we want to prove that the students of University A perform better on a standardized test than the average student in City X. The board of the standardized test's testing committee analyzed all the test scores and told us that students from City X average 75 points on the standardized test.

To see if our students actually perform better, we poll 100 students to share their test scores and find out that the average is 78 points with a standard deviation of 2.5 points. We also set a significance level (α) value of 0.05, which means the results are significant only if the P-value is below 0.05..

Since we are trying to prove that our students perform better on the test, our null hypothesis is that the average score of students at University A is not above the city average.

We begin by calculating the Z-score for this test by subtracting the population mean (the City X average of 75) from our measured value (78) and dividing by the standard deviation (2.5) over the square root of the number of samples (100).

$Z = (x – \mu) / (\sigma / \sqrt{n}) = (78-75) / (2.5 / \sqrt{100})$

This gives us a Z-score of 12. Converting this Z-score gives us a very small P-value that's less than 0.00001. That means that we can reject the null hypothesis. In other words, there's statistically significant evidence that students of

WORKSHEET- VI

***Q.13: ans**. - Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.*

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

***Q.14: ans**. - Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.*

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side,

then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

***Q.15: ans**. For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects 60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.*

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Many random variables have distributions that are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

For me, the best examples come from my field of research (astrophysical data analysis). For example, something that comes up all the time is that we detect stars in astronomical images and solve for their celestial coordinates. My current project uses images about 1.5 degrees on a side and typically detects

WORKSHEET- VI

60 to 80 thousand stars per image, with the number well modeled as a Poisson Distribution, assuming that the image is not of a star cluster surrounded by mostly empty space. That's about 8 or 9 stars per square arcminute. If we cut out "postage stamps"from the image that are half an arcminute per side, then the mean number of detected stars in them is about 2. If we do that for (say) 1000 postage stamps and make a histogram of the number of detected stars in them, it will not look very Gaussian, but as we increase the size of the postage stamps, it becomes asymptotically Gaussian.

What generally never becomes Gaussian, however, is the Uniform Distribution. A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.

Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.

The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.