Name : Abhisher Srivastava
Reg. No : 19BCE10071
Subject : NASSOM
Date : Mar 11, 2022
Slot : E21+E22+E23

Q1]

Categorical variables represent types of data that may be divided into groups. Example categorical variables are race, sex, age group and education level which means class level.

Nominal Attributes - related to names.

The name of things or symbols, are the values of nominal attribute. Nominal attribute values reflect a category or state, which is why nominal attributes are sometime known as categorical numbers or attributes, and there is no order (rank, position) among nominal attribute

| Attribute. | Values. |
|---|---|
| Colours. | Red, blue, green |
| Categorical Data. | Lecturer, student, professor. |

## Binary Attributes - has only two states.

- Symmetric : Both values are equally important.
  For Example: Gender.
- Asymmetric : Both values are not equally important.
  For Example : Result.

| Attribute | Values. |
|---|---|
| Gender | Male, Female |
| Cancer Detected | Yes, No |
| Result | Pass, Fail. |

## Ordinal Attributes

It comprises of values that have a logical sequence or ranking between them but the magnitude between them is unknown, the order of values indicate why it is important and what is important.

| Attribute | Value |
|---|---|
| Grade | A, B, C, D, E |
| PayScale | 17, 18, 19 |

## Numeric

Because it is a measurable quantity represented in integer or real values, a numeric quantity is quantitative. There are two sorts of numeric attributes: Interval and ratio.

- The numeric characteristic do not have a correct reference point, or what we can term zero point, while interval scaled attributes contains value with interpretable differences. In interval scale, data can be added or subtracted but not multiplied or divided.

  Example : Temperature in °C.

- A numeric attribute having a fixed zero-point is known as a ratio scale attribute. We can claim that the value is a multiple of another value if a measurement is ratio-scaled. The numbers are arranged, and we may compute the difference between them as well as calculate mean, median, quantile-range, etc.

Q2]



Scores : 85, 86, 100, 76, 81, 93, 84, 99, 71, 69, 93, 85, 81, 87, 89, 91 and 88

Total Score = 1458

Frequency = 17

$$\text{Mean} = \frac{1458}{17} = 85.76$$

Let's Subtract Mean from every Score.

| -Mean + Score | -Mean + Score | $(\text{Mean} - \text{Score})^2$ |
|---|---|---|
| 85 - 85.76 | -0.76 | 0.5776 |
| 86 - 85.76 | 0.24 | 0.0576 |
| 100 - 85.76 | 14.24 | 202.77 |
| 76 - 85.76 | -9.76 | 95.25 |
| 81 - 85.76 | -4.76 | 22.65 |
| 93 - 85.76 | 10.24 | 104.85 |
| 84 - 85.76 | -1.76 | 3.09 |
| 99 - 85.76 | 13.24 | 175.29 |
| 71 - 85.76 | 14.76 | 217.85 |
| 69 - 85.76 | 16.76 | 279.89 |
| 93 - 85.76 | 7.24 | 52.41 |
| 85 - 85.76 | -0.76 | 0.57 |
| 81 - 85.76 | -5.76 | 33.17 |
| 87 - 85.76 | 1.24 | 1.53 |
| 89 - 85.76 | 3.24 | 10.49 |
| 91 - 85.76 | 5.24 | 27.45 |
| 88 - 85.76 | 2.24 | 5.0176 |

$\sum (-\text{Mean} + \text{Score})^2 = 1232.91$

$$\text{Variance} = \frac{\sum (\text{Score} - \text{Mean})^2}{n}$$

$$= \frac{1232.91}{17} = 72.52$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{72.52}$$

$$= 8.51$$

## Conclusion

Since variation is somehow low, teacher knows the most student are performing around same level.

The standard devaition is 8.51 which is somehow low. The teacher know that the result of sample size are somehow reliable.

Q5]

Classification is a process of categorizing a given set of data into classes. In Machine Learning, we frame a problem, collect and clean the data, and some necessary feature variables (if any), train the model, measure its performance, improve it by using some cost function and then it is ready to be deployed.

A great way to measure the performance of a classified is through the use of a confusion matrix.

CONFUSION MATRIX.

A confusion matrix is a summary of predictions result on a ~~the~~ classification problem. A number of correct and incorrect predictions are summerized with count values and broken down by each class. This is the key to the Confusion Matrix.

Confusion matrix shows the way in which your classification model is confused when it makes prediction.

It gives you insight not only into the errors being made by your classifier but more importantly the type of error that are being made. It is this breakdown that overcomes the limitation of using a classifier accuracy alone.

## Real-World Problem.

Let us predict whether a photograph contains a man or women.

We have test dataset of 10 records, with their expected and predicted outcomes.

| Expected | Predicted |
|---|---|
| M | F |
| M | M |
| F | F |
| M | M |
| F | M |
| F | F |
| F | F |
| M | M |
| M | F |
| F | F |

Accuracy = $\frac{\text{total correct prediction}}{\text{prediction (total)}} * 100$

$= \frac{7}{10} \times 100 = 70\%$.

Let's turn our errors into confusion Matrix.

POINTS

→ 1 M classified as 3 M

→ 2 F classified as 4 F

We will calculate incorrect prediction.

→ 1 M classified as 2 F

→ 2 F classified as 1 M

Confusion Matrix →

| | Men (M) | Women (F) |
|---|---|---|
| 2-Men | 3 | 1 |
| 3-Female | 2 | 4 |

Q4]

## Performing One-Way ANOVA test in R

1. Installing the Package.

```
> install.package("dplyr")
```

2. Loading the Package.

```
> library("dplyr")
```

3. Plotting variance in mean value within group and between groups.

```
> boxplot (mtcars$disp~factor(mtcars$gear),
          xlab="gear", ylab="disp")
```

4. Set up null hypothesis and Alternal hypothesis.

H0 = mu = mu01 = mu02

There is no difference between avg displacement for different year.

H1 = Not all means are equal.

5. Calculate test statistics using aov function.

```
> mtcars_aov = aov(mtcars$disp~factor(mtcars$gear))
> summary(mtcars_aov)
```

6. Calculate F-Critical value.

For 0.05: significant value, critical value = alpha
= 0.05

Now, compare test statistics with F-Critical value and conclude test p < alpha, Reject Null hypothesis.

## Performing Two-Way ANOVA in R.

1. Installing the package.

```
> install.package("dplyr")
```

2. Loading the package

```
> Library("dplyr")
```

3. Plotting variance in mean within group and between group.

```
> boxplot(mtcars$disp~mtcars$gear,
          subset=(mtcars$am==0),
          xlab="gear", ylab="disp",
          main="Automatic")
```

```
> boxplot(mtcars$disp ~ mtcars$gear,
          subset = (mtcars$am == 1),
          xlab = "gear", ylab = "disp",
          main = "Manual")
```

4. Setup Null hypothesis & Alternate hypothesis

H0 = mu0 = mu01 = mu02

5. Calculate test statistics.

```
> mtcars$aov2 = aov(mtcars$disp ~ factor(mtcars$gear),
                    factor(mtcars$am)).
```

6. Calculate F-Critical Value

Q3]

## Covariance.

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

It can be simplified in computation as:

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

Correlation coefficient $\Rightarrow r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where,

$n$ = no. of tuples

$\bar{A}, \bar{B}$ = respective mean or expected value of A & B.

$\sigma_A, \sigma_B$ = respective standard deviation of A and B.

### Positive Covariance

- If $Cov_{A,B} > 0$, then A and B tend to be larger than the expected values.

### Negative Covariance.

- If $Cov_{A,B} < 0$, then if A is larger than it's expected value, B is likely to be smaller than its expected value.

### Independeance

- $Cov_{A,B} = 0$ but converse is not true.
- Some pair of random value may have a covariance of 0 but are not independent. Under some additional assumptions, covariance 0 imply independence.

we have two Stack A and B with values :-

(2, -5), (3, 8), (5, 10), (4, -11), (6, 14), (7, -15)

$E(A) = (2 + 3 + 5 + 4 + 6 + 7)/6$

$= 27/6 = 4.5$

$E(B) = (-5 + 8 + 10 + (-11) + 14 + (-15))/6$

$= (22 - 31)/6$

$= +1/6 = 0.16$

$Cov(A, B) = (-10 + 24 + 50 - 44 + 84 - 105) - (4.5 \times 0.16)$

$= (-1) - (0.75) = -1.75$

$\because Cov(A, B) < 0$

$\therefore$ A and B will not Rise together.