**Advance Statistics - Final Assignment**

**Question 1 (20 marks)**

A salary survey was conducted to explore the monthly salary of a sample of employees from different education background, who are working in public and private organizations.

Use the **Salary Data** Sheet provided to perform the following tasks and comment on the results:

1. Identify the type of each parameter (Qualitative & Quantitative).
2. Create summary table for Job Level.
3. Draw bar graph for Education.
4. Draw pie chart for Sector.
5. Draw a histogram for salary.
6. Calculate descriptive statistics of Salary and test its normality.
7. Draw box blot for Age and determine the existence of outliers.
8. Draw a Pareto chart for total salary of each Discipline & present your conclusion about the vital few.

**Answer:**

1. In the **Salary Data** provided, an analysis was conducted to discern the nature of each parameter, categorizing them into qualitative and quantitative types. This classification aids in understanding the fundamental characteristics of the dataset and informs subsequent analytical approaches.

**Qualitative Parameters:**

Qualitative parameters, also referred to as categorical data, encompass attributes that represent distinct categories or labels rather than numerical values. In the dataset, the following parameters were identified as qualitative:

['Job Level', 'Sector', 'Education', 'Discipline', 'Salary']

**Quantitative Parameters:**

Quantitative parameters entail measurable quantities, typically expressed as numerical values. These parameters facilitate quantitative analysis and statistical computations. Within the dataset, the following parameters were recognized as quantitative:

['Name', 'Age', 'Customer Satisfaction']

1. To create a summary table for the Job Level in your dataset, we use the groupby function in pandas to group the data by the Job Level and then calculate summary statistics or counts. The code by Python:

A close-up of a computer code

Description automatically generated

After we import our data and the necessary library we see that we have **83 junior ,**

**295 manager and 305 senior .**

1. **To draw a bar graph for the 'Education' we can use Plotly Express it`s strong library in python to draw graphs:**

**A screenshot of a computer

Description automatically generated**

1. **To draw pie chart for sector:**

**We use px.pie() to create a pie chart using Plotly Express.**

**Then, names='Sector' specifies the 'Sector' column as the data for the pie chart.**

**We sets the title of the pie chart**

**title='Pie Chart of Sector' sets the title of the pie chart.**

**Finaly, fig.show() displays the pie chart.**

**The final pie chart for Sector:**

**A blue and red circle with a red circle

Description automatically generated**

1. we can use Plotly Express again to draw a histogram for the 'Salary' column in our dataset. Here's how we do it:

**First,** we need to **convert the salary data from strings to integers by removing the commas and converting them to numeric format.**

**Second,** create a DataFrame from the numeric salary data.

**Finally, we create a histogram using Plotly Express with the 'Salary' column as the data for the histogram, using the numeric values as bins, and count the frequency of occurrence of each salary value.**

**A graph with blue lines and text

Description automatically generated with medium confidenceThe Final histogram:**

1. To Calculate descriptive statistics of Salary and test normality we use **Minitab**

**First**, To Calculate Descriptive Statistics:

We Open Minitab and import our dataset.

Go to "Stat" > "Basic Statistics" > "Display Descriptive Statistics".

In the "Variables" box, select the 'Salary' column.

**And there is the final answer by the steps from Minitab:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Second, Test Normality:**

**After obtaining the descriptive statistics we test the normality assumption.**

**Go to "Stat" > "Basic Statistics" > "Normality Test".**

**In the "Variable" box, select the 'Salary' column.**

**And there is the final answer by the steps from Minitab:**

**A screenshot of a computer

Description automatically generated**

****

**We can also use python to answer the above question:**

we use the describe() function to obtain summary statistics and perform a normality test using statistical methods

**A screenshot of a computer

Description automatically generated**

Since the p-value is extremely small (close to zero), much smaller than any conventional significance level (such as 0.05), it provides strong evidence against the null hypothesis. Therefore, you would reject the null hypothesis and conclude that the data significantly deviates from a normal distribution.

**In summary, based on the normality test result with the given p-value, you would conclude that the 'Salary' data is not normally distributed.**

1. **To draw a boxplot for the 'Age' column in our dataset and determine the existence of outliers, we use seaborn, a Python data visualization library. Here's how you can do it:**

**A screen shot of a diagram

Description automatically generated**

**To identify the outliers, we use the interquartile range (IQR) method. Here's how we do it:**

**A screenshot of a computer

Description automatically generated**

**We conclude that there are no outliers in the “Age” factor.**

8)Draw a Pareto chart for total salary of each Discipline & present your conclusion about the vital few.

To create a Pareto chart for the total salary of each discipline, We do the following steps using Python:

1.Calculate the total salary for each discipline.

2.Sort the disciplines based on their total salary.

3.Create a Pareto chart showing the cumulative percentage of total salary.

4.Analyze the vital few disciplines based on their contribution to the total salary.

**As the following:**

**A screenshot of a computer program

Description automatically generated**

**A graph with a red line

Description automatically generatedAnd there is the output:**

**And Based on the Output , we can draw the following conclusions:**

**Based on these findings, it's evident that Engineering, IT, and Business are the vital few disciplines, collectively contributing to a significant portion of the total salary. These disciplines play a crucial role in shaping the overall salary distribution within the dataset. Therefore, focusing on these disciplines may be essential for strategic decision-making or resource allocation within the organization.**

**...**

**Question 2 (20 marks)**

1. You are working in TV set factory. The manufactured TV has a normal distribution life with *m* = 3,500 working hours and *s* = 200 hours.
2. What is the probability that a TV will work less than 3,350 hours?

We will answer the following question using Minitab:

**"Calc" > "Probability Distributions" > "Normal".**

**And we will input data as the following:**

A screenshot of a computer

Description automatically generated

**The output:**

**A screenshot of a computer function

Description automatically generated**

**The probability obtained from the Z-table is approximately 0.2266**

**Finally, probability as a percentage. Therefore, the probability that a TV will work less than 3350 hours is approximately 22.66%**

1. What is the probability that a TV will work more than 3,750 hours?

As the previous question we will do the same steps

A screenshot of a computer

Description automatically generated**input data as the following:**

A white background with black text

Description automatically generated

But in this question, we don’t want to calculate the cumulative probability of less than 3750

So we know that the sum of probability = 1 , So to find the probability that a TV will work more than 3,750 hours, we subtract this probability from 1.

**It will be = 1-0.894350 = 0.10565**

**Finally, probability as a percentage. Therefore, the probability that a TV will work more than 3750 hours is approximately 10.565%**

1. What is the probability that a TV will work between 3,350 & 3,750 hours?

To find the probability of a TV working between 3,350 and 3,750 hours, **we can subtract the probability of the TV working less than 3,750 hours from the probability of the TV working less than 3,350 hours**. Since we calculated these values previously, the calculation would be as follows:

**0.894350−0.2266=0.66775**

**Therefore, the probability that a TV will work between 3,350 and 3,750 hours is approximately** **66.775%."**

1. What is TV life that you are confident 95% it will keep working?

To find the TV life such that you are confident 95% it will keep working, we need to find the value of the TV's working hours at the 95th percentile of the normal distribution. This value represents the point below which 95% of the TV working hours fall.

And we will use Minitab to find it

A screenshot of a computer

Description automatically generatedWe will change the cumulative probability to the inverse cumulative probability and input value will be 1 – 0.95 = 0.05 as the following:

And the final result is

A white background with black text

Description automatically generated

Therefore, with a confidence level of 95%, we can expect the majority of TVs to continue working for at **least 3171.03 hours.**

1. You are working in a bank. You have collected enough data to determine the average time needed to serve one customer and found that it follows a normal distribution with *m* = 4.78 minutes and *s* = 1.32 minutes.
2. What is the probability that you will serve 10 customers every hour?

Mean (μ) = 4.78 minutes

Standard deviation (σ) = 1.32 minutes

**Let's calculate the total time to serve 10 customers:**

Since 10 customers are to serve in every hour , 60 minutes, each customer will be served in an average of **6 minutes**

To find this probability, we can use the cumulative distribution function (CDF) of the normal distribution, and we will use Minitab:

A screenshot of a computer

Description automatically generated

And the output is:

A black text on a white background

Description automatically generated

**Therefore, the probability that you will serve 10 customers every hour is nearly** **88.2320%.**

1. What is the probability that you will serve more than 15 customers every hour?

We will use Minitab But, in this question, we don’t want to calculate the cumulative probability of less than 15 customers every hour .

So we know that the sum of probability = 1 , So to find the probability that you will serve more than 15 customers every hour, we subtract this probability from 1.

Since 15 customers or more are to serve in every hour , 60 minutes, each customer will be served in an average of 4 minutes. The probability that you will serve more than 15 customers every hour by Minitab:

A black text on a white background

Description automatically generated

**P (X>4) = 1-0.27729 = 0.72271**

Hence, the probability that you will serve more than 15 customers every hour is **0.72271 or 72.271%**

1. What is the probability that you will serve between 10 & 15 customers every hour?

To find the probability of serve between 10 & 15 customers every hour, **we can subtract the probability of less than 15 customers every hour from the probability of serve 10 customers every hour**. Since we calculated these values previously, the calculation would be as follows:

**0.882320- 0.27729 =0.60503**

**Therefore, the probability that** serve between 10 & 15 customers every hour **approximately** **60.503%."**

1. What is the number of customers you will be 95% confident that you will serve every hour?

To find the number of customers you will be 95% confident that you will serve every hour

we will use Minitab to find it

We will change the cumulative probability to the inverse cumulative probability and input value **will be 1 – 0.95 = 0.05 as the following:**

A white background with black text

Description automatically generated

Hence, the number of customers at which 95% confident is achieved that you will serve every hour is **nearly 3 customers.**

1. You are thinking about signing a contract, as a supplier for one of the biggest global exporting company. The draft contract obligates you to deliver 20 tons of orange every week. The delivery process of orange during this season follows a normal distribution with *m* = 22.5 tons every week and *s* = 3.2 tons.
2. What is the probability that you will achieve the contract terms?

To find the probability of achieving the contract terms, which is delivering at least 20 tons of oranges every week, we need to calculate the probability of the delivery process being greater than or equal to 20 tons.

**Mean (μ) = 22.5 tons**

**Standard deviation (σ) = 3.2 tons**

This represents the probability of delivering at least 20 tons of oranges every week.

In Minitab, you can use the Probability Distribution Calculator for the normal distribution to find this probability we will input data as the following:

A screenshot of a computer

Description automatically generated

And the output is :

A white background with black text

Description automatically generated

Based on the calculation, the probability of achieving the contract terms, which is delivering at least 20 tons of oranges every week, is approximately **0.21738 or 21.738%.**

1. What is the orange quantity that you will be 95% confident that you will deliver every week?

To find the orange quantity that you will be 95% confident that you will deliver every week we will use Minitab to find it

We will change the cumulative probability to the inverse cumulative probability and input value **will be 1 – 0.95 = 0.05 as the following:**

**A white background with black text

Description automatically generated**

**Hence, the orange quantity at which 95% confident is achieved that you will deliver every week is 17.2365, nearly 17 tons oranges.**

**Question 3 (15 marks)**

A supplier was requested to deliver order within 25 to 35 days after receiving the Purchase Order.

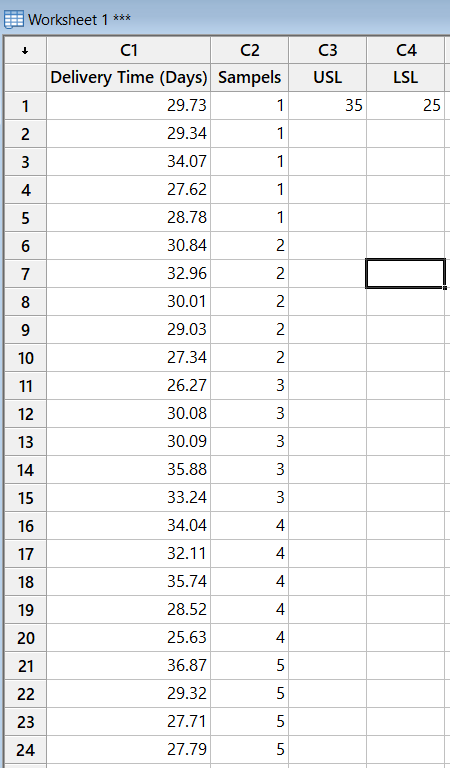
Use **Delivery Time Data** to perform the following:

1. Calculate the capability indices and provide your comments on the results.

We will use Minitab to calculate the capability indices so we will import the **Delivery Time Data**

And we will divide the data we have into **13 groups**, each group containing **5 values as wrote in the Delivery Time Data**

**The data in Minitab will be like that :**



After that we going to “State” 🡪 “Capability Analysis” 🡪 “Normal”

As the following:

A screenshot of a computer

Description automatically generated

The output will be:

**Based on the provided data and the estimated capability indices in Minitab, we analyze the results and provide it:**

**1)Overall Capability:**

* **Pp:** 0.51
* **PPL:** 0.59
* **PPU:** 0.43
* **Ppk:** 0.43
* **Cp:** 0.50
* **CPL:** 0.58
* **CPU:** 0.42
* **Cpk:** 0.42

2) **Process Data:**

* **Sample Mean:** 30.7539
* **Sample N:** 64
* **StDev(Overall):** 3.27027
* **StDev(Within):** 3.33123

3) **Performance:**

* **PPM < LSL:** 15625.00
* **PPM > USL:** 109375.00
* **PPM Total:** 125000.00
* **Expected Within:** 42059.40

**Conclusion:**

* The calculated capability indices indicate that the supplier's process has moderate capability to meet the specified delivery time requirements.
* The Ppk value of 0.43 suggests that the process is slightly off-center with respect to the specification limits, indicating the need for improvement.
* While the Ppk is slightly lower than desirable, it still indicates that the process is capable of meeting the customer requirements to a reasonable extent.
* The process data shows that the sample mean is close to the target, but there is variation within the process, as indicated by the standard deviations.

1. Provide your recommendation to improve the supplier performance.

**from my point of view to improve supplier performance in a statistical way, we can focus on utilizing statistical tools and methodologies to identify areas for improvement and optimize the delivery process, like:**

1. **Statistical Process Control (SPC):** Implement SPC techniques such as control charts to monitor delivery times over time. By analyzing control charts, the supplier can identify trends, shifts, or anomalies in the process that may indicate underlying issues affecting performance. This proactive approach allows for early detection of problems and timely intervention.
2. **Regression Analysis:** Perform regression analysis to identify factors influencing delivery times. By analyzing historical data, regression models can help identify correlations between various factors such as order volume, transportation mode, or supplier location, and delivery performance. This insight can inform strategic decisions to optimize delivery processes and mitigate delays.
3. **Design of Experiments (DOE):** Utilize DOE techniques to systematically evaluate the impact of different process variables on delivery performance. By conducting controlled experiments, the supplier can identify the most effective combinations of factors that lead to improved delivery times. This empirical approach enables data-driven decision-making and optimization of delivery processes.

**There are many statistical methods other than that, but I think they are the most important to improve the supplier performance.**

**Question 4 (10 marks)**

Cycle Time was measured and found to be a non-normal distribution.

Use **Cycle Time** **Data** to perform the following:

1. Transform Cycle Time data using Box-Cox transformation.
2. Transform Cycle Time data using Johnson transformation.