# **Module 3 - Statistics Essential for Data Science – Assignment**

# **-** Diwakar Sinha



[This Photo](https://online.stat.psu.edu/statprogram/reviews) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/)

# Question 1. Pizza Delivery Outlets analysis

In question 1 we are given 20 random instances of time taken delivery times of pizza within a radius of 1 km for 3 pizza outlets and we are tasked with analyzing the delivery performance of the 3 outlets and their summary statistics

**a. Summary statistics for all the outlets:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Outlet 1* | *Outlet 2* | *Outlet 3* |
|  |  |  |  |
| Mean | 25.5 | 29.75 | 27.85 |
| Standard Error | 0.941723 | 1.033326 | 0.968653 |
| Median | 24 | 31 | 27.5 |
| Mode | 23 | 30 | 33 |
| Standard Deviation | 4.211513 | 4.621176 | 4.33195 |
| Sample Variance | 17.73684 | 21.35526 | 18.76579 |
| Kurtosis | -0.88491 | -0.5089 | -1.38871 |
| Skewness | 0.671702 | -0.81798 | -0.14566 |
| Range | 13 | 15 | 12 |
| Minimum | 20 | 20 | 21 |
| Maximum | 33 | 35 | 33 |
| Sum | 510 | 595 | 557 |
| Count | 20 | 20 | 20 |
| Critical value **CV =** | 2.093 | from t-distribution with 19 [n-1] degrees of freedom | |
| Margin of Error = CV \* Std. Error | 1.971026 | 2.162752 | 2.027392 |
| b) | *Outlet 1* | *Outlet 2* | *Outlet 3* |
| 95% confidence interval Upper Limit | 27.47103 | 31.91275 | 29.87739 |
| 95% confidence interval Lower Limit | 23.52897 | 27.58725 | 25.82261 |

**Method to calculate 95% confidence interval:**

1. Determine the critical value for a 95% confidence level using a t-distribution table with n-1 degrees of freedom (where n is the sample size [20 in our case]).
2. Calculate the margin of error using the formula:
3. margin of error = critical value \* standard error (from table above)
4. Calculate the upper and lower limits of the confidence interval using the formulas:
   1. upper limit = sample mean + margin of error
   2. lower limit = sample mean - margin of error

**c. We can interpret each of the above summary statistics as follows:**

1. **Mean**: The mean is the average value of the variable across all the observations in each outlet. Outlet 2 has the highest mean (29.75), which suggests that it has the highest average value of the variable among the three outlets.
2. **Standard** **Error**: The standard error measures the variability of the sample mean. A smaller standard error indicates more precision in the estimate of the population mean. The standard error is similar for all three outlets, indicating that the sample means are equally precise.
3. **Median**: The median is the middle value of the variable when the observations are arranged in order. Outlet 2 has the highest median (31), which suggests that it has a higher number of observations with values above the median compared to the other two outlets.
4. **Mode**: The mode is the value that occurs most frequently in the variable. Outlet 3 has the highest mode (33), which suggests that this value is the most common among the observations in Outlet 3.
5. **Standard Deviation**: The standard deviation measures the variability of the variable within each outlet. Outlet 2 has the highest standard deviation (4.621), which suggests that it has more variability in the values of the variable compared to the other two outlets.
6. **Sample Variance**: The sample variance is a measure of the spread of the variable around the mean. Outlet 2 has the highest sample variance (21.355), indicating that the values of the variable in Outlet 2 are more spread out compared to the other two outlets.
7. **Kurtosis**: The kurtosis measures the degree of peakedness of the distribution of the variable. Outlet 3 has the lowest kurtosis (-1.389), which indicates that its distribution is flatter and more spread out compared to the other two outlets.
8. **Skewness**: The skewness measures the degree of asymmetry of the distribution of the variable. Outlet 2 has negative skewness (-0.818), indicating that its distribution is skewed to the left (i.e., has more observations with low values).
9. **Range**: The range is the difference between the maximum and minimum values of the variable. Outlet 2 has the highest range (15), indicating that it has the highest difference between the highest and lowest values of the variable among the three outlets.

***BUSINESS CONCLUSION:*** *Outlet 2 is struggling with the on time delivery of its pizzas and needs most attention in terms of process improvement and variability reduction . next is outlet 3 and best performing out of these 3 is outlet 1*

**d. Hypothesis test:**

**i.** Null Hypothesis (H0) is that there is no significant difference in the mean delivery time among the three pizza outlets.

Alternate Hypothesis (H1) is that there is a significant difference in the mean delivery time among the three pizza outlets.

**ii.**. The appropriate hypothesis test to be used is a **one-way ANOVA** (analysis of variance) test to determine whether there is a significant difference in the mean delivery time among the three pizza outlets.

**Results of one-way ANOVA are summarized below.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *Df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 181.3 | 2 | 90.65 | 4.700309 | 0.012898 | 3.158843 |
| Within Groups | 1099.3 | 57 | 19.28596 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 1280.6 | 59 |  |  |  |  |
|  |  |  |  |  |  |  |

**Explanation of terms in ANOVA output:**

The source of variation is decomposed into two components: between groups and within groups. The sum of squares (SS) represents the variation in the data that is due to the different sources of variation. The degrees of freedom (df) are the number of groups minus one for the between-group variation, and the total number of observations minus the number of groups for the within-group variation. The mean squares (MS) are the SS divided by the df.

The F-test is used to compare the variability between groups to the variability within groups. The F-statistic is calculated by dividing the MS between groups by the MS within groups. The p-value is the probability of observing an F-statistic as large or larger than the one calculated, assuming that the null hypothesis is true. The null hypothesis is that there is no difference between the means of the groups.

iii. **Interpretation**: In this case, the F-statistic is 4.700309, and the p-value is 0.012989, which is less than the significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant difference between the means of the three outlets. The critical value of F for this test, with 2 and 57 degrees of freedom, is 3.159, which is higher than the calculated F-value. This means that we can reject the null hypothesis at the 5% level of significance.

*In conclusion, this ANOVA table suggests that there is a significant difference in the means of the three outlets, based on the observed data.* Further analysis, such as post hoc tests, would be required to determine which outlets have significantly different mean delivery times.

# Question 2 Teller vs ATM

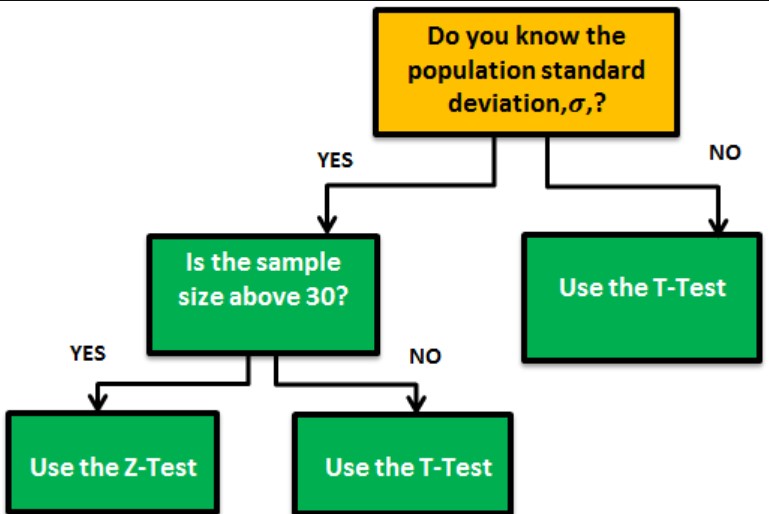
In question 2 we are given 45 independent observations of time taken in minutes to withdraw cash using u traditional manned teller at the bank vs 53 observations of time taken by people to withdraw cash from an ATM or an automated teller machine. We are tasked with analyzing the average turnaround time of withdrawing cash via a bank teller vs at the ATM

**a.** **Hypotheses**:

Null Hypothesis (H0): There is no significant difference in the turnaround time for withdrawing cash via a Teller and ATM.

Alternative Hypothesis (H1): There is a significant difference in the turnaround time for withdrawing cash via a Teller and ATM.

**b.** to decide the appropriate hypothesis test to be used, let us review the use cases for the z-test and t-test:



Since the data given is sample data, and Population std dev & var are unknown, the appropriate hypothesis test to be used is the **two-sample t-test,** assuming the samples are independent, and the variances are not equal. we are assuming unequal variances because the processes are very different and can have different variances. The results of our t-test are as follows:

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Unequal Variances | | |
|  |  |  |
|  | *Teller* | *ATM* |
| Mean | 1.4506 | 0.9851 |
| Variance | 0.1054 | 0.0440 |
| Observations | 45 | 53 |
| Hypothesized Mean Difference | 0 |  |
| df | 73 |  |
| t Stat | 8.2637 |  |
| P(T<=t) one-tail | 2.2491E-12 |  |
| t Critical one-tail | 1.6660 |  |
| P(T<=t) two-tail | 4.49821E-12 |  |
| t Critical two-tail | 1.9930 |  |

**c.** **Inference**:

Based on the output of the t-Test: Two-Sample Assuming Unequal Variances:

The sample mean turnaround time for cash withdrawal through Teller is 1.4506, while for ATM it is 0.9851.

The variances of the two samples are different (0.1054 for Teller and 0.0440 for ATM), and the degrees of freedom (df) are 73.

*The calculated t-statistic is 8.2637, and the p-value is 4.49821E-12 (which is less than the level of significance of 0.05 for 95% confidence).*

***Therefore, we reject the null hypothesis (H0) and conclude that there is a significant difference in the turnaround time for withdrawing cash via a Teller and ATM.***

The t-critical values for a one-tailed test with 73 degrees of freedom at 95% confidence are ±1.6660. The t-critical values for a two-tailed test with 73 degrees of freedom at 95% confidence are ±1.9930.

*Since the calculated t-statistic (8.2637) is greater than the t-critical value (1.9930***),** **we can conclude that the difference in the mean turnaround time between Teller and ATM is statistically significant.**

**XTRA work** - Although the hypothesis test, we have chosen is completely justified and fit for purpose, for the sake of curiosity and completeness we shall try the others as well.

2. Two sample t-test assuming equal variances, we observe that even if we assume equal variances the variances calculated in the result are actually NOT equal

*BOTH of these cases lead to the same conclusion i.e. we reject the null hypothesis (HO) there is significant statistical evidence that the there is a difference between turnaround time of teller and ATM*

3. z-test – although the population variance is not known, if we approximate population variance by sample variance and calculate