

**Answer**:

1. **T² Statistic Overview**: The Hotelling's T² statistic is a crucial measure in multivariate statistical analysis. It is used primarily for hypothesis testing and accurate inference, especially under the multivariate general linear hypothesis, such as testing whether a vector of means equals a given vector.

* **Multivariate Normality**: The T² statistic assumes that the data follows a multivariate normal distribution, indicating that all involved variables are individually normally distributed, and any linear combination of these variables is also normally distributed. This assumption forms the bedrock of many classical statistical methods, including the T² statistic.
* **Known Covariance Matrix**: Another critical assumption for the T² statistic to follow a χ² distribution is the knowledge of the true covariance matrix of the population. This covariance matrix forms the basis for calculating the T² statistic accurately, which measures the squared Mahalanobis distance between an observation and the population mean, considering the correlation structure among the variables.
* **Chi-Square Distribution**: When these assumptions are met, the T² statistic follows a χ² distribution. If we compute a T² statistic for each observation, the sum of these T² statistics across all observations will exhibit a χ² distribution. The degrees of freedom for this χ² distribution depend on the number of variables involved in the analysis.
* **Unknown Covariance Matrix and Asymptotic Behaviour**: In real-world scenarios, it is often the case that the true covariance matrix is unknown and must be estimated from the sample data. While this does cause the distribution of the T² statistic to deviate from a χ² distribution, it does still asymptotically approach a χ² distribution as the sample size increases indefinitely, under mild regularity conditions.
* **Transformed T² Statistic**: More accurately, for finite samples when the population covariance matrix is unknown, the T² statistic does not follow a χ² distribution. Instead, a transformed version of it, usually denoted as F, follows an F-distribution. This transformation compensates for the uncertainty introduced by estimating the population covariance matrix from the data.
* **Sample Size and Dimensionality**: The approximation of the T² statistic to a χ² distribution becomes more accurate as the sample size increases and/or the dimensionality (number of variables in the multivariate data) decreases. Similarly, the degrees of freedom for the F-distribution depend on the sample size and the number of variables.
* **Assumption Evaluation**: As a data scientist, it's critical to evaluate these assumptions before applying the T² statistic. If the assumptions are violated, the T² statistic may not follow a χ² or F distribution. In such cases, non-parametric methods, or alternative approaches such as bootstrap methods or permutation tests may be necessary.

**Conclusion**: In summary, the T² statistic follows a χ² distribution under specific assumptions - primarily when data conform to multivariate normality, and the population covariance matrix is known. However, when the covariance matrix is unknown and estimated from the data, the T² statistic follows an F-distribution. It's vital to verify these assumptions to make valid conclusions based on the T² statistic.

**b)**

**Dataset and Definitions:** Suppose you have a dataset with m variables and n observations. Let X be the data matrix, where each row represents an observation, and each column represents a variable. Define μ as the mean vector of the population and Σ as the known covariance matrix.

**Sample Mean Vector:** Calculate the sample mean vector, x̄ , using the formula:

x̄ = (1/n) \* Σx

**Sample Covariance Matrix**:Then, compute the sample covariance matrix, S, through the following formula:

S = (1/(n-1)) \* ((X - x̄)^T \* (X - x̄))

**T² Statistic:** For each observation, calculate the T² statistic using the equation:

T² = n \* (x - μ)^T \* Σ^(-1) \* (x - μ)

**Sum of T² Statistics:** Compute the sum of these T² statistics:

T²\_sum = ΣT²

**Distribution of T²\_sum:** Now, you can express T²\_sum as follows:

T²\_sum = n \* (x̄ - μ)^T \* (Σ^(-1)/n) \* (x̄ - μ)

Now, let us consider A = Σ^(-1)/n, then, we have:

T²\_sum = n \* (x̄ - μ)^T \* A \* (x̄ - μ)

**Quadratic Form:** Substitute y = A^(1/2) \* (x̄ - μ) to rewrite T²\_sum in terms of a quadratic form:

T²\_sum = n \* y^T \* y

**T² Distribution**: As y follows a multivariate normal distribution, the quadratic form n \* y^T \* y follows a χ² distribution with m degrees of freedom. Therefore, under the assumptions of multivariate normality and known covariance matrix, the T² statistic follows a χ² distribution with m degrees of freedom.

**Assumptions and Alternatives:** It's essential to ensure the validity of the assumptions, like multivariate normality and a known covariance matrix. If these assumptions are violated, the χ² distribution assumption for the T² statistic may no longer hold, and alternative approaches may be required for hypothesis testing and inference.

As a summary we conclude that, the T² statistic follows a χ² distribution under the conditions of multivariate normality and a known covariance matrix, the proof of which relies on manipulating the T² statistic to express it as a quadratic form of a multivariate normal vector.

A white paper with black text

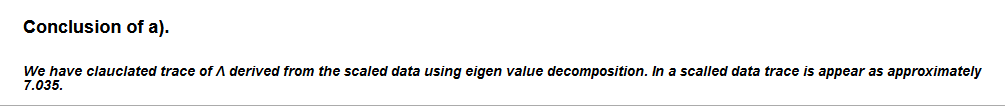
Description automatically generated

**Answer:**

**Note** : Entire report has been created with the seed value of 30. Any further modification done on the seed value may change the result and graph in the notebook.



**a)**



**b)**

A black text on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

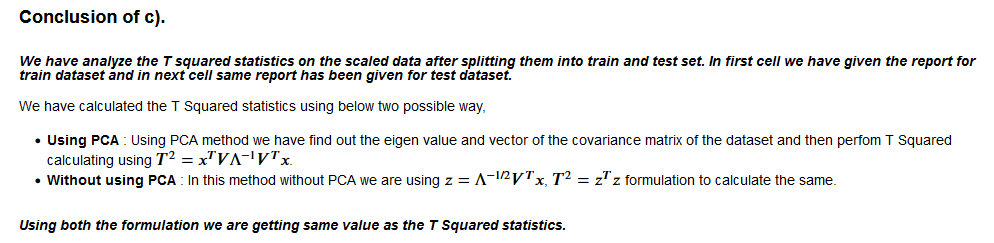
A screenshot of a computer screen

Description automatically generated

5-point statistical summary has been stored for entire dataset in attached excel file,



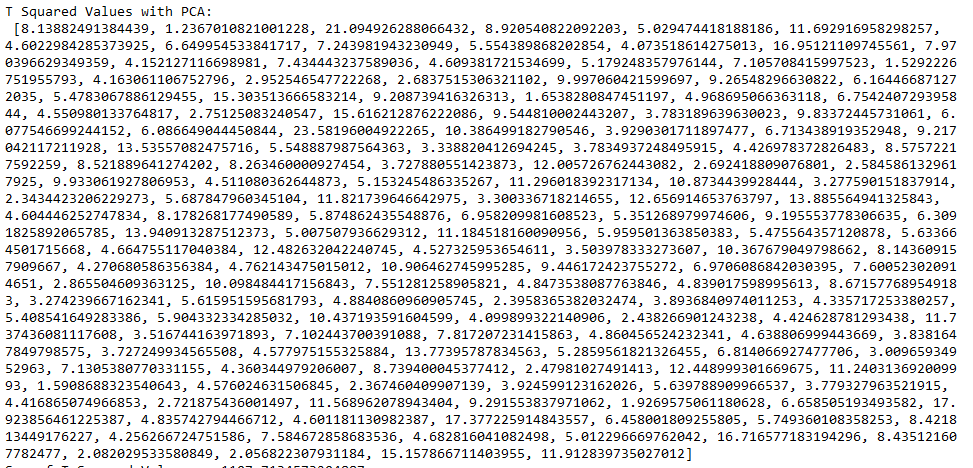
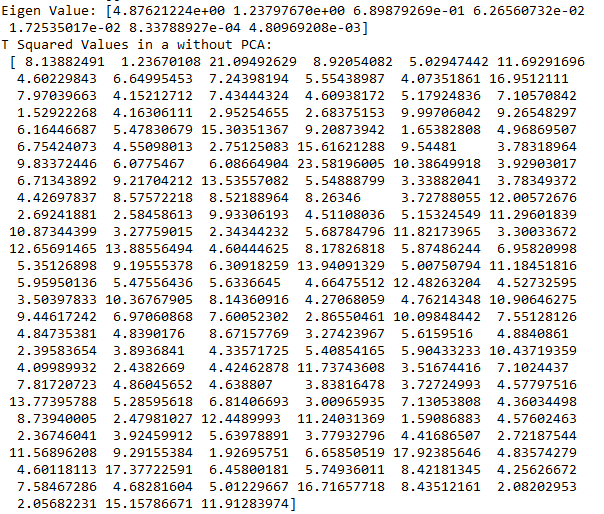
**c)**



**Analysis on Train dataset :**

A number on a white background

Description automatically generated

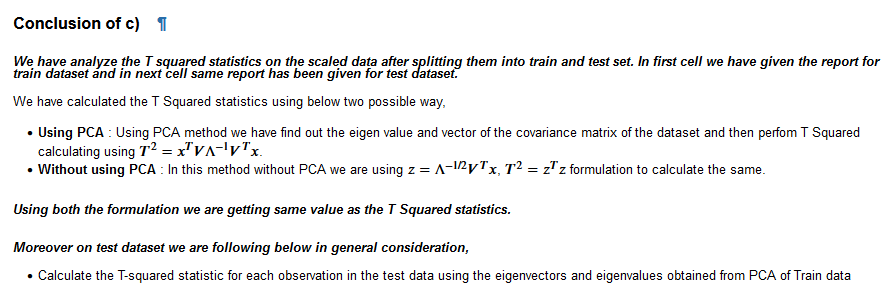


**Analysis on Test dataset :**

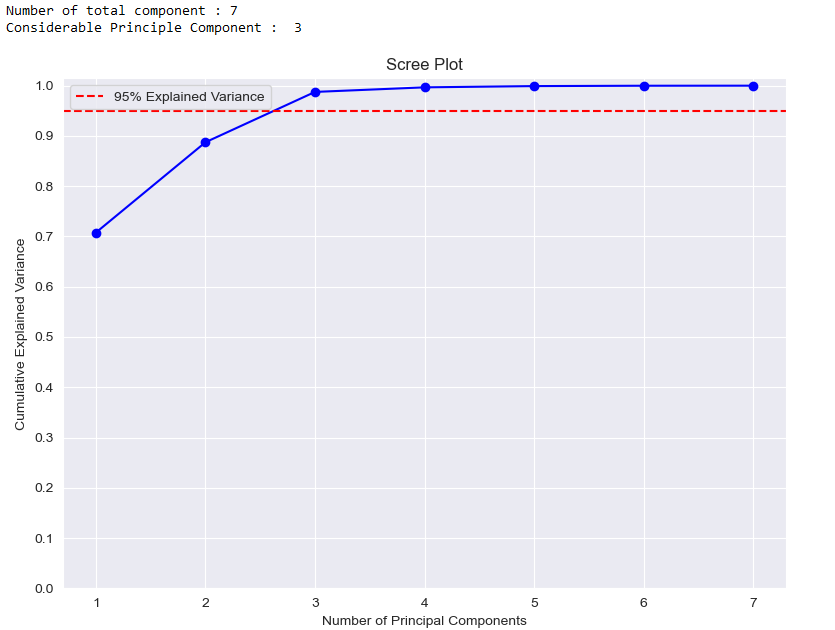
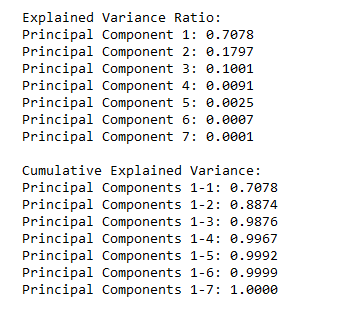


A white background with black numbers

Description automatically generated



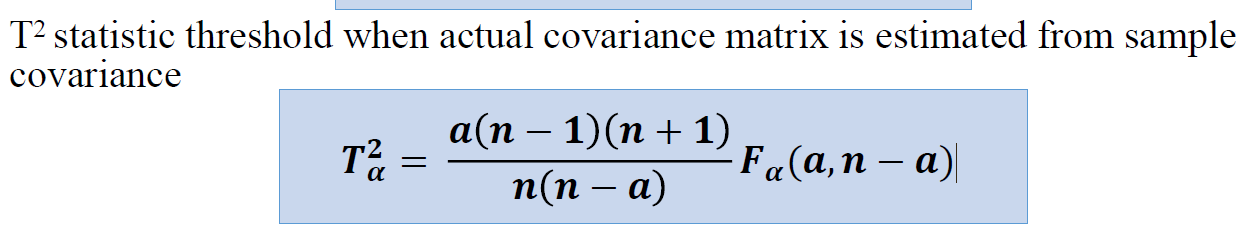
**d)**



Black text on a white background

Description automatically generated

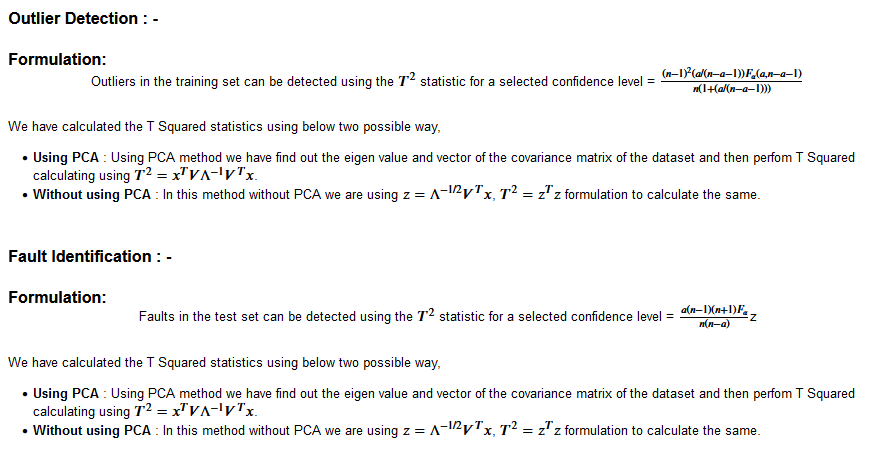
**e)**



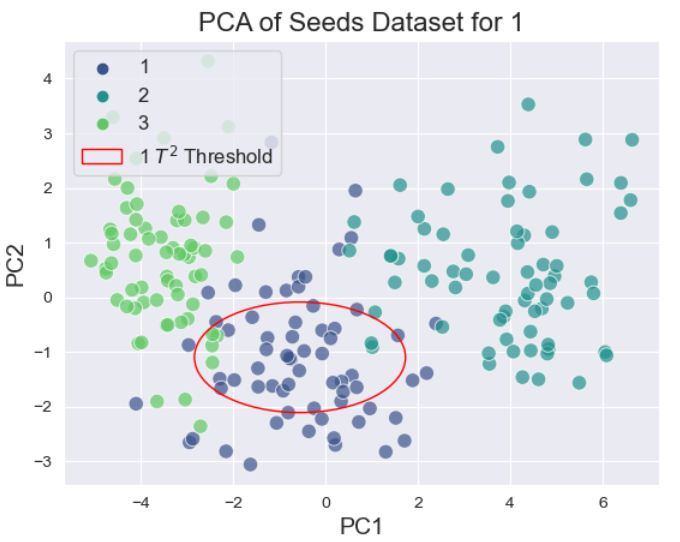
A white background with black text

Description automatically generated

**f)**



**Type 1 :**



A green sphere with orange dots

Description automatically generated

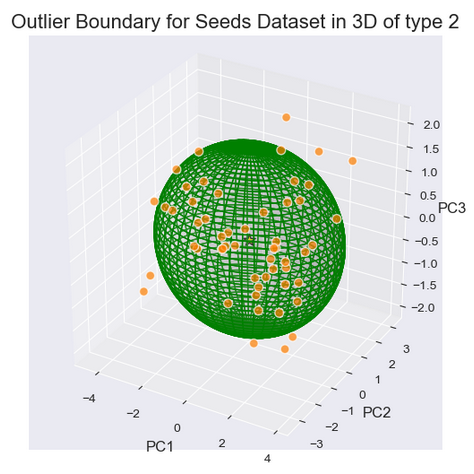
A green sphere with orange dots

Description automatically generated

**Type 2 :**

A graph showing seeds and numbers

Description automatically generated



A green sphere with lines and dots

Description automatically generated

**Type 3 :**

A graph showing seeds and numbers

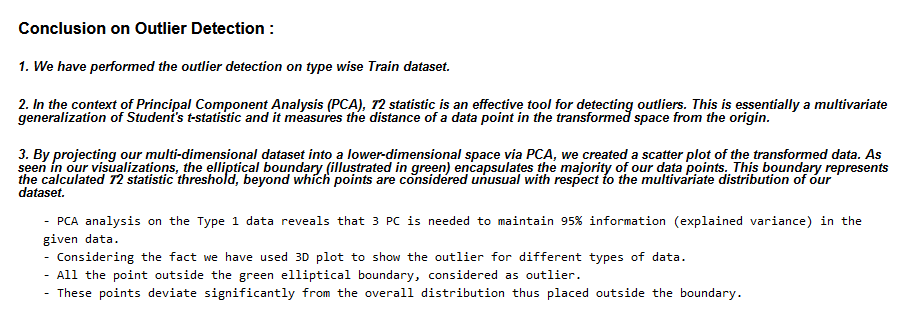
Description automatically generated

A green sphere with orange dots

Description automatically generated

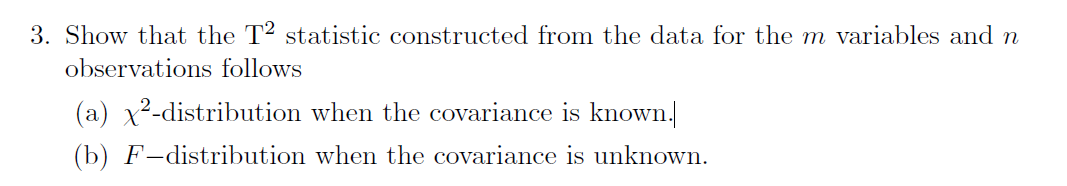
A green sphere with lines and dots

Description automatically generated



A white text on a white background

Description automatically generated



**Answer**:

a)

**T² Statistic Overview**: The Hotelling's T² statistic is a crucial measure in multivariate statistical analysis. It is used primarily for hypothesis testing and accurate inference, especially under the multivariate general linear hypothesis, such as testing whether a vector of means equals a given vector.

* **Multivariate Normality**: The T² statistic assumes that the data follows a multivariate normal distribution, indicating that all involved variables are individually normally distributed, and any linear combination of these variables is also normally distributed. This assumption forms the bedrock of many classical statistical methods, including the T² statistic.
* **Known Covariance Matrix**: Another critical assumption for the T² statistic to follow a χ² distribution is the knowledge of the true covariance matrix of the population. This covariance matrix forms the basis for calculating the T² statistic accurately, which measures the squared Mahalanobis distance between an observation and the population mean, considering the correlation structure among the variables.
* **Chi-Square Distribution**: When these assumptions are met, the T² statistic follows a χ² distribution. If we compute a T² statistic for each observation, the sum of these T² statistics across all observations will exhibit a χ² distribution. The degrees of freedom for this χ² distribution depend on the number of variables involved in the analysis.
* **Unknown Covariance Matrix and Asymptotic Behaviour**: In real-world scenarios, it is often the case that the true covariance matrix is unknown and must be estimated from the sample data. While this does cause the distribution of the T² statistic to deviate from a χ² distribution, it does still asymptotically approach a χ² distribution as the sample size increases indefinitely, under mild regularity conditions.
* **Transformed T² Statistic**: More accurately, for finite samples when the population covariance matrix is unknown, the T² statistic does not follow a χ² distribution. Instead, a transformed version of it, usually denoted as F, follows an F-distribution. This transformation compensates for the uncertainty introduced by estimating the population covariance matrix from the data.
* **Sample Size and Dimensionality**: The approximation of the T² statistic to a χ² distribution becomes more accurate as the sample size increases and/or the dimensionality (number of variables in the multivariate data) decreases. Similarly, the degrees of freedom for the F-distribution depend on the sample size and the number of variables.
* **Assumption Evaluation**: As a data scientist, it's critical to evaluate these assumptions before applying the T² statistic. If the assumptions are violated, the T² statistic may not follow a χ² or F distribution. In such cases, non-parametric methods, or alternative approaches such as bootstrap methods or permutation tests may be necessary.

In summary, the T² statistic follows a χ² distribution under specific assumptions - primarily when data conform to multivariate normality, and the population covariance matrix is known. However, when the covariance matrix is unknown and estimated from the data, the T² statistic follows an F-distribution. It's vital to verify these assumptions to make valid conclusions based on the T² statistic.

**Proof for the same is as below**,

**Dataset and Definitions:** Suppose you have a dataset with m variables and n observations. Let X be the data matrix, where each row represents an observation, and each column represents a variable. Define μ as the mean vector of the population and Σ as the known covariance matrix.

**Sample Mean Vector:** Calculate the sample mean vector, x̄ , using the formula:

x̄ = (1/n) \* Σx

**Sample Covariance Matrix**:Then, compute the sample covariance matrix, S, through the following formula:

S = (1/(n-1)) \* ((X - x̄)T \* (X - x̄))

**T² Statistic:** For each observation, calculate the T² statistic using the equation:

T² = n \* (x - μ)T \* Σ-1 \* (x - μ)

**Sum of T² Statistics:** Compute the sum of these T² statistics:

T²\_sum = ΣT²

**Distribution of T²\_sum:** Now, you can express T²\_sum as follows:

T²\_sum = n \* (x̄ - μ)T \* (Σ-1/n) \* (x̄ - μ)

Now, let us consider A = Σ^(-1)/n, then, we have:

T²\_sum = n \* (x̄ - μ)T \* A \* (x̄ - μ)

**Quadratic Form:** Substitute y = A^(1/2) \* (x̄ - μ) to rewrite T²\_sum in terms of a quadratic form:

T²\_sum = n \* yT \* y

**T² Distribution**: As y follows a multivariate normal distribution, the quadratic form n \* yT \* y follows a χ² distribution with m degrees of freedom. Therefore, under the assumptions of multivariate normality and known covariance matrix, the T² statistic follows a χ² distribution with m degrees of freedom.

**Assumptions and Alternatives:** It's essential to ensure the validity of the assumptions, like multivariate normality and a known covariance matrix. If these assumptions are violated, the χ² distribution assumption for the T² statistic may no longer hold, and alternative approaches may be required for hypothesis testing and inference.

As a summary we conclude that, the T² statistic follows a χ² distribution under the conditions of multivariate normality and a known covariance matrix, the proof of which relies on manipulating the T² statistic to express it as a quadratic form of a multivariate normal vector.

b)

In most practical data analysis scenarios, we encounter situations where the covariance is unknown. When this happens, we need to estimate the population covariance matrix, typically denoted as S, from our sample data. Naturally, estimating S introduces a certain level of uncertainty into the analysis, causing the distribution of Hotelling's T² statistic to follow an F-distribution rather than a Chi-Square distribution.

The T² statistic in this scenario takes the form:

T² = n \* (x̄ - μ)T \* S-1 \* (x̄ - μ)

* The Sample Mean Vector: x̄, is computed directly from the dataset.
* The Sample Covariance Matrix: S, is also derived from the dataset.
* The T² Statistic: For each observation, we compute the T² statistic using the given formula.

Degrees of Freedom: With the F-distribution, we deal with two degrees of freedom parameters. The first one corresponds to the number of variables m, while the second is n - m, with n being the sample size.

Given that the covariance matrix is estimated from the data, the T² statistic now adheres to an F-distribution, specifically with m and n - m degrees of freedom. This change is rooted in the following factors:

* **The Sample Mean Vector**: x̄ is distributed according to a multivariate normal distribution, centered around the population mean vector μ, with a covariance matrix of Σ/n.
* **The Sample Covariance Matrix**: S follows a Wishart distribution with n - 1 degrees of freedom and a scale matrix of Σ.

The T² statistic can be re-expressed as a ratio of two quantities:

T² = n \* (x̄ - μ)' \* S-1 \* (x̄ - μ) = [(n-1) \* m / (n-m)] \* (F)

In this equation, F adheres to an F-distribution with m and n - m degrees of freedom.

To understand the logic behind this, consider that the T² statistic can be conceptualized as the squared Mahalanobis distance between the sample mean vector x̄ and the population mean vector μ. This distance is divided by the variability of the distance, gauged by the inverse of the sample covariance matrix S. As both the numerator and denominator of this ratio are dependent on the data and hence are random variables, their ratio's distribution follows an F-distribution.

First, let's define the population multivariate data vector as X = (X1, X2, ..., Xm)', where m is the number of variables. The multivariate sample mean vector and covariance matrix are x̄ = (x̄1, x̄2, ..., x̄m)' and S, respectively.

The Hotelling's T^2 statistic is defined as:

T2 = n(x̄ - μ)'S-1(x̄ - μ)

where μ is the population mean vector and S is the sample covariance matrix.

If X ~ N(μ, Σ) where X is the vector of observations, μ is the vector of means, and Σ is the covariance matrix, we know that (x̄ - μ)'S-1(x̄ - μ) ~ χ2 (m, a) chi-square distribution with m degrees of freedom, from the properties of the multivariate normal distribution.

We also know that S ~ W(Σ, n-m), a Wishart distribution with parameters Σ and n-m. This implies that S/(n-m) ~ W(Σ/(n-m), n-m), which follows an F-distribution.

From the previous steps, we know that T2 = n(x̄ - μ)'S-1(x̄ - μ). Rearranging, we can write T^2 as:

T2 = m(n - m)(x̄ - μ)'[S/(n - m)]-1(x̄ - μ)

This can be recognized as a ratio of two quadratic forms in the multivariate normal distribution. Given the independence of the two quadratic forms, it can be shown that the ratio follows an F-distribution.

Therefore, if we divide T2 by m(n - m), it will follow an F distribution with m and (n - m) degrees of freedom:

T2 / m(n - m) ~ F(m, n - m)

We base our conclusions on several assumptions, including the multivariate normality of the data. If these assumptions do not hold, the T² statistic may not follow an F-distribution, warranting alternative approaches for hypothesis testing and inference.

In conclusion, under the conditions of multivariate normality and an unknown covariance matrix, the T² statistic will follow an F-distribution with m and n - m degrees of freedom. The proof of this relies on the relationship between multivariate normal and Wishart distributions, and the characteristics of the F-distribution as the ratio of two scaled chi-squared random variables. As data scientists, we must always validate these assumptions to ensure the credibility of our analyses.

**End of Report**