Topic for the class-ANOVA

Unit _3 : Title-Descriptive statistics

Date & Time : 2.9.24 11.00 AM – 11.50 AM

Dr. Bhramaramba Ravi
Professor
Department of Computer Science and Engineering
GITAM School of Technology (GST)
Visakhapatnam – 530045
Email: **bravi@gitam.edu**

Unit3-syllabus

- **UNIT 3 Descriptive statistics 9 hours, P - 2 hours**
- Measures of Central Tendency – Measures of Variation – Quartiles and Percentiles –

Moments – Skewness and Kurtosis. Exploratory Data Analytics Descriptive Statistics – Mean,

Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA, Random variable, Variance, covariance, and correlation- Linear transformations of random variables, Regression.

- https://www.coursera.org/learn/data-visualization-r

Definition of ANOVA

- ANOVA, which stands for **Analysis of Variance**, is a statistical method used to determine whether there are statistically significant differences between the means of three or more independent groups.
- It assesses the impact of one or more factors by comparing the means of different samples.
- The fundamental idea behind ANOVA is to analyze the variance within each group and between the groups to ascertain if any observed differences in sample means are due to random chance or if they reflect true differences in population means.

Assumptions of ANOVA

- To properly apply ANOVA, several assumptions must be met:

1. **Random Sampling**: The data should be collected through a process that ensures each member of the population has an equal chance of being selected.

2. **Normal Distribution**: The residuals (the differences between observed and predicted values) should be normally distributed. This assumption can often be relaxed with larger sample sizes due to the Central Limit Theorem.

Assumptions contd.

3.Equal Variances: Also known as homogeneity of variances, this assumption states that the variance among the groups should be approximately equal.

4. Independence: Observations must be independent; that is, the value of one observation does not influence another.

Components of ANOVA

- In conducting an ANOVA test, we typically define:
- Let K represent the number of groups (or factor levels).
- Let n denote the number of replicates (sets of identical observations) within each group.
- Let yij represent the j-th observation within factor level i.
- The analysis involves calculating two types of variances:

1. **Between-group variance**: This measures how much group means differ from the overall mean.
2. **Within-group variance**: This measures how much individual observations within each group differ from their respective group mean.

Types of ANOVA

- There are several types of ANOVA tests depending on experimental design:

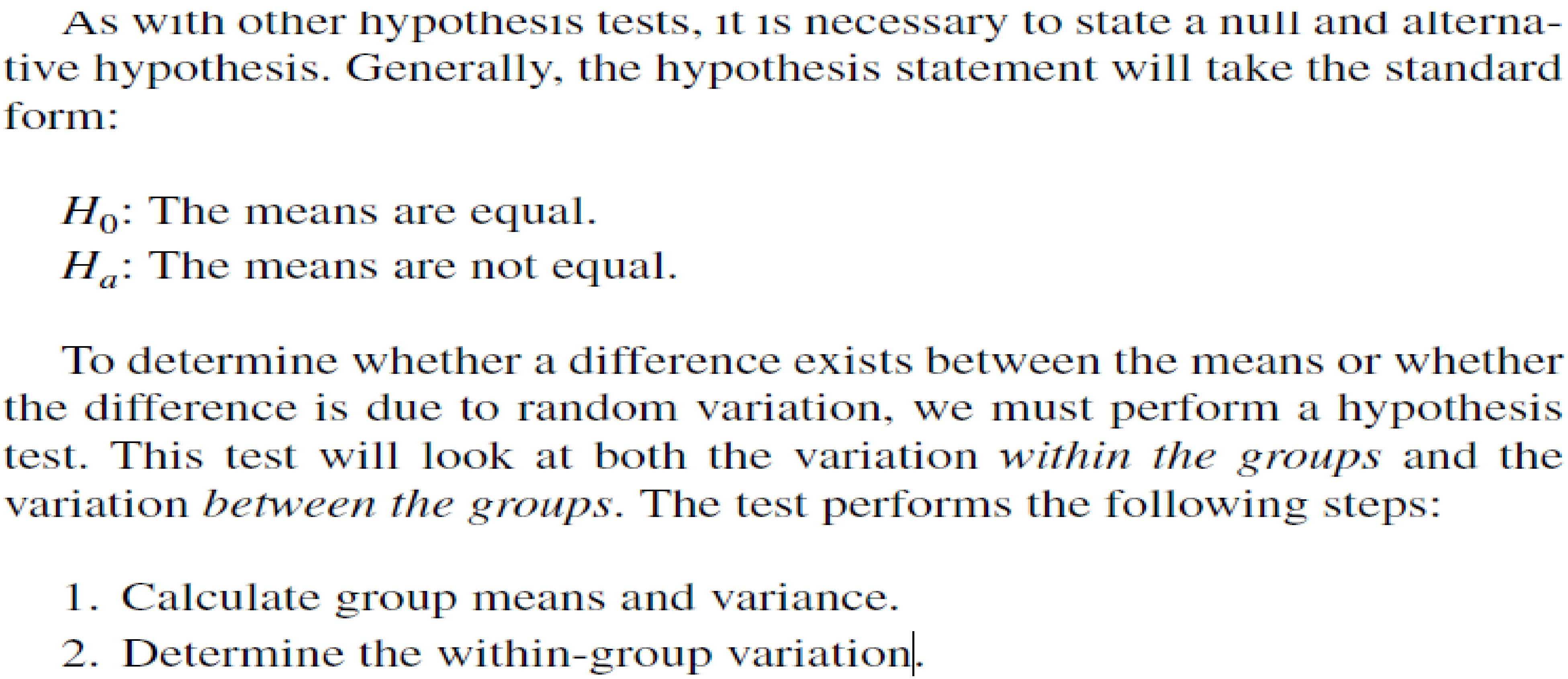
1. **One-way ANOVA**: Used when comparing means across a single factor with multiple levels.
2. **Two-way ANOVA**: Used when examining two factors simultaneously and their interaction effects.
3. **MANOVA (Multivariate Analysis of Variance)**: An extension that allows for multiple dependent variables.

- Each type serves specific research questions and designs, allowing researchers to analyze complex datasets effectively.

ANOVA

- A technique called the completely randomized *one-way analysis of variance* that compares the means of three or more different groups is reviewed.
- The test determines whether there is a difference between the groups. This method can be applied to cases where the groups are independent and random, the distributions are normal and the populations have similar variances.
- For example, an online computer retail company has call centers in four different locations.
- These call centers are approximately the same size and handle a certain number of calls each day.
- An analysis of the different call centers based on the average number of calls processed each day is required to understand whether one or more of the call centers are under- or over-performing.

ANOVA



As with other hypothesis tests, it is necessary to state a null and alternative hypothesis. Generally, the hypothesis statement will take the standard form:

H_0: The means are equal.
H_a: The means are not equal.

To determine whether a difference exists between the means or whether the difference is due to random variation, we must perform a hypothesis test. This test will look at both the variation *within the groups* and the variation *between the groups*. The test performs the following steps:

1. Calculate group means and variance.
2. Determine the within-group variation.

ANOVA

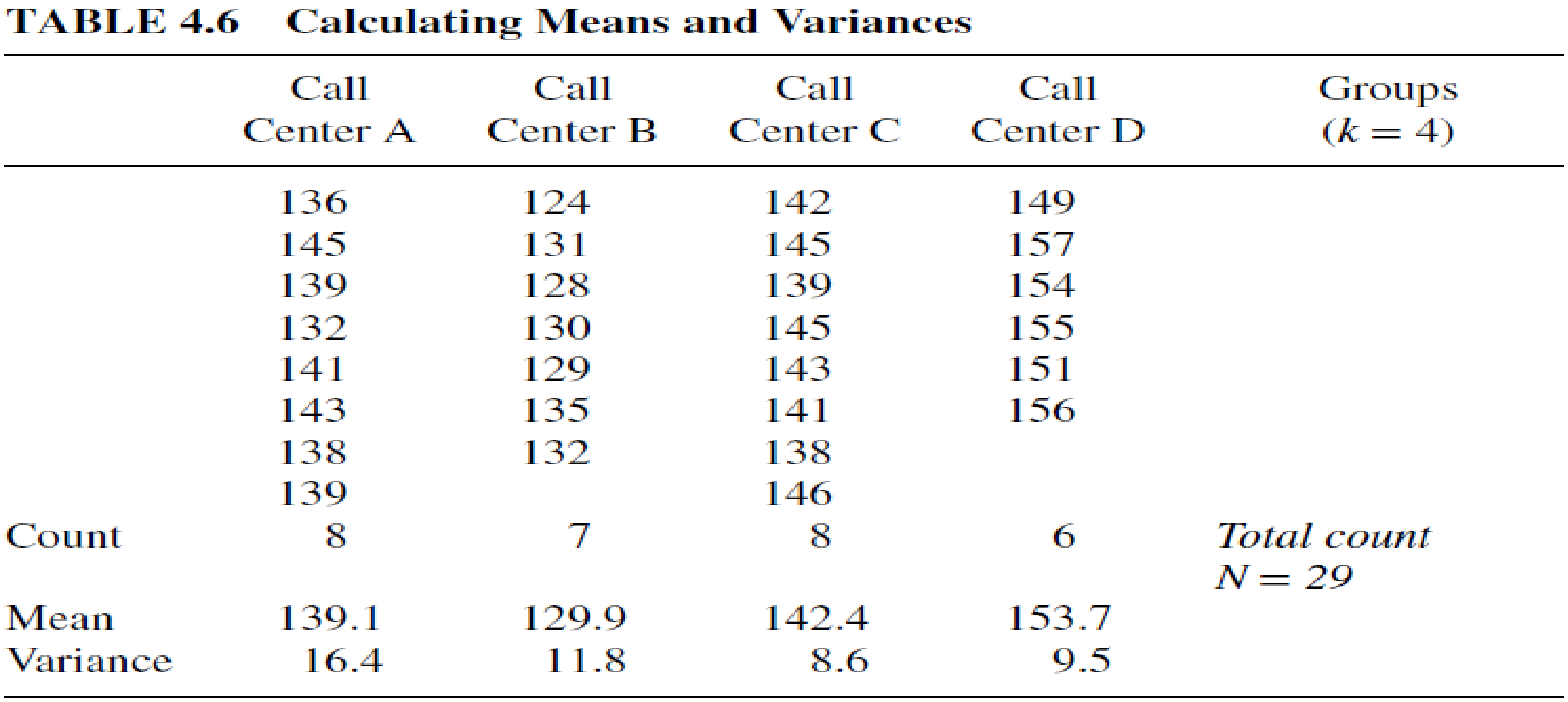


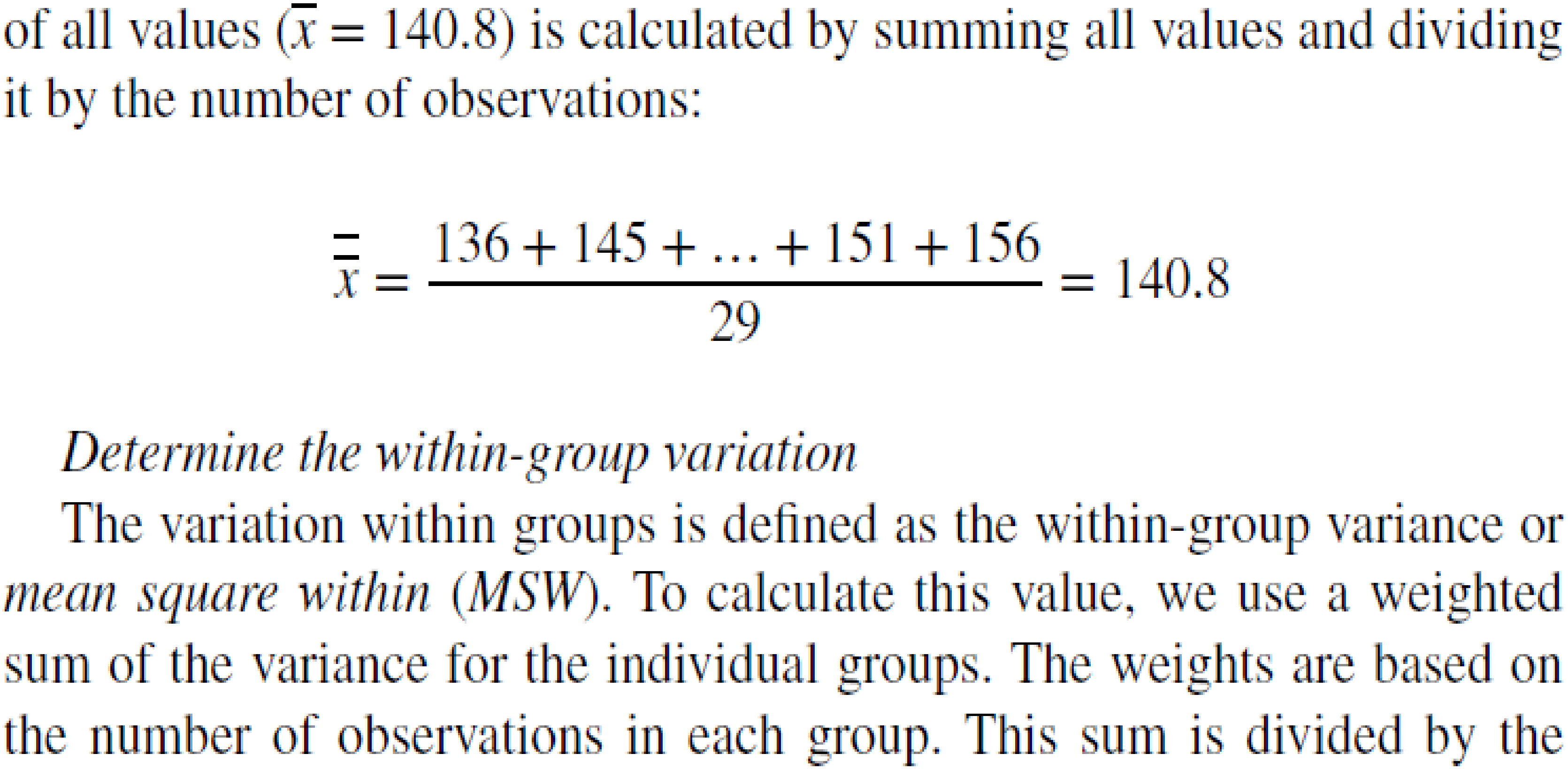
TABLE 4.6 Calculating Means and Variances

	Call Center A	Call Center B	Call Center C	Call Center D	Groups ($k = 4$)
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		
Count	8	7	8	6	*Total count* *$N = 29$*
Mean	139.1	129.9	142.4	153.7	
Variance	16.4	11.8	8.6	9.5	

ANOVA

- 3. Determine the between-group variation.
- 4. Determine the *F*-statistic, which is based on the between-group and within group ratio.
- 5. Test the significance of the *F*-statistic.
- The following sections describe these steps in detail:
- *Calculate group means and variances*
- In Table 4.6, for each call center a count along with the mean and variance has been calculated.
- In addition, the total number of groups ($k = 4$) and the total number of observations ($N = 29$) is listed. An average

ANOVA



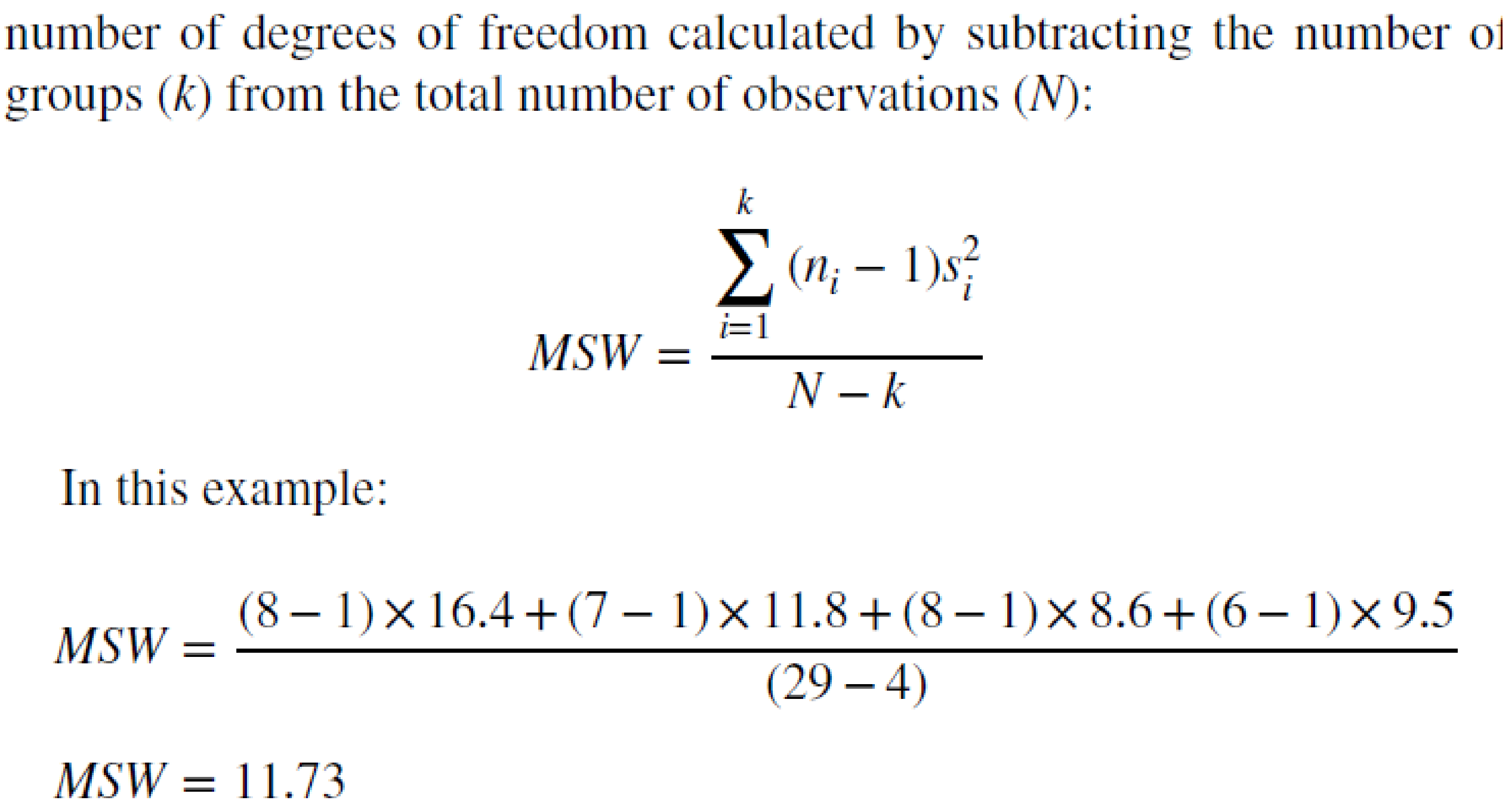
of all values ($\bar{x} = 140.8$) is calculated by summing all values and dividing it by the number of observations:

$$\bar{\bar{x}} = \frac{136 + 145 + \ldots + 151 + 156}{29} = 140.8$$

Determine the within-group variation

The variation within groups is defined as the within-group variance or *mean square within* (*MSW*). To calculate this value, we use a weighted sum of the variance for the individual groups. The weights are based on the number of observations in each group. This sum is divided by the

ANOVA



number of degrees of freedom calculated by subtracting the number of groups (k) from the total number of observations (N):

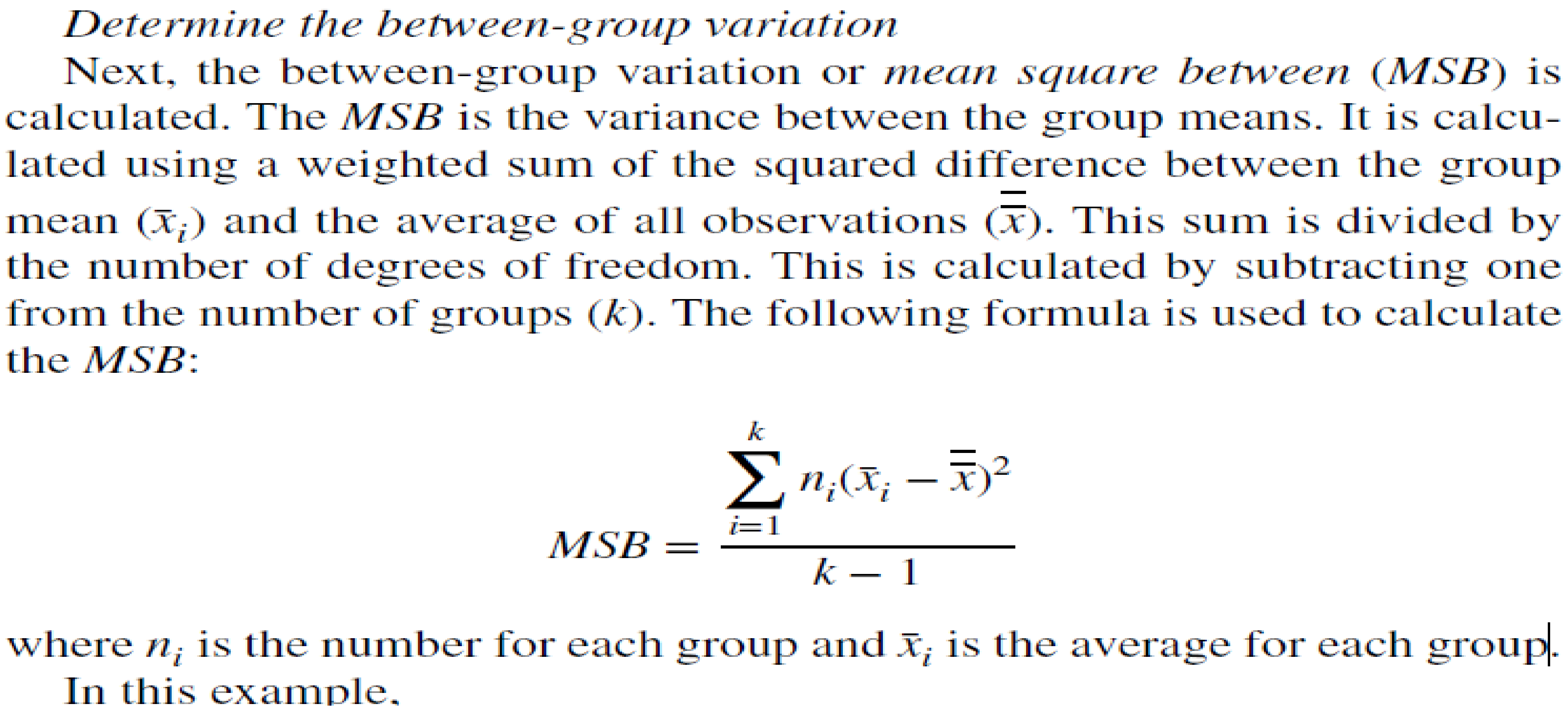
$$MSW = \frac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{N - k}$$

In this example:

$$MSW = \frac{(8-1)\times 16.4 + (7-1)\times 11.8 + (8-1)\times 8.6 + (6-1)\times 9.5}{(29-4)}$$

$$MSW = 11.73$$

ANOVA



Determine the between-group variation

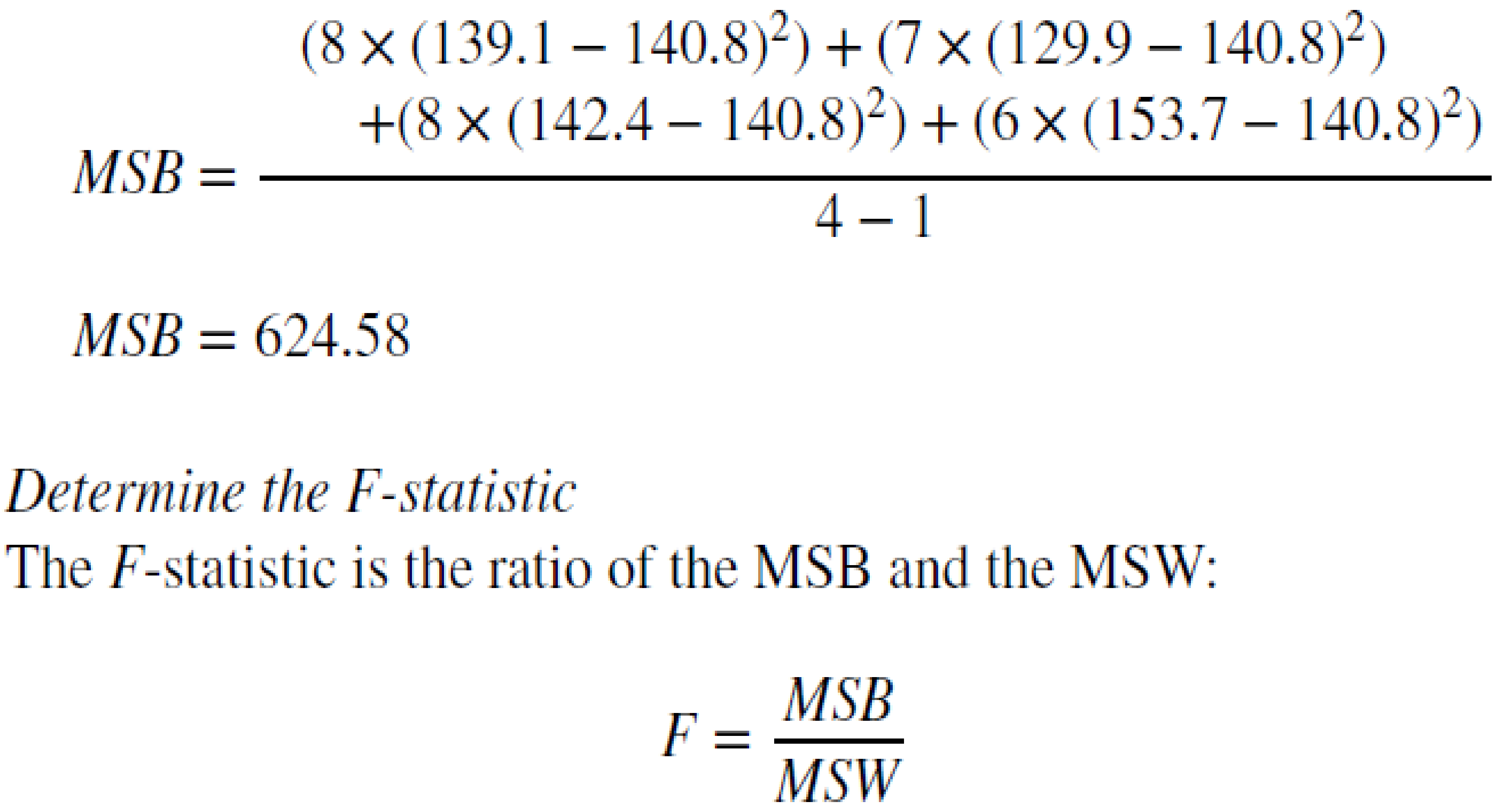
Next, the between-group variation or *mean square between* (*MSB*) is calculated. The *MSB* is the variance between the group means. It is calculated using a weighted sum of the squared difference between the group mean ($\bar{x}_i$) and the average of all observations ($\bar{\bar{x}}$). This sum is divided by the number of degrees of freedom. This is calculated by subtracting one from the number of groups (k). The following formula is used to calculate the *MSB*:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k - 1}$$

where n_i is the number for each group and $\bar{x}_i$ is the average for each group.

In this example,

ANOVA



$$MSB = \frac{(8 \times (139.1 - 140.8)^2) + (7 \times (129.9 - 140.8)^2) + (8 \times (142.4 - 140.8)^2) + (6 \times (153.7 - 140.8)^2)}{4 - 1}$$

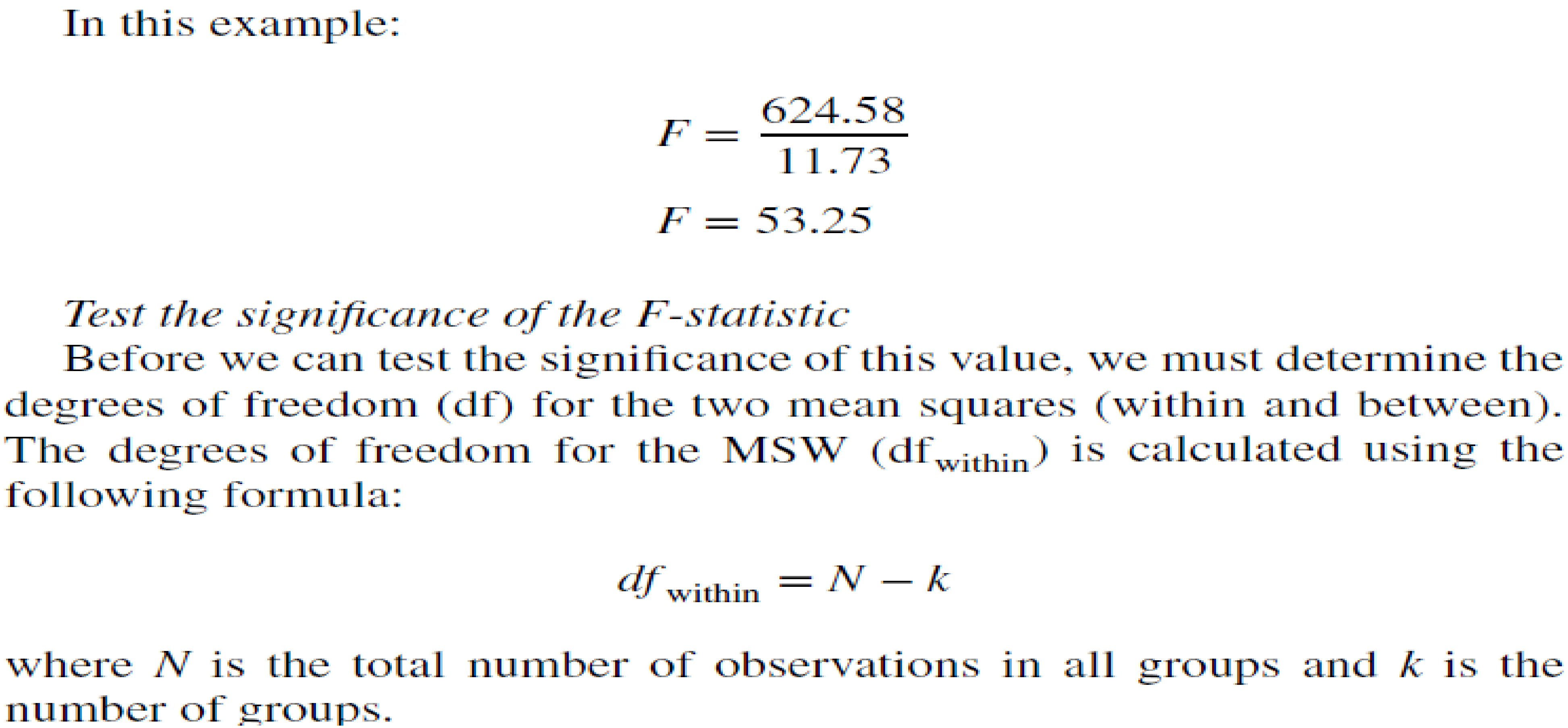
$$MSB = 624.58$$

Determine the F-statistic

The *F*-statistic is the ratio of the MSB and the MSW:

$$F = \frac{MSB}{MSW}$$

ANOVA



In this example:

$$F = \frac{624.58}{11.73}$$

$$F = 53.25$$

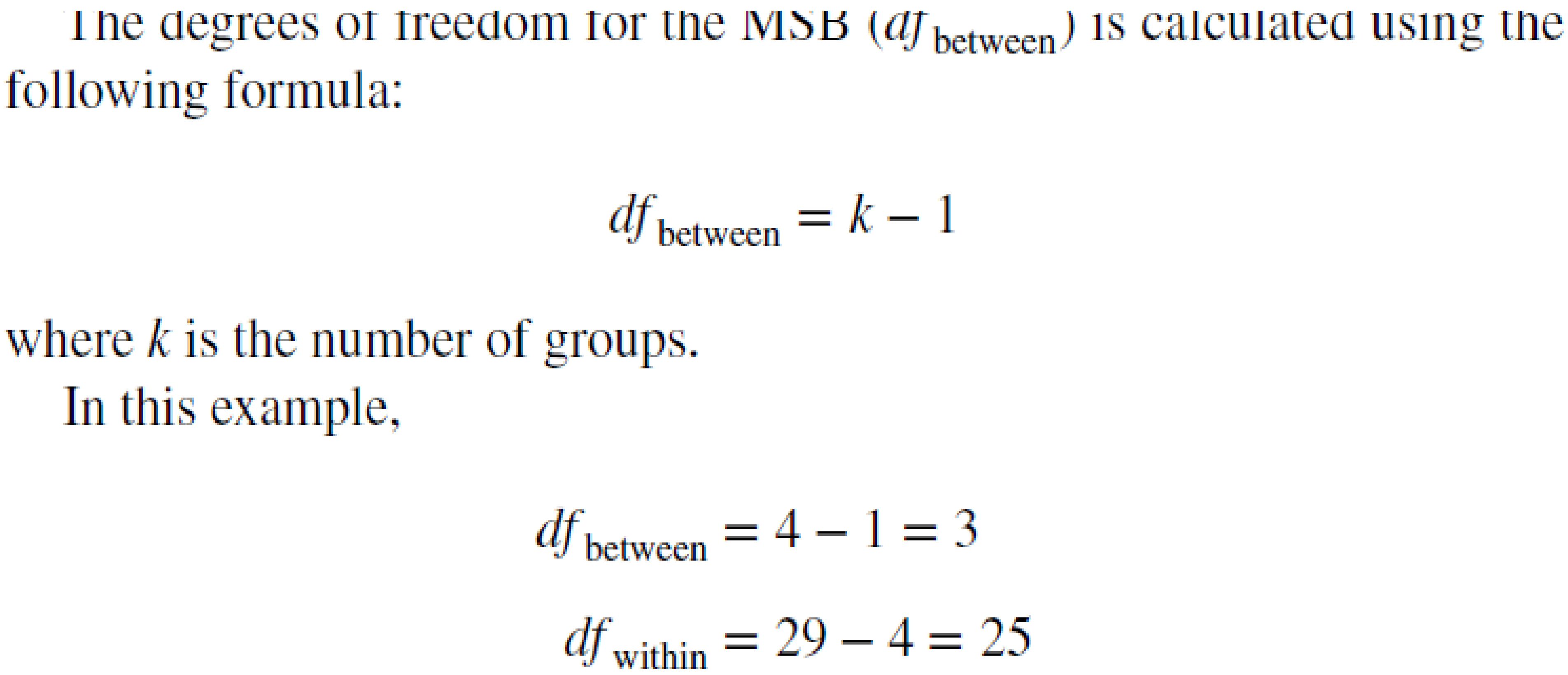
Test the significance of the F-statistic

Before we can test the significance of this value, we must determine the degrees of freedom (df) for the two mean squares (within and between). The degrees of freedom for the MSW ($\mathrm{df}_{\mathrm{within}}$) is calculated using the following formula:

$$df_{\mathrm{within}} = N - k$$

where N is the total number of observations in all groups and k is the number of groups.

ANOVA



The degrees of freedom for the MSB (df_{between}) is calculated using the following formula:

$$df_{\text{between}} = k - 1$$

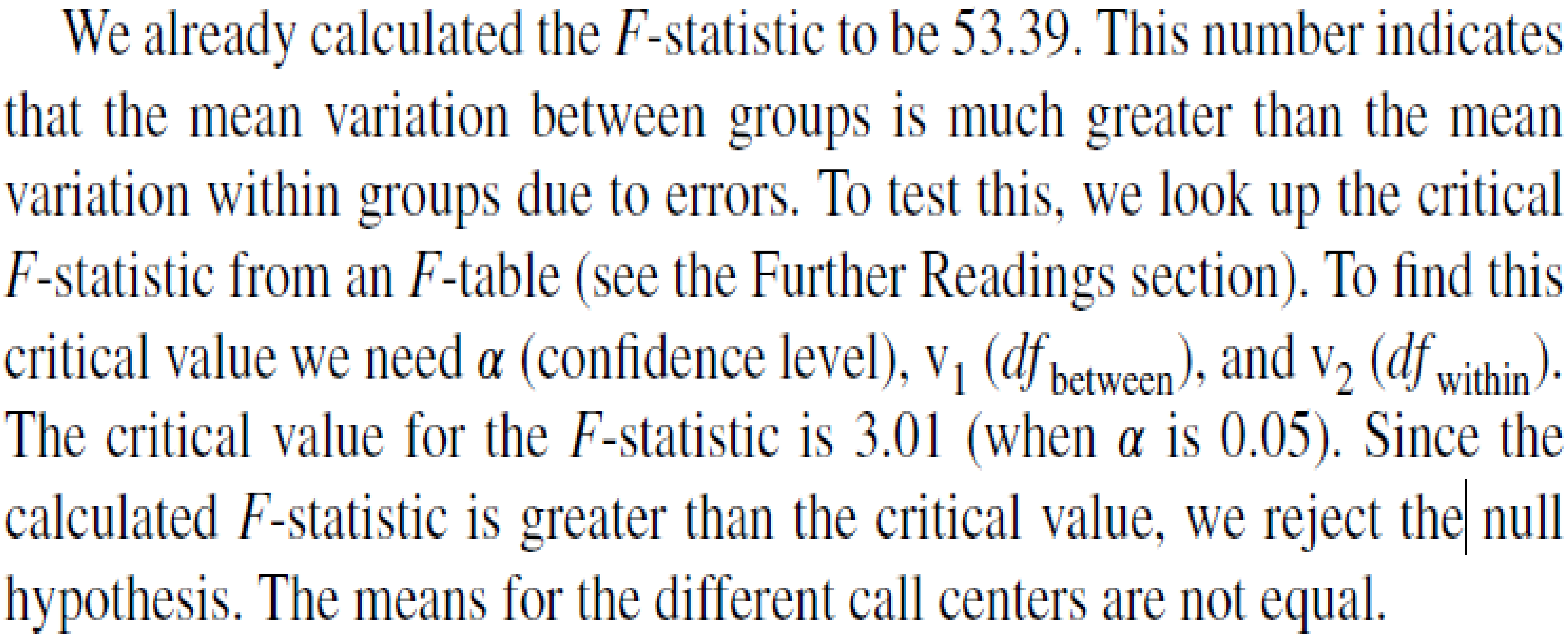
where k is the number of groups.

In this example,

$$df_{\text{between}} = 4 - 1 = 3$$

$$df_{\text{within}} = 29 - 4 = 25$$

ANOVA



We already calculated the *F*-statistic to be 53.39. This number indicates that the mean variation between groups is much greater than the mean variation within groups due to errors. To test this, we look up the critical *F*-statistic from an *F*-table (see the Further Readings section). To find this critical value we need α (confidence level), v_1 (df_{between}), and v_2 (df_{within}). The critical value for the *F*-statistic is 3.01 (when α is 0.05). Since the calculated *F*-statistic is greater than the critical value, we reject the null hypothesis. The means for the different call centers are not equal.

THANK YOU