Applied Statistics and Data Analysis
Analysis of Variance I

Joaquin Ortega

Fall 2020

Analysis of Variance

Introduction

The purpose of statistical experiments is to explore the effect that several **levels** of one or more categorical variables, known as **factors**, have on the outcome of the experiment.

A factor may be the type of fertilizer for a crop, a specific treatment for a patient suffering from a given disease, the gender or race of a subject, the type of surface over which a race is run, or the diet of an animal in an experiment.

Factors are frequently called **treatments**.

Analysis of Variance

The analysis of variance (Anova) is particularly important in statistically designed experiments, where, depending on the characteristics of the process being tested, a scheme is adopted, which determines how the experiment must be carried out and how the results should be analyzed.

Anova was initially proposed by Ronald Fisher nearly a century ago in the context of experiments in Agriculture.

Analysis of Variance

The main idea of Anova is to compare several means, which are related to the levels of the different factors.

This will allow us to differentiate the effects of several factors being varied together in a single experiment.

However, to do this, we need to analyze the variances associated with the different means.

One-way Analysis of Variance

One-way Analysis of Variance

Let us start with a simple example to introduce the main ideas. This example is elaborated from *The R Book, 2nd. Edition*, M.J. Crawley, Wiley 2013.

Suppose we have only one factor and two levels, L_1 and L_2, and we have repeated the experiment five times for each level. The measurements are:

$$y_{11}, \ldots, y_{15}, \quad y_{21}, \ldots, y_{25}.$$

In general, if we have n_1 measurements for the first level and n_2 for the second, we would have

$$y_{11}, \ldots, y_{1n_1}, \quad y_{21}, \ldots, y_{2n_2}.$$

Notation

An index replaced by a $\bullet$ indicates that we sum the values corresponding to that index, leaving all other indices fixed:

$$y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}, \qquad y_{\bullet j} = \sum_{i=1}^{2} y_{ij}, \qquad y_{\bullet\bullet} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}.$$

A bar over the variable indicates that we divide by the number of terms that are being added:

$$\bar{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \qquad \bar{y}_{\bullet j} = \frac{1}{2} \sum_{i=1}^{2} y_{ij}, \qquad \bar{y}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij},$$

where $n = n_1 + n_2$.

This notation can be used with any number of levels.

One-way Analysis of Variance

Figure 1 shows the ten values obtained in the experiment.

The horizontal line corresponds to the overall mean $\bar{y}_{\bullet\bullet}$ for the Response.

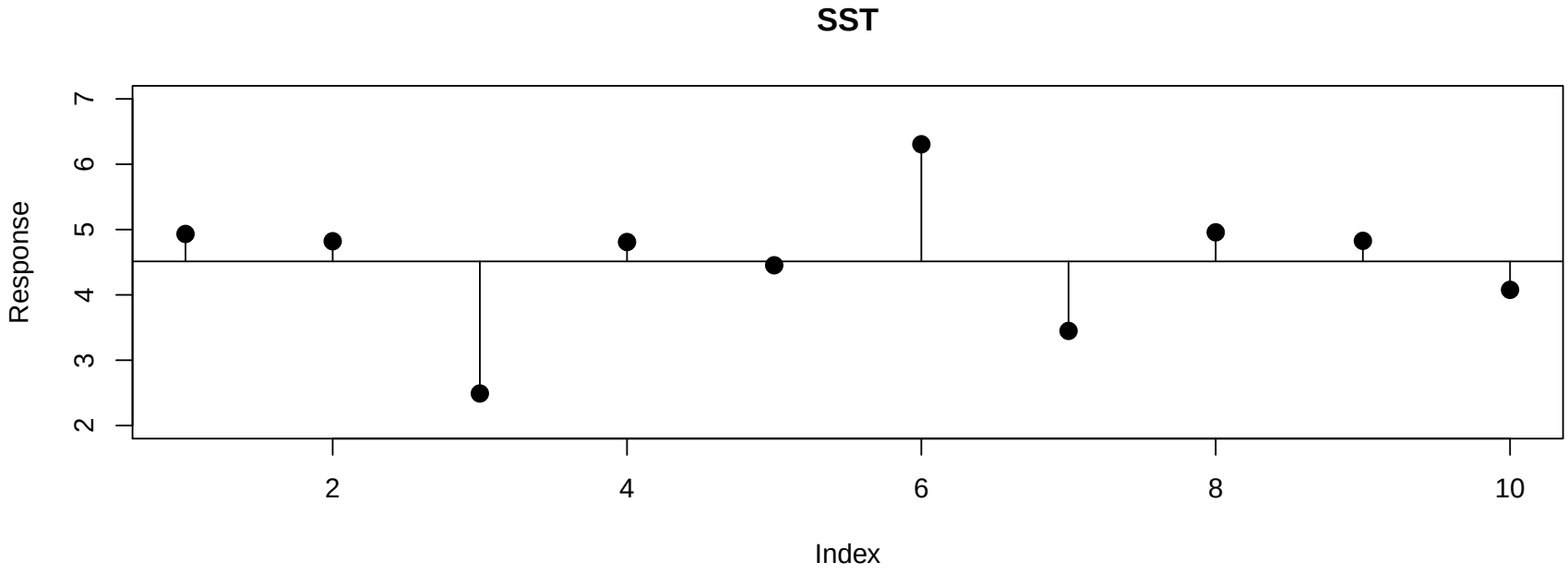


Figure 1: Outcomes of ten replications of the experiment.

One-way Analysis of Variance

The vertical segments are the differences between the observed values and the average,

$$y_{ij} - \bar{y}_{\bullet\bullet},$$

for $i = 1, 2, j = 1, \ldots, 5$.

The **total sum of squares** SST is defined as

$$SST = \sum_{i=1}^{2} \sum_{j=1}^{5} (y_{ij} - \bar{y}_{\bullet\bullet})^2.$$

One-way Analysis of Variance

If we color the points in the graph according to the level of the factor, we get

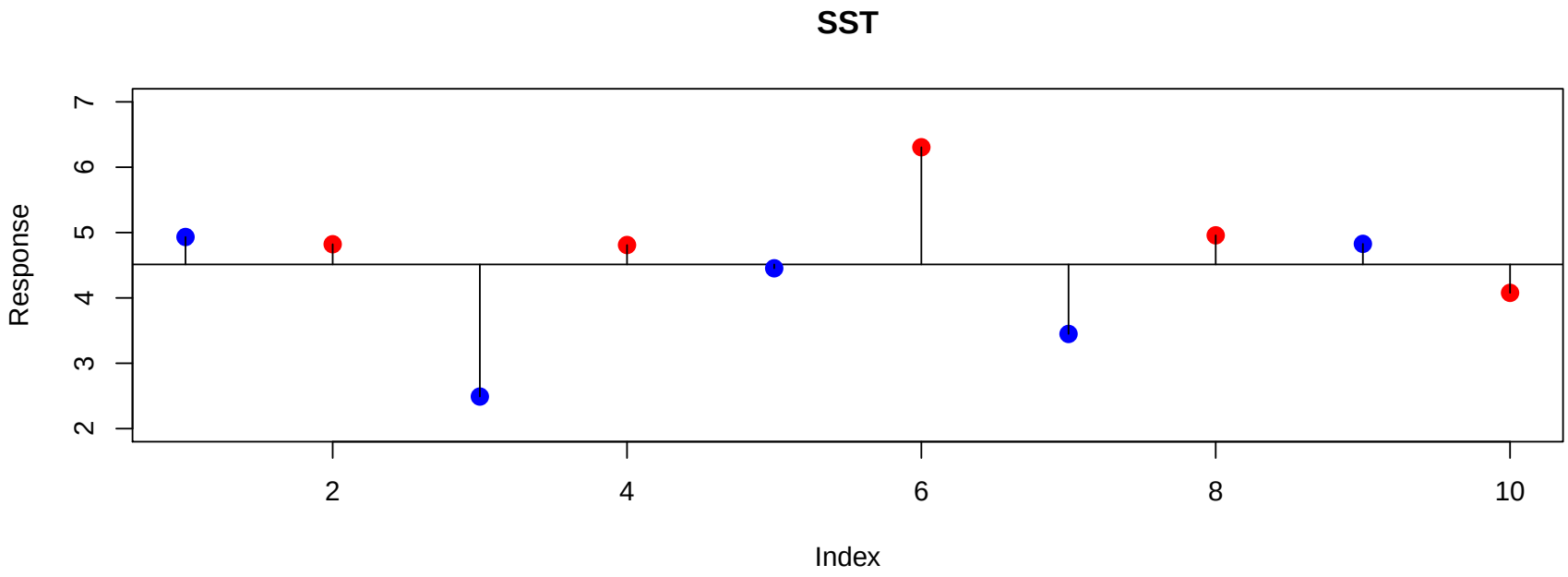


Figure 2: Total sum of squares. Colors correspond to the two treatment levels.

One-way Analysis of Variance

We see that the blue points, which correspond to L_1, tend to be lower than the red points.

What we would like to know is whether this is evidence of a real difference between the effect of the two levels or is simply due to chance.

Anova will help us decide which case it is.

One-way Analysis of Variance

Next, let us consider the data for each level and calculate a separate mean, $\bar{y}_{i\bullet}$ for the values corresponding to $L_i, i = 1, 2$. The graph is

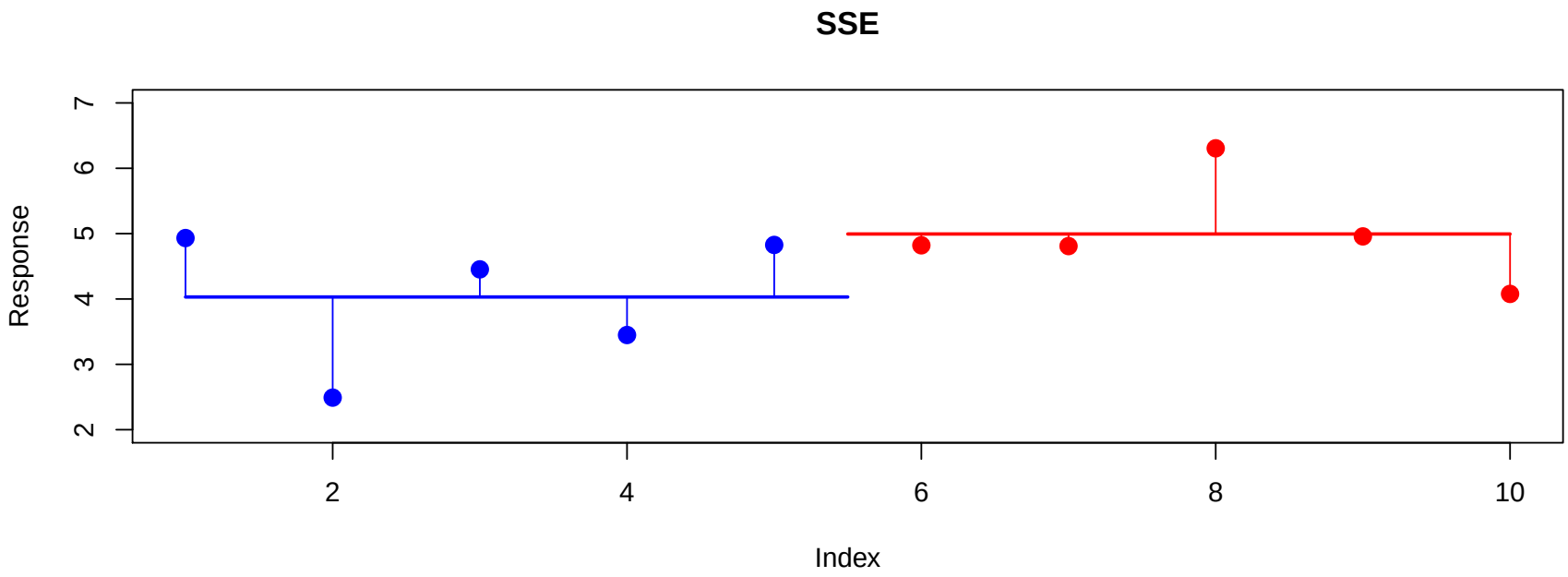


Figure 3: Error sum of squares.

One-way Analysis of Variance

The sum of the squares of the differences between the observed values and the corresponding treatment mean is known as the **error sum of squares**

$$SSE = \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\bullet})^2.$$

SSE is the sum of the squares with respect to the red and blue lines in figure 3.

One-way Analysis of Variance

If the treatment we are considering has no effect on the outcome, we would expect the red and blue lines to be equal and also to coincide with the overall mean represented in the first figure by a horizontal black line.

This is the same as saying that $SST = SSE$.

If the means for the two treatment groups were the same, we would get a graph like figure 4.

One-way Analysis of Variance

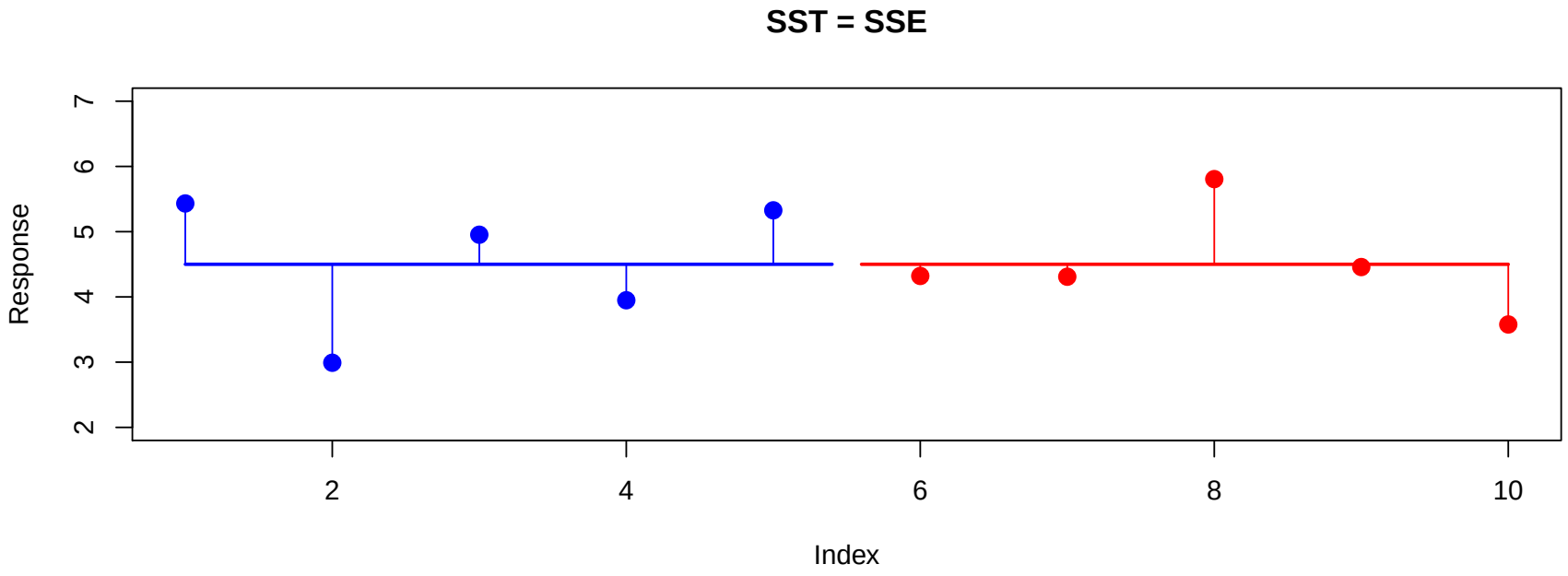


Figure 4: Total and error sum of squares.

One-way Analysis of Variance

On the other hand, if the treatments have an effect on the response, we would expect the red and blue lines to be different and SSE to be less than SST.

In fact, SSE could be zero if all the values were equal to their treatment means, as in figure 5(r).

We see that the difference in the sums of squares SST and SSE is linked to the differences in the treatment means.

One-way Analysis of Variance

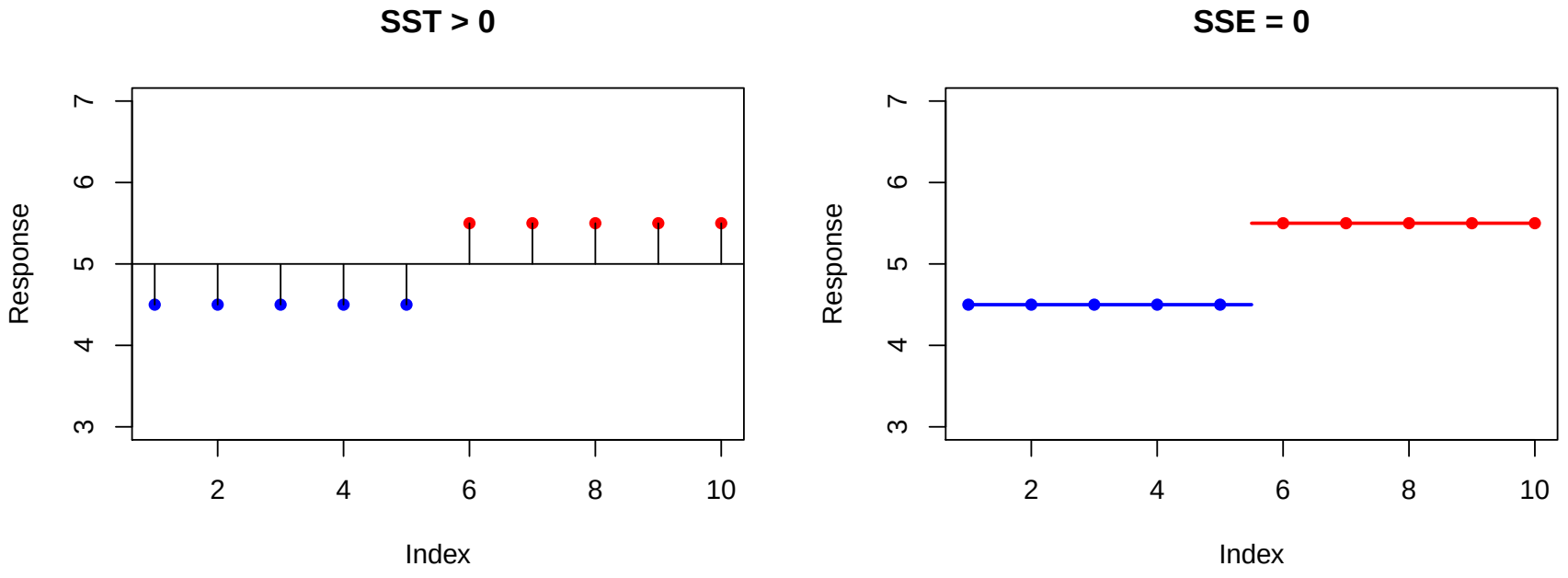


Figure 5: Total and error sum of squares.

Sums of Squares

One-way Analysis of Variance

To see that $SSE \leq SST$ always, start with the definition of SST:

$$\begin{aligned} SST &= \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet} + \bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\bullet} + \bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet})^2 \end{aligned} \tag{1}$$

One-way Analysis of Variance

Let us look at the first sum; the second one is similar.

$$\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet} + \bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2 = \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n_1}(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2 + \sum_{j=1}^{n_1} 2(y_{1j} - \bar{y}_{1\bullet})(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})$$

The terms in the second sum above do not depend on the index of summation j, and the sum is equal to

$$n_1(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2.$$

One-way Analysis of Variance

For the third term in the sum we have

$$\sum_{j=1}^{n_1} 2(y_{1j} - \bar{y}_{1\bullet})(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet}) = 2(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})\Big(\sum_{j=1}^{n_1} y_{1j} - n_1\bar{y}_{1\bullet}\Big) = 0$$

because $\sum_{1j}^{n_1} y_i = n_1\bar{y}_{1\bullet}$. A similar argument is true for the second sum in (1) and we get

One-way Analysis of Variance

$$
\begin{aligned}
SST &= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet})^2 + n_1(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})^2 \\
&\quad + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\bullet})^2 + n_2(\bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1\bullet})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2\bullet})^2 + \sum_{i=1}^{2} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \\
&= SSE + \sum_{i=1}^{2} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2. \qquad (2)
\end{aligned}
$$

Since the second sum in (2) is positive, we see that $SSE \leq SST$.

One-way Analysis of Variance

The difference between SST and SSE is known as the **treatment sum of squares** and will be denoted by SSA.

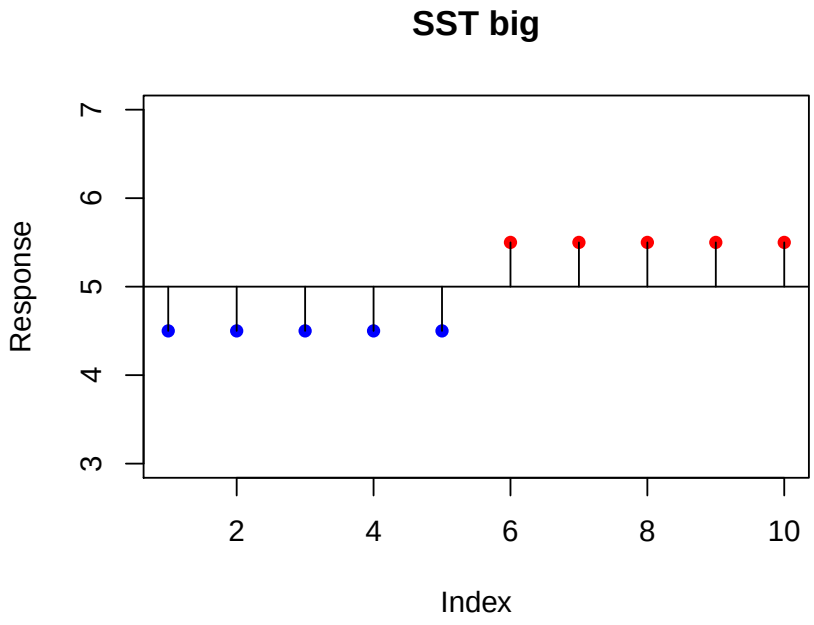
$$SSA = \sum_{i=1}^{2} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2.$$

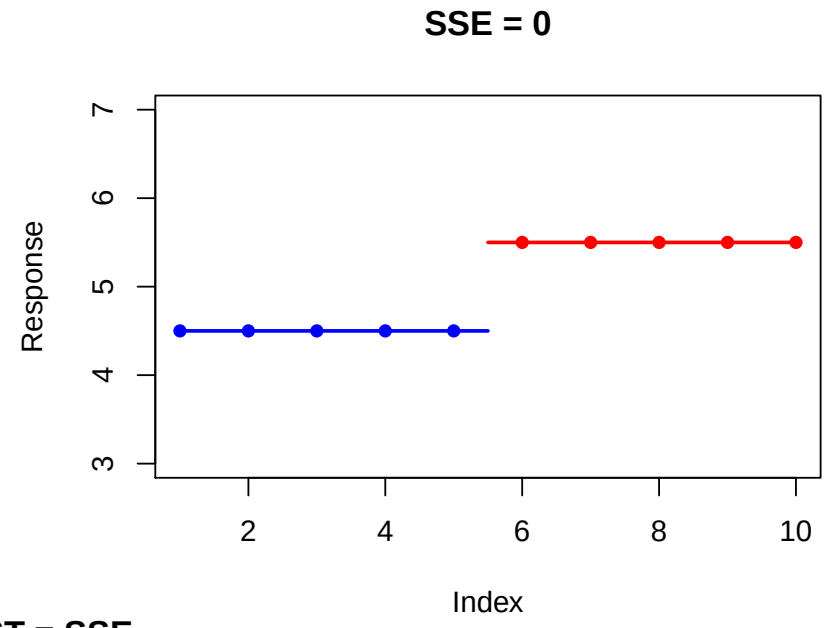
When differences in mean between treatments are significant, SSA will be big with respect to SSE. If, on the contrary, the means are similar, then SSA will be small with respect to SSE.

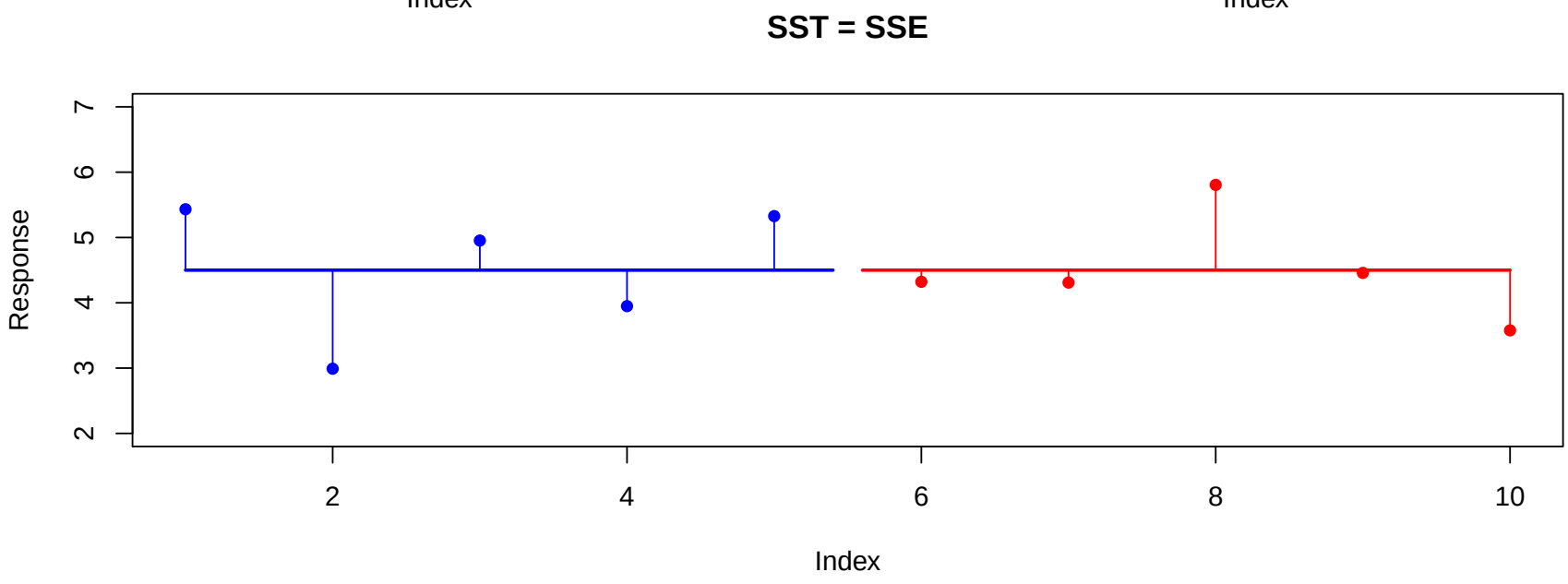
Figure 6 presents extreme cases of these situations. On the top, all the variation in the response is explained by the difference in the treatment means, and $SSE = 0$, so $SST = SSA$.

On the bottom, the treatment means are equal and $SST = SSE$, so $SSA = 0$, and the treatment does not affect the response.

One-way Analysis of Variance







One-way Analysis of Variance

Observe that *SSE* represents the variability *within* each group or level of the treatment factor, while *SSA* represents the variability *between* different groups or factor levels.

This equation can be seen as a decomposition of the observed variability in the sample into the variability within each factor level and the difference between the factors.

As we will see later, to estimate variances, we divide sums of squares by the corresponding degrees of freedom (d.f.).

In our example, we have two levels for the treatment, and we lose one degree of freedom estimating the overall mean, so there are $2 - 1 = 1$ degrees of freedom for the treatments.

One-way Analysis of Variance

In general, if we had k treatment levels, we would have $k-1$ degrees of freedom for treatments.

If each factor level were replicated r times, then there would be $r-1$ degrees of freedom for each level, since we lose one for each treatment mean. Considering that there are k levels, there are $k(r-1)$ d.f. for error in the whole experiment.

Finally, the total number of data points in the experiment is $n = rk$, and we lose one for the overall mean, so there are in total $rk-1$ d.f. As a verification, it is always useful to check that the degrees of freedom for the components add up to the correct total:

$$k - 1 + k(r-1) = k - 1 + kr - k = kr - 1.$$

One-way Analysis of Variance

The following expressions give the empirical variance for treatment, MSA, and the errors MSE.

$$MSE = \frac{SSE}{k(r-1)}; \qquad MSA = \frac{SSA}{k-1}.$$

One-way Analysis of Variance

The usual way to sum up these results is through an Analysis of Variance (Anova) table.

Table 1: Anova table for example 1.

Source	SS	d.f.	MS	F_{obs}	Critical F
Treatment	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	`qf(1-`α`, k-1, k(r-1))`
Error	SSE	$k(r-1)$	$MSE = \frac{SSE}{k(r-1)}$		
Total	SST	$kr-1$			

One-way Analysis of Variance

The significance of the difference between the means for the different treatment levels is assessed using an F test, which we will describe later on in detail.

The null hypothesis is that the means are equal versus the alternative that at least two of them are different.

One-way Analysis of Variance

Another way of thinking about the comparison we made is in terms of the relative amounts of sampling variability between replicates that correspond to the same treatment level and between different levels.

If the variation between replicates of a given level is large when compared to the variability between treatments, we will probably conclude that the difference between treatments is not significant.

On the other hand, if the variability within treatment levels is small compared to the differences between treatments, we will likely reject the null hypothesis of equal treatment means.