Chapter 13

Analysis of Variance and Experimental Design

In this chapter we introduce a statistical procedure called *analysis of variance* (ANOVA).

First, we show how ANOVA can be used to test for the equality of three or more population means using data obtained from an observational study. Then, we discuss the use of ANOVA for analyzing data obtained from three types of experimental studies: a completely randomized design, a randomized block design and a factorial experiment. In the following chapters we will see that ANOVA plays a key role in analyzing the results of regression analysis involving both experimental and observational data.

13.1 An introduction to analysis of variance

National Computer Products (NCP) manufactures printers and fax machines at plants located in Ayr, Dusseldorf and Stockholm. To measure how much employees at these plants know about total quality management, a random sample of six employees was selected from each plant and given a quality awareness examination. The examination scores obtained for these 18 employees are listed in Table 13.1. The sample means, sample variances and sample standard deviations for each group are also provided. Managers want to use these data to test the hypothesis that the mean examination score is the same for all three plants.

We will define population 1 as all employees at the Ayr plant, population 2 as all employees at the Dusseldorf plant, and population 3 as all employees at the Stockholm plant. Let

μ_1 = mean examination score for population 1
μ_2 = mean examination score for population 2
μ_3 = mean examination score for population 3

Although we will never know the actual values of μ_1, μ_2 and μ_3, we want to use the sample results to test the following hypotheses.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_1: \text{Not all population means are equal}$$

Table 13.1 Examination scores for 18 employees

Observation	Plant 1 Ayr	Plant 2 Dusseldorf	Plant 3 Stockholm
1	85	71	59
2	75	75	64
3	82	73	62
4	76	74	69
5	71	69	75
6	85	82	67
Sample mean	79	74	66
Sample variance	34	20	32
Sample standard deviation	5.83	4.47	5.66

The two variables in the NCP example are plant location and score on the quality awareness examination. Because the objective is to determine whether the mean examination score is the same for plants located in Ayr, Dusseldorf and Stockholm, examination score is referred to as the dependent or *response variable* and plant location as the independent variable or *factor*. In general, the values of a factor selected for investigation are referred to as levels of the factor or *treatments*. Thus, in the NCP example the three treatments are Ayr, Dusseldorf and Stockholm. These three treatments define the populations of interest in the NCP example. For each treatment or population, the response variable is the examination score.

Assumptions for analysis of variance

Three assumptions are required to use analysis of variance.

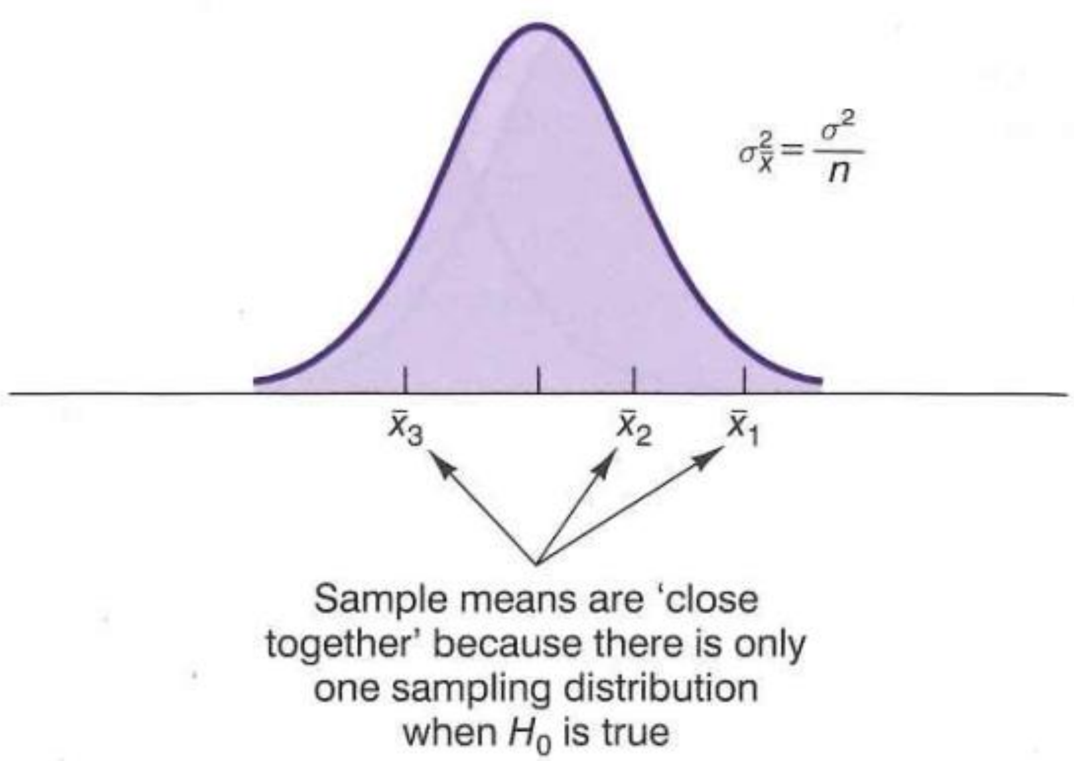
1. **For each population, the response variable is normally distributed**. Implication: In the NCP example, the examination scores (response variable) must be normally distributed at each plant.
2. **The variance of the response variable, denoted σ^2, is the same for all of the populations**. Implication: In the NCP example, the variance of examination scores must be the same for all three plants.
3. **The observations must be independent**. Implication: In the NCP example, the examination score for each employee must be independent of the examination score for any other employee.

A conceptual overview

If the means for the three populations are equal, we would expect the three sample means to be close together. In fact, the closer the three sample means are to one another, the more evidence we have for the conclusion that the population means are equal. Alternatively, the more the sample means differ, the more evidence we have for the conclusion that the population means are not equal. In other words, if the variability among the sample means is 'small', it supports H_0; if the variability among the sample means is 'large', it supports H_1.

If the null hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3$, is true, we can use the variability among the sample means to develop an estimate of σ^2. First, note that if the assumptions for analysis of variance are satisfied, each sample will have come from the same normal distribution with mean μ and variance σ^2. Recall from Chapter 7 that the sampling distribution of the sample mean for a simple random sample of size n from a normal population will be normally distributed with mean μ and variance σ^2/n. Figure 13.1 illustrates such a sampling distribution.

Figure 13.1 Sampling distribution of $\overline{X}$ given H_0 is true



Therefore, if the null hypothesis is true, we can think of each of the three sample means, $\bar{x}_1 = 79$, $\bar{x}_2 = 74$, and $\bar{x}_3 = 66$, from Table 13.1 as values drawn at random from the sampling distribution shown in Figure 13.1. In this case, the mean and variance of the three values can be used to estimate the mean and variance of the sampling distribution. When the sample sizes are equal, as in the NCP example, the best estimate of the mean of the sampling distribution of $\bar{X}$ is the mean or average of the sample means. Thus, in the NCP example, an estimate of the mean of the sampling distribution of $\bar{X}$ is $(79 + 74 + 66)/3 = 73$. We refer to this estimate as the *overall sample mean.* An estimate of the variance of the sampling distribution of $\bar{X}$, $\sigma^2_{\bar{x}}$ is provided by the variance of the three sample means.

$$s^2_{\bar{x}} = \frac{(79-73)^2 + (74-73)^2 + (66-73)^2}{3-1} = \frac{86}{2} = 43$$

Because $\sigma_{\bar{x}}^2 = \sigma^2 / n$, solving for σ^2 gives

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

Hence,

$$\text{Estimate of } \sigma^2 = n\,(\text{Estimate of } \sigma_{\bar{x}}^2) = ns_{\bar{x}}^2 = 6(43) = 258$$

The result, $ns_{\bar{x}}^2 = 258$, is referred to as the *between-treatments* estimate of σ^2.

The between-treatments estimate of σ^2 is based on the assumption that the null hypothesis is true. In this case, each sample comes from the same population, and there is only one sampling distribution of $\overline{X}$. To illustrate what happens when H_0 is false, suppose the population means all differ. Note that because the three samples are from normal populations with different means, they will result in three different sampling distributions. Figure 13.2 shows that in this case, the sample means are not as close together as they were when H_0 was true. Thus, $s_{\bar{x}}^2$ will be larger, causing the between-treatments estimate of σ^2 to be larger. In general, when the population means are not equal, the between-treatments estimate will overestimate the population variance σ^2.

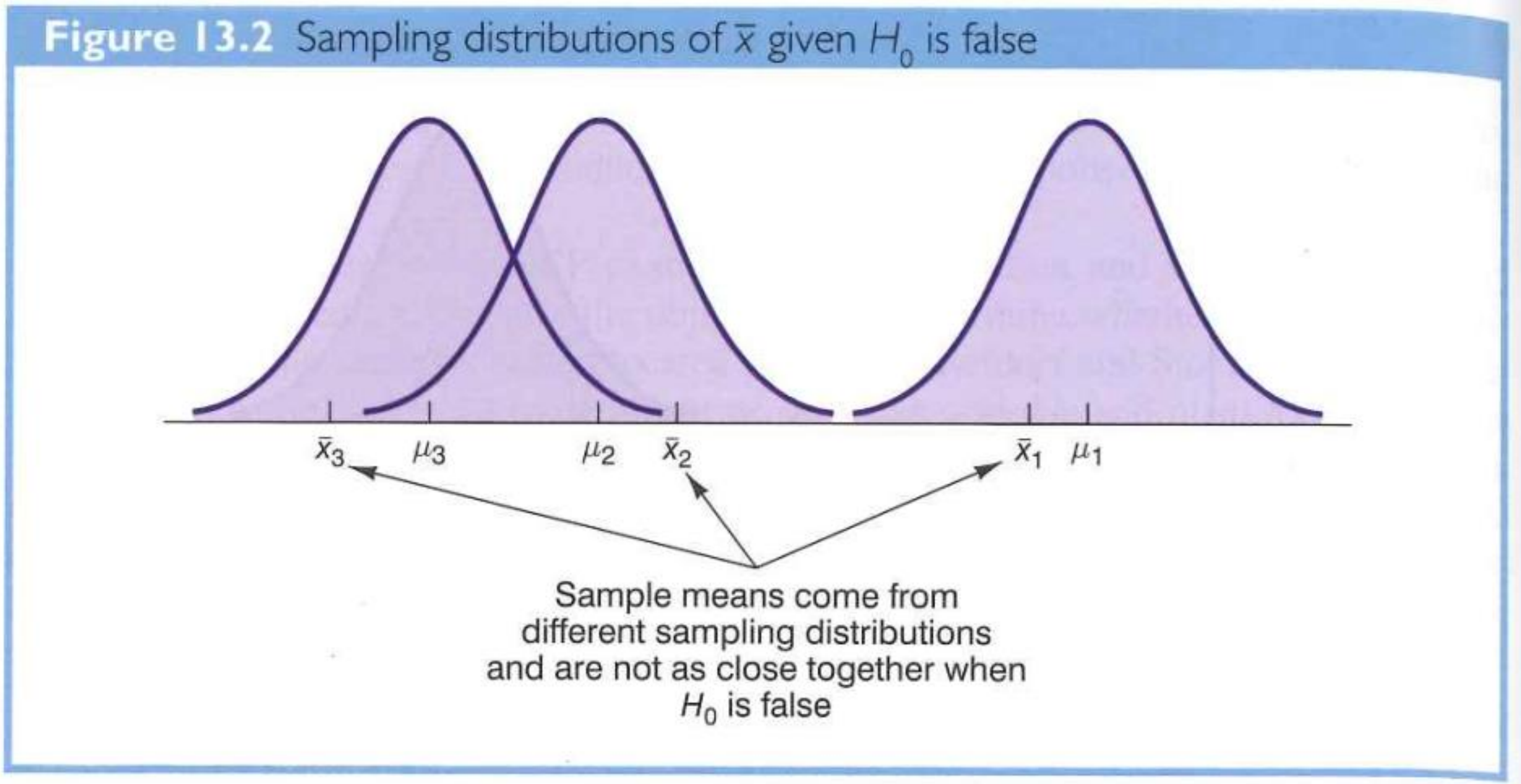


Figure 13.2 Sampling distributions of $\bar{x}$ given H_0 is false
$\bar{x}_3$
μ_3
μ_2
$\bar{x}_2$
$\bar{x}_1$
μ_1
Sample means come from different sampling distributions and are not as close together when H_0 is false

The variation within each of the samples also has an effect on the conclusion we reach in analysis of variance. When a simple random sample is selected from each population, each of the sample variances provides an unbiased estimate of σ^2. Hence, we can combine or pool the individual estimates of σ^2 into one overall estimate. The estimate of σ^2 obtained in this way is called the *pooled* or *within-treatments* estimate of σ^2. Because each sample variance provides an estimate of σ^2 based only on the variation within each sample, the within-treatments estimate of σ^2 is not affected by whether the population means are equal.

When the sample sizes are equal, the within-treatments estimate of σ^2 can be obtained by computing the average of the individual sample variances. For the NCP example we obtain

$$\text{Within-treatments estimate of } \sigma^2 = \frac{34 + 20 + 32}{3} = \frac{86}{3} = 28.67$$

In the NCP example, the between-treatments estimate of σ^2 (258) is much larger than the within-treatments estimate of σ^2 (28.67). In fact, the ratio of these two estimates is $258/28.67 = 9.00$. Recall, however, that the between-treatments approach provides a good estimate of σ^2 only if the null hypothesis is true; if the null hypothesis is false, the between treatments approach overestimates σ^2. The within-treatments approach provides a good estimate of σ^2 in either case. Thus, if the null hypothesis is true, the two estimates will be similar and their ratio will be close to 1. If the null hypothesis is false, the between-treatments estimate will be larger than the within-treatments estimate, and their ratio will be large. In the next section we will show how large this ratio must be to reject H_0.

In summary, the logic behind ANOVA is based on the development of two independent estimates of the common population variance σ^2. One estimate of σ^2 is based on the variability among the sample means themselves, and the other estimate of σ^2 is based on the variability of the data within each sample. By comparing these two estimates of σ^2, we will be able to determine whether the population means are equal.