Chapter 20

ANALYSIS OF VARIANCE

In Chapter 19, we examine how a quantitative independent variable x can be used for predicting the value of a quantitative dependent variable y. In this chapter we would like to examine whether one or more independent (or predictor) variable affects a dependent (or response) variable y. This chapter differs from the last chapter because the independent variable may now be either quantitative or qualitative. It also differs from the last chapter in assuming that the response measurements were obtained for specific settings of the independent variables. Selecting the settings of the independent variables is another aspect of experimental design. It enables us to tell whether changes in the independent variables cause changes in the mean response and it permits us to analyze the data using a method known as analysis of variance (or ANOVA). Sir Ronald Aylmer Fisher (1890-1962) developed the analysis of variance in 1920's and used it to analyze data from agricultural experiments.

The ANOVA investigates independent measurements from several treatments or levels of one or more than one factors (that is, the predictor variables). The technique of ANOVA consists of partitioning the total sum of squares into component sum of squares due to different factors and the error. For instance, suppose there are Q factors. Then the total sum of squares ($\mathrm{SS_T}$) is partitioned as

$$\mathrm{SS_T} = \mathrm{SS_A} + \mathrm{SS_B} + \cdots + \mathrm{SS_Q} + \mathrm{SS_{Error}},$$

where $\mathrm{SS_A}$, $\mathrm{SS_B}$, ..., and $\mathrm{SS_Q}$ represent the sum of squares associated with the factors A, B, ..., and Q, respectively. If the ANOVA involves only one factor, then it is called one-way analysis of variance. Similarly if it involves two factors, then it is called the two-way analysis of variance. If it involves more then two factors, then the corresponding ANOVA is called the higher order analysis of variance. In this chapter we only treat the one-way analysis of variance.

The analysis of variance is a special case of the linear models that represent the relationship between a continuous response variable y and one or more predictor variables (either continuous or categorical) in the form

$$y = X\,\beta + \epsilon \tag{1}$$

where y is an $m \times 1$ vector of observations of response variable, X is the $m \times n$ design matrix determined by the predictor variables, β is $n \times 1$ vector of parameters, and ϵ is an $m \times 1$ vector of random error (or disturbances) independent of each other and having distribution.

20.1. One-Way Analysis of Variance with Equal Sample Sizes

The standard model of one-way ANOVA is given by

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n, \tag{2}$$

where $m \geq 2$ and $n \geq 2$. In this model, we assume that each random variable

$$Y_{ij} \sim N(\mu_i, \sigma^2) \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n. \tag{3}$$

Note that because of (3), each ϵ_{ij} in model (2) is normally distributed with mean zero and variance σ^2.

Given m independent samples, each of size n, where the members of the i^{th} sample, $Y_{i1}, Y_{i2}, ..., Y_{in}$, are normal random variables with mean μ_i and unknown variance σ^2. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n.$$

We will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a} : \text{not all the means are equal.}$$

In the following theorem we present the maximum likelihood estimators of the parameters $\mu_1, \mu_2, ..., \mu_m$ and σ^2.

Theorem 20.1. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Then the MLE's of the parameters μ_i $(i = 1, 2, ..., m)$ and σ^2 of the model are given by

$$\begin{aligned}\widehat{\mu_i} &= \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,\\ \widehat{\sigma^2} &= \frac{1}{nm}\mathrm{SS_W},\end{aligned}$$

where $\overline{Y}_{i\bullet} = \frac{1}{n}\sum\limits_{j=1}^{n} Y_{ij}$ and $\mathrm{SS_W} = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ is the within samples sum of squares.

Proof: The likelihood function is given by

$$\begin{aligned}L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) &= \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}}\right\}\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}(Y_{ij}-\mu_i)^2}.\end{aligned}$$

Taking the natural logarithm of the likelihood function L, we obtain

$$\ln L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = -\frac{nm}{2}\ln(2\,\pi\,\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2. \tag{4}$$

Now taking the partial derivative of (4) with respect to μ_1, $\mu_2, ..., \mu_m$ and σ^2, we get

$$\frac{\partial lnL}{\partial \mu_i} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(Y_{ij}-\mu_i) \tag{5}$$

and

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{nm}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2. \tag{6}$$

Equating these partial derivatives to zero and solving for μ_i and σ^2, respectively, we have

$$\begin{aligned}\mu_i &= \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,\\ \sigma^2 &= \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2,\end{aligned}$$

where

$$\overline{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}.$$

It can be checked that these solutions yield the maximum of the likelihood function and we leave this verification to the reader. Thus the maximum likelihood estimators of the model parameters are given by

$$\begin{aligned} \widehat{\mu_i} &= \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m, \\ \widehat{\sigma^2} &= \frac{1}{nm}\mathrm{SS_W}, \end{aligned}$$

where $\mathrm{SS_W} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The proof of the theorem is now complete.

Define

$$\overline{Y}_{\bullet\bullet} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}. \tag{7}$$

Further, define

$$\mathrm{SS_T} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{8}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \tag{9}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{10}$$

Here $\mathrm{SS_T}$ is the total sum of square, $\mathrm{SS_W}$ is the within sum of square, and $\mathrm{SS_B}$ is the between sum of square.

Next we consider the partitioning of the total sum of squares. The following lemma gives us such a partition.

Lemma 20.1. The total sum of squares is equal to the sum of within and between sum of squares, that is

$$\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}. \tag{11}$$

Proof: Rewriting (8) we have

$$\begin{aligned}
\mathrm{SS_T} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[(Y_{ij} - \overline{Y}_{i\bullet}) + (Y_{i\bullet} - \overline{Y}_{\bullet\bullet})\right]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \\
&\qquad\qquad + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) \\
&= \mathrm{SS_W} + \mathrm{SS_B} + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}).
\end{aligned}$$

The cross-product term vanishes, that is

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) = \sum_{i=1}^{m}(Y_{i\bullet} - Y_{\bullet\bullet})\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet}) = 0.$$

Hence we obtain the asserted result $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$ and the proof of the lemma is complete.

The following theorem is a technical result and is needed for testing the null hypothesis against the alternative hypothesis.

Theorem 20.2. Consider the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where $Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\left(m(n-1)\right)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\, m(n-1)}{\mathrm{SS_W}\,(m-1)} \sim F(m-1, m(n-1))$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(nm-1)$.

Proof: In Chapter 13, we have seen in Theorem 13.7 that if $X_1, X_2, ..., X_n$ are independent random variables each one having the distribution $N(\mu, \sigma^2)$, then their mean $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ have the following properties:

(i) $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independent, and

(ii) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$.

Now using (i) and (ii), we establish this theorem.

(a) Using (ii), we see that

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(n-1)$$

for each $i = 1, 2, ..., m$. Since

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{j=1}^{n}\left(Y_{i'j} - \overline{Y}_{i'\bullet}\right)^2$$

are independent for $i' \neq i$, we obtain

$$\sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).$$

Hence

$$\begin{aligned}\frac{SS_W}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \\ &= \sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).\end{aligned}$$

(b) Since for each $i = 1, 2, ..., m$, the random variables $Y_{i1}, Y_{i2}, ..., Y_{in}$ are independent and

$$Y_{i1}, Y_{i2}, ..., Y_{in} \sim N\left(\mu_i, \sigma^2\right)$$

we conclude by (i) that

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i\bullet}$$

are independent. Further

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i'\bullet}$$

are independent for $i' \neq i$. Therefore, each of the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

is independent of the statistics $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$, and the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

are independent. Thus it follows that the sets

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m \qquad \text{and} \qquad \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m$$

are independent. Thus

$$\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2$$

are independent. Hence by definition, the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Suppose the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true.

(c) Under $\mathrm{H_o}$, the random variables $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$ are independent and identically distributed with $N\left(\mu, \frac{\sigma^2}{n}\right)$. Therefore by (ii)

$$\frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1).$$

Hence

$$\begin{aligned} \frac{\mathrm{SS_B}}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{m}\sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \\ &= \frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1). \end{aligned}$$

(d) Since

$$\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\left(m(n-1)\right)$$

and

$$\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$$

therefore

$$\frac{\frac{\mathrm{SS_B}}{(m-1)\,\sigma^2}}{\frac{\mathrm{SS_W}}{(n(m-1)\,\sigma^2}} \sim F(m-1, m(n-1)).$$

That is

$$\frac{\frac{\mathrm{SS_B}}{(m-1)}}{\frac{\mathrm{SS_W}}{(n(m-1)}} \sim F(m-1, m(n-1)).$$

(e) Under $\mathrm{H_o}$, the random variables Y_{ij}, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$ are independent and each has the distribution $N(\mu, \sigma^2)$. By (ii) we see that

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(nm-1).$$

Hence we have

$$\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(nm-1)$$

and the proof of the theorem is now complete.

From Theorem 20.1, we see that the maximum likelihood estimator of each μ_i $(i = 1, 2, ..., m)$ is given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet},$$

and since $\overline{Y}_{i\bullet} \sim N\left(\mu_i, \frac{\sigma^2}{n}\right)$,

$$E\left(\widehat{\mu_i}\right) = E\left(\overline{Y}_{i\bullet}\right) = \mu_i.$$

Thus the maximum likelihood estimators are unbiased estimator of μ_i for $i = 1, 2, ..., m$.

Since

$$\widehat{\sigma^2} = \frac{\mathrm{SS_W}}{mn}$$

and by Theorem 20.2, $\frac{1}{\sigma^2}\mathrm{SS_W} \sim \chi^2(m(n-1))$, we have

$$E\left(\widehat{\sigma^2}\right) = E\left(\frac{\mathrm{SS_W}}{mn}\right) = \frac{1}{mn}\,\sigma^2\,E\left(\frac{1}{\sigma^2}\mathrm{SS_W}\right) = \frac{1}{mn}\,\sigma^2\,m(n-1) \neq \sigma^2.$$

Thus the maximum likelihood estimator $\widehat{\sigma^2}$ of σ^2 is biased. However, the estimator $\frac{\mathrm{SS_W}}{m(n-1)}$ is an unbiased estimator. Similarly, the estimator $\frac{\mathrm{SS_T}}{mn-1}$ is an unbiased estimator where as $\frac{\mathrm{SS_T}}{mn}$ is a biased estimator of σ^2.

Theorem 20.3. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. The null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > F_\alpha(m-1,\, m(n-1)), \tag{12}$$

where α is the significance level of the hypothesis test and $F_\alpha(m-1,\, m(n-1))$ denotes the $100(1-\alpha)$ percentile of the F-distribution with $m-1$ numerator and $nm-m$ denominator degrees of freedom.

Proof: Under the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$, the likelihood function takes the form

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_{ij}-\mu)^2}{2\sigma^2}}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu)^2}. \end{aligned}$$

Taking the natural logarithm of the likelihood function and then maximizing it, we obtain

$$\widehat{\mu} = \overline{Y}_{\bullet\bullet} \qquad \text{and} \qquad \widehat{\sigma_{\mathrm{H_o}}} = \frac{1}{mn}\mathrm{SS_T}$$

as the maximum likelihood estimators of μ and σ^2, respectively. Inserting these estimators into the likelihood function, we have the maximum of the likelihood function, that is

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{\mathrm{H_o}}}}}\right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2_{\mathrm{H_o}}}}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{\mathrm{H_o}}}}}\right)^{nm} e^{-\frac{mn}{2\,\mathrm{SS_T}}\,\mathrm{SS_T}}$$

which is

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2_{\mathrm{H_o}}}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{13}$$

When no restrictions imposed, we get the maximum of the likelihood function from Theorem 20.1 as

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2\,\mathrm{SS_W}}\,\mathrm{SS_W}}$$

which is

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{14}$$

Next we find the likelihood ratio statistic W for testing the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$. Recall that the likelihood ratio statistic W can be found by evaluating

$$W = \frac{\max L(\mu, \sigma^2)}{\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2)}.$$

Using (13) and (14), we see that

$$W = \left(\frac{\widehat{\sigma^2}}{\widehat{\sigma^2_{\mathrm{H_o}}}} \right)^{\frac{mn}{2}}. \tag{15}$$

Hence the likelihood ratio test to reject the null hypothesis $\mathrm{H_o}$ is given by the inequality

$$W < k_0$$

where k_0 is a constant. Using (15) and simplifying, we get

$$\frac{\widehat{\sigma^2_{\mathrm{H_o}}}}{\widehat{\sigma^2}} > k_1$$

where $k_1 = \left(\frac{1}{k_0}\right)^{\frac{2}{mn}}$. Hence

$$\frac{\mathrm{SS_T}/mn}{\mathrm{SS_W}/mn} = \frac{\widehat{\sigma^2_{\mathrm{H_o}}}}{\widehat{\sigma^2}} > k_1.$$

Using Lemma 20.1 we have

$$\frac{\mathrm{SS_W} + \mathrm{SS_B}}{\mathrm{SS_W}} > k_1.$$

Therefore

$$\frac{\mathrm{SS_B}}{\mathrm{SS_W}} > k \tag{16}$$

where $k = k_1 - 1$. In order to find the cutoff point k in (16), we use Theorem 20.2 (d). Therefore

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > \frac{m(n-1)}{m-1} k$$

Since $\mathcal{F}$ has F distribution, we obtain

$$\frac{m(n-1)}{m-1} k = F_\alpha(m-1,\, m(n-1)).$$

Thus, at a significance level α, reject the null hypothesis $\mathrm{H_o}$ if

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(m(n-1))} > F_\alpha(m-1,\, m(n-1))$$

and the proof of the theorem is complete.

The various quantities used in carrying out the test described in Theorem 20.3 are presented in a tabular form known as the ANOVA table.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	$\mathrm{SS_B}$	$m-1$	$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{m-1}$	$\mathcal{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$
Within	$\mathrm{SS_W}$	$m(n-1)$	$\mathrm{MS_W} = \frac{\mathrm{SS_W}}{m(n-1)}$	
Total	$\mathrm{SS_T}$	$mn-1$		

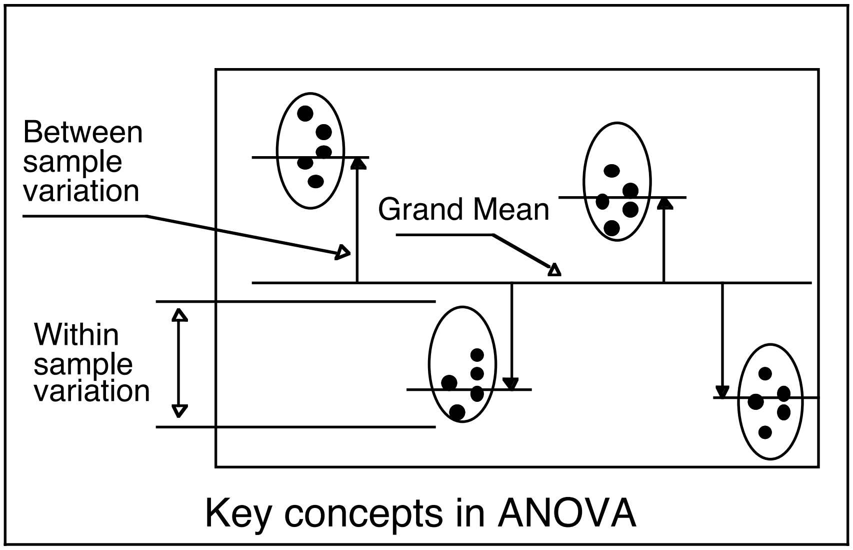
Table 20.1. One-Way ANOVA Table

At a significance level α, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $\mathcal{F} > F_\alpha(m-1,\, m(n-1))$." One can also use the notion of $p-$value to perform this hypothesis test. If the value of the test statistics is $\mathcal{F} = \gamma$, then the p-value is defined as

$$p - \text{value} = P(F(m-1, m(n-1)) \geq \gamma).$$

Alternatively, at a significance level α, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $p - \text{value} < \alpha$."

The following figure illustrates the notions of between sample variation and within sample variation.



The ANOVA model described in (2), that is

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

can be rewritten as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{for } i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where μ is the mean of the m values of μ_i, and $\sum_{i=1}^{m} \alpha_i = 0$. The quantity α_i is called the effect of the i^{th} treatment. Thus any observed value is the sum of an overall mean μ, a treatment or class deviation α_i, and a random element from a normally distributed random variable ϵ_{ij} with mean zero and variance σ^2. This model is called model I, the fixed effects model. The effects of the treatments or classes, measured by the parameters α_i, are regarded as fixed but unknown quantities to be estimated. In this fixed effect model the null hypothesis H_0 is now

$$\mathrm{H_o} : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

and the alternative hypothesis is

$$\mathrm{H_a} : \text{not all the } \alpha_\mathrm{i} \text{ are zero.}$$

The random effects model, also known as model II, is given by

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where μ is the overall mean and

$$A_i \sim N(0, \sigma_A^2) \qquad \text{and} \qquad \epsilon_{ij} \sim N(0, \sigma^2).$$

In this model, the variances σ_A^2 and σ^2 are unknown quantities to be estimated. The null hypothesis of the random effect model is $H_o : \sigma_A^2 = 0$ and the alternative hypothesis is $H_a : \sigma_A^2 > 0$. In this chapter we do not consider the random effect model.

Before we present some examples, we point out the assumptions on which the ANOVA is based on. The ANOVA is based on the following three assumptions:

(1) *Independent Samples:* The samples taken from the population under consideration should be independent of one another.

(2) *Normal Population:* For each population, the variable under consideration should be normally distributed.

(3) *Equal Variance:* The variances of the variables under consideration should be the same for all the populations.

Example 20.1. The data in the following table gives the number of hours of relief provided by 5 different brands of headache tablets administered to 25 subjects experiencing fevers of 38^oC or more. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is same for all 5 brands.

Tablets				
A	B	C	D	F
5	9	3	2	7
4	7	5	3	6
8	8	2	4	9
6	6	3	1	4
3	9	7	4	7

Answer: Using the formulas (8), (9) and (10), we compute the sum of squares $\mathrm{SS_W}$, $\mathrm{SS_B}$ and $\mathrm{SS_T}$ as

$$\mathrm{SS_W} = 57.60, \qquad \mathrm{SS_B} = 79.94, \qquad \text{and} \qquad \mathrm{SS_T} = 137.04.$$

The ANOVA table for this problem is shown below.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	79.94	4	19.86	6.90
Within	57.60	20	2.88	
Total	137.04	24		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(4,\, 20) = 2.8661$. Since

$$6.90 = \mathcal{F} > F_{0.05}(4,\, 20) = 2.8661$$

we reject the null hypothesis that the mean number of hours of relief provided by the tablets is same for all 5 brands.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the p-value to be

$$p - \text{value} = P(F(4, 20) \geq 6.90) = 0.001.$$

Hence again we reach the same conclusion since p-value is less then the given α for this problem.

Example 20.2. Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following two data sets.

Data Set 1			Data Set 2		
Sample			Sample		
A	B	C	A	B	C
8.1	8.0	14.8	9.2	9.5	9.4
4.2	15.1	5.3	9.1	9.5	9.3
14.7	4.7	11.1	9.2	9.5	9.3
9.9	10.4	7.9	9.2	9.6	9.3
12.1	9.0	9.3	9.3	9.5	9.2
6.2	9.8	7.4	9.2	9.4	9.3

Answer: Computing the sum of squares SS_W, SS_B and SS_T, we have the following two ANOVA tables:

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.3	2	0.1	0.01
Within	187.2	15	12.5	
Total	187.5	17		

and

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find from the F-table that $F_{0.05}(2,\,15) = 3.68$. For the first data set, since

$$0.01 = \mathcal{F} < F_{0.05}(2,\,15) = 3.68$$

we do not reject the null hypothesis whereas for the second data set,

$$35.0 = \mathcal{F} > F_{0.05}(2,\,15) = 3.68$$

we reject the null hypothesis.

Remark 20.1. Note that the sample means are same in both the data sets. However, there is a less variation among the sample points in samples of the second data set. The ANOVA finds a more significant differences among the means in the second data set. This example suggests that the larger the variation among sample means compared with the variation of the measurements within samples, the greater is the evidence to indicate a difference among population means.

20.2. One-Way Analysis of Variance with Unequal Sample Sizes

In the previous section, we examined the theory of ANOVA when samples are same sizes. When the samples are same sizes we say that the ANOVA is in the balanced case. In this section we examine the theory of ANOVA for unbalanced case, that is when the samples are of different sizes. In experimental work, one often encounters unbalance case due to the death of experimental animals in a study or drop out of the human subjects from a study or due to damage of experimental materials used in a study. Our analysis of the last section for the equal sample size will be valid but have to be modified to accommodate the different sample size.

Consider m independent samples of respective sizes $n_1, n_2, ..., n_m$, where the members of the i^{th} sample, $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, are normal random variables with mean μ_i and unknown variance σ^2. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n_i.$$

Let us denote $N = n_1 + n_2 + \cdots + n_m$. Again, we will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a} : \text{not all the means are equal.}$$

Now we defining

$$\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n} Y_{ij}, \tag{17}$$

$$\overline{Y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}, \tag{18}$$

$$\mathrm{SS_T} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\bullet\bullet} \right)^2, \tag{19}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2, \tag{20}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet} \right)^2 \tag{21}$$

we have the following results analogous to the results in the previous section.

Theorem 20.4. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. Then the MLE's of the parameters μ_i $(i = 1, 2, ..., m)$ and σ^2 of the model are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{N} \mathrm{SS_W},$$

where $\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ and $\mathrm{SS_W} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2$ is the within samples sum of squares.

Lemma 20.2. The total sum of squares is equal to the sum of within and between sum of squares, that is $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$.

Theorem 20.5. Consider the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ..., m, \quad j = 1, 2, ..., n_i,$$

where $Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\,(N-m)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\, m(n-1)}{\mathrm{SS_W}\,(m-1)} \sim F(m-1, N-m)$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(N-1)$.

Theorem 20.6. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. The null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(N-m)} > F_\alpha(m-1,\, N-m),$$

where α is the significance level of the hypothesis test and $F_\alpha(m-1,\, N-m)$ denotes the $100(1-\alpha)$ percentile of the F-distribution with $m-1$ numerator and $N-m$ denominator degrees of freedom.

The corresponding ANOVA table for this case is

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	$\mathrm{SS_B}$	$m-1$	$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{m-1}$	$\mathcal{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$
Within	$\mathrm{SS_W}$	$N-m$	$\mathrm{MS_W} = \frac{\mathrm{SS_W}}{N-m}$	
Total	$\mathrm{SS_T}$	$N-1$		

Table 20.2. One-Way ANOVA Table with unequal sample size

Example 20.3. Three sections of elementary statistics were taught by different instructors. A common final examination was given. The test scores are given in the table below. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that there is a difference in the average grades given by the three instructors.

Elementary Statistics		
Instructor A	Instructor B	Instructor C
75	90	17
91	80	81
83	50	55
45	93	70
82	53	61
75	87	43
68	76	89
47	82	73
38	78	58
	80	70
	33	
	79	

Answer: Using the formulas (17) - (21), we compute the sum of squares SS_W, SS_B and SS_T as

$$SS_W = 10362, \qquad SS_B = 755, \qquad \text{and} \qquad SS_T = 11117.$$

The ANOVA table for this problem is shown below.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	755	2	377	1.02
Within	10362	28	370	
Total	11117	30		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\, 28) = 3.34$. Since

$$1.02 = \mathcal{F} < F_{0.05}(2,\, 28) = 3.34$$

we accept the null hypothesis that there is no difference in the average grades given by the three instructors.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the p-value to be

$$p - \text{value} = P(F(2, 28) \geq 1.02) = 0.374.$$

Hence again we reach the same conclusion since p-value is less then the given α for this problem.

We conclude this section pointing out the advantages of choosing equal sample sizes (balance case) over the choice of unequal sample sizes (unbalance case). The first advantage is that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. The second advantage is that the choice of equal sample size minimizes the probability of committing a type II error.

20.3. Pair wise Comparisons

When the null hypothesis is rejected using the F-test in ANOVA, one may still wants to know where the difference among the means is. There are several methods to find out where the significant differences in the means lie after the ANOVA procedure is performed. Among the most commonly used tests are Scheffé test and Tuckey test. In this section, we give a brief description of these tests.

In order to perform the Scheffé test, we have to compare the means two at a time using all possible combinations of means. Since we have m means, we need $\binom{m}{2}$ pair wise comparisons. A pair wise comparison can be viewed as a test of the null hypothesis $H_0 : \mu_i = \mu_k$ against the alternative $H_a : \mu_i \neq \mu_k$ for all $i \neq k$.

To conduct this test we compute the statistics

$$F_s = \frac{\left(\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}\right)^2}{MS_W \left(\frac{1}{n_i} + \frac{1}{n_k}\right)},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, n_i and n_k are the respective sample sizes, and MS_W is the mean sum of squared of within group. We reject the null hypothesis at a significance level of α if

$$F_s > (m-1) F_\alpha(m-1, N-m)$$

where $N = n_1 + n_2 + \cdots + n_m$.

Example 20.4. Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following data given in the table below. Further perform a Scheffé test to determine where the significant differences in the means lie.

Sample		
1	2	3
9.2	9.5	9.4
9.1	9.5	9.3
9.2	9.5	9.3
9.2	9.6	9.3
9.3	9.5	9.2
9.2	9.4	9.3

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\,15) = 3.68$. Since

$$35.0 = \mathcal{F} > F_{0.05}(2,\,15) = 3.68$$

we reject the null hypothesis. Now we perform the Scheffé test to determine where the significant differences in the means lie. From given data, we obtain $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$. Since $m = 3$, we have to make 3 pair wise comparisons, namely μ_1 with μ_2, μ_1 with μ_3, and μ_2 with μ_3. First we consider the comparison of μ_1 with μ_2. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}\right)^2}{MS_W\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(9.2 - 9.5)^2}{0.004\,\left(\frac{1}{6} + \frac{1}{6}\right)} = 67.5.$$

Since

$$67.5 = F_s > 2\,F_{0.05}(2,\,15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we consider the comparison of μ_1 with μ_3. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W \, \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = \frac{(9.2 - 9.3)^2}{0.004 \, \left(\frac{1}{6} + \frac{1}{6}\right)} = 7.5.$$

Since

$$7.5 = F_s > 2\, F_{0.05}(2,\, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we consider the comparison of μ_2 with μ_3. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W \, \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = \frac{(9.5 - 9.3)^2}{0.004 \, \left(\frac{1}{6} + \frac{1}{6}\right)} = 30.0.$$

Since

$$30.0 = F_s > 2\, F_{0.05}(2,\, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Next consider the Tukey test. Tuckey test is applicable when we have a balanced case, that is when the sample sizes are equal. For Tukey test we compute the statistics

$$Q = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}}{\sqrt{\frac{MS_W}{n}}},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, n is the size of the samples, and MS_W is the mean sum of squared of within group. At a significance level α, we reject the null hypothesis H_0 if

$$|Q| > Q_\alpha(m, \nu)$$

where ν represents the degrees of freedom for the error mean square.

Example 20.5. For the data given in Example 20.4 perform a Tukey test to determine where the significant differences in the means lie.

Answer: We have seen that $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$.

First we compare μ_1 with μ_2. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.3}{\sqrt{\frac{0.004}{6}}} = -11.6189.$$

Since

$$11.6189 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we compare μ_1 with μ_3. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.5}{\sqrt{\frac{0.004}{6}}} = -3.8729.$$

Since

$$3.8729 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we compare μ_2 with μ_3. For this we compute

$$Q = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.5 - 9.3}{\sqrt{\frac{0.004}{6}}} = 7.7459.$$

Since

$$7.7459 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Often in scientific and engineering problems, the experiment dictates the need for comparing simultaneously each treatment with a control. Now we describe a test developed by C. W. Dunnett for determining significant differences between each treatment mean and the control. Suppose we wish to test the m hypotheses

$$H_0 : \mu_0 = \mu_i \quad \text{versus} \quad H_a : \mu_0 \neq \mu_i \qquad \text{for } i = 1, 2, ..., m,$$

where μ_0 represents the mean yield for the population of measurements in which the control is used. To test the null hypotheses specified by H_0 against two-sided alternatives for an experimental situation in which there are m treatments, excluding the control, and n observation per treatment, we first calculate

$$D_i = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\, MS_W}{n}}}, \quad i = 1, 2, ..., m.$$

At a significance level α, we reject the null hypothesis H_0 if

$$|D_i| > D_{\frac{\alpha}{2}}(m, \nu)$$

where ν represents the degrees of freedom for the error mean square. The values of the quantity $D_{\frac{\alpha}{2}}(m, \nu)$ are tabulated for various α, m and ν.

Example 20.6. For the data given in the table below perform a Dunnett test to determine any significant differences between each treatment mean and the control.

Control	Sample 1	Sample 2
9.2	9.5	9.4
9.1	9.5	9.3
9.2	9.5	9.3
9.2	9.6	9.3
9.3	9.5	9.2
9.2	9.4	9.3

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find that $D_{0.025}(2, 15) = 2.44$. Since

$$35.0 = D > D_{0.025}(2,\ 15) = 2.44$$

we reject the null hypothesis. Now we perform the Dunnett test to determine if there is any significant differences between each treatment mean and the control. From given data, we obtain $\overline{Y}_{0\bullet} = 9.2$, $\overline{Y}_{1\bullet} = 9.5$ and $\overline{Y}_{2\bullet} = 9.3$. Since $m = 2$, we have to make 2 pair wise comparisons, namely μ_0 with μ_1, and μ_0 with μ_2. First we consider the comparison of μ_0 with μ_1. For this case, we find

$$D_1 = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}} = \frac{9.5 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 8.2158.$$

Since

$$8.2158 = D_1 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_0$ in favor of the alternative $H_a : \mu_1 \neq \mu_0$.

Next we find

$$D_2 = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\, MS_W}{n}}} = \frac{9.3 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 2.7386.$$

Since

$$2.7386 = D_2 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_0$ in favor of the alternative $H_a : \mu_2 \neq \mu_0$.

20.4. Tests for the Homogeneity of Variances

One of the assumptions behind the ANOVA is the equal variance, that is the variances of the variables under consideration should be the same for all population. Earlier we have pointed out that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. Nevertheless it is advisable to run a preliminary test for homogeneity of variances. Such a test would certainly be advisable in the case of unequal sample sizes if there is a doubt concerning the homogeneity of population variances.

Suppose we want to test the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$$

versus the alternative hypothesis

$$H_a : \text{not all variances are equal.}$$

A frequently used test for the homogeneity of population variances is the Bartlett test. Bartlett (1937) proposed a test for equal variances that was modification of the normal-theory likelihood ratio test.

We will use this test to test the above null hypothesis H_0 against H_a. First, we compute the m sample variances $S_1^2, S_2^2, ..., S_m^2$ from the samples of size $n_1, n_2, ..., n_m$, with $n_1 + n_2 + \cdots + n_m = N$. The test statistics B_c is given by

$$B_c = \frac{(N-m)\,\ln S_p^2 - \sum_{i=1}^{m}(n_i - 1)\,\ln S_i^2}{1 + \frac{1}{3\,(m-1)}\left(\sum_{i=1}^{m}\frac{1}{n_i - 1} - \frac{1}{N-m}\right)}$$

where the pooled variance S_p^2 is given by

$$S_p^2 = \frac{\sum_{i=1}^{m}(n_i - 1)\,S_i^2}{N-m} = \mathrm{MS_W}.$$

It is known that the sampling distribution of B_c is approximately chi-square with $m-1$ degrees of freedom, that is

$$B_c \sim \chi^2(m-1)$$

when $(n_i - 1) \geq 3$. Thus the Bartlett test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level α if

$$B_c > \chi^2_{1-\alpha}(m-1),$$

where $\chi^2_{1-\alpha}(m-1)$ denotes the upper $(1-\alpha)100$ percentile of the chi-square distribution with $m-1$ degrees of freedom.

Example 20.7. For the following data perform an ANOVA and then apply Bartlett test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Data			
Sample 1	Sample 2	Sample 3	Sample 4
34	29	32	34
28	32	34	29
29	31	30	32
37	43	42	28
42	31	32	32
27	29	33	34
29	28	29	29
35	30	27	31
25	37	37	30
29	44	26	37
41	29	29	43
40	31	31	42

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	16.2	3	5.4	0.20
Within	1202.2	44	27.3	
Total	1218.5	47		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\, 44) = 3.23$. Since

$$0.20 = \mathcal{F} < F_{0.05}(2,\, 44) = 3.23$$

we do not reject the null hypothesis.

Now we compute Bartlett test statistic B_c. From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Further, the pooled variance is

$$S_p^2 = \mathrm{MS}_W = 27.3.$$

The statistics B_c is

$$\begin{aligned} B_c &= \frac{(N-m)\,\ln S_p^2 - \displaystyle\sum_{i=1}^{m}(n_i-1)\,\ln S_i^2}{1 + \frac{1}{3\,(m-1)}\left(\displaystyle\sum_{i=1}^{m}\frac{1}{n_i-1} - \frac{1}{N-m}\right)} \\ &= \frac{44\,\ln 27.3 - 11\,[\ln 35.2836 - \ln 30.1401 - \ln 19.4481 - \ln 24.4036\,]}{1 + \frac{1}{3\,(4-1)}\left(\frac{4}{12-1} - \frac{1}{48-4}\right)} \\ &= \frac{1.0537}{1.0378} = 1.0153. \end{aligned}$$

From chi-square table we find that $\chi^2_{0.95}(3) = 7.815$. Hence, since

$$1.0153 = B_c < \chi^2_{0.95}(3) = 7.815,$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

The Bartlett test assumes that the m samples should be taken from m normal populations. Thus Bartlett test is sensitive to departures from normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality. Levene (1960) proposed a test for the homogeneity of population variances that considers the random variables

$$W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$$

and apply a one-way analysis of variance to these variables. If the F-test is significant, the homogeneity of variances is rejected.

Levene (1960) also proposed using F-tests based on the variables

$$W_{ij} = |Y_{ij} - \overline{Y}_{i\bullet}|, \qquad W_{ij} = \ln|Y_{ij} - \overline{Y}_{i\bullet}|, \qquad \text{and} \quad W_{ij} = \sqrt{|Y_{ij} - \overline{Y}_{i\bullet}|}.$$

Brown and Forsythe (1974c) proposed using the transformed variables based on the absolute deviations from the median, that is $W_{ij} = |Y_{ij} - Med(Y_{i\bullet})|$, where $Med(Y_{i\bullet})$ denotes the median of group i. Again if the F-test is significant, the homogeneity of variances is rejected.

Example 20.8. For the data in Example 20.7 do a Levene test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Answer: From data we find that $\overline{Y}_{1\bullet} = 33.00$, $\overline{Y}_{2\bullet} = 32.83$, $\overline{Y}_{3\bullet} = 31.83$, and $\overline{Y}_{4\bullet} = 33.42$. Next we compute $W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The resulting values are given in the table below.

Transformed Data			
Sample 1	Sample 2	Sample 3	Sample 4
1	14.7	0.0	0.3
25	0.7	4.7	19.5
16	3.4	3.4	2.0
16	103.4	103.4	29.3
81	3.4	0.0	2.0
36	14.7	1.4	0.3
16	23.4	8.0	19.5
4	8.0	23.4	5.8
64	17.4	26.7	11.7
16	124.7	34.0	12.8
64	14.7	0.0	91.8
49	3.4	0.7	73.7

Now we perform an ANOVA to the data given in the table above. The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	1430	3	477	0.46
Within	45491	44	1034	
Total	46922	47		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(3,\, 44) = 2.84$. Since

$$0.46 = \mathcal{F} < F_{0.05}(3,\, 44) = 2.84$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

Although Bartlet test is most widely used test for homogeneity of variances a test due to Cochran provides a computationally simple procedure. Cochran test is one of the best method for detecting cases where the variance of one of the groups is much larger than that of the other groups. The test statistics of Cochran test is give by

$$C = \frac{\max\limits_{1\leq i\leq m} \; S_i^2}{\sum\limits_{i=1}^{m} S_i^2}.$$

The Cochran test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level α if

$$C > C_\alpha.$$

The critical values of C_α were originally published by Eisenhart *et al* (1947) for some combinations of degrees of freedom ν and the number of groups m. Here the degrees of freedom ν are

$$\nu = \max_{1\leq i\leq m} (n_i - 1).$$

Example 20.9. For the data in Example 20.7 perform a Cochran test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Answer: From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Hence the test statistic for Cochran test is

$$C = \frac{35.2836}{35.2836 + 30.1401 + 19.4481 + 24.4036} = \frac{35.2836}{109.2754} = 0.3328.$$

The critical value $C_{0.5}(3, 11)$ is given by 0.4884. Since

$$0.3328 = C < C_{0.5}(3, 11) = 0.4884.$$

At a significance level $\alpha = 0.05$, we do not reject the null hypothesis that the variances are equal. Hence Cochran test suggests that the homogeneity of variances condition is met.

20.5. Exercises

1. A consumer organization wants to compare the prices charged for a particular brand of refrigerator in three types of stores in Louisville: discount stores, department stores and appliance stores. Random samples of 6 stores of each type were selected. The results were shown below.

Discount	Department	Appliance
1200	1700	1600
1300	1500	1500
1100	1450	1300
1400	1300	1500
1250	1300	1700
1150	1500	1400

At the 0.05 level of significance, is there any evidence of a difference in the average price between the types of stores?

2. It is conjectured that a certain gene might be linked to ovarian cancer. The ovarian cancer is sub-classified into three categories: stage I, stage II and stage III-IV. There are three random samples available; one from each stage. The samples are labelled with three colors dyes and hybridized on a four channel cDNA microarray (one channel remains unused). The experiment is repeated 5 times and the following data were obtained.

Microarray Data

Array	mRNA 1	mRNA 2	mRNA 3
1	100	95	70
2	90	93	72
3	105	79	81
4	83	85	74
5	78	90	75

Is there any difference between the averages of the three mRNA samples at 0.05 significance level?

3. A stock market analyst thinks 4 stock of mutual funds generate about the same return. He collected the accompaning rate-of-return data on 4 different mutual funds during the last 7 years. The data is given in table below.

Mutual Funds

Year	A	B	C	D
2000	12	11	13	15
2001	12	17	19	11
2002	13	18	15	12
2004	18	20	25	11
2005	17	19	19	10
2006	18	12	17	10
2007	12	15	20	12

Do a one-way ANOVA to decide whether the funds give different performance at 0.05 significance level.

4. Give a proof of the Theorem 20.4.

5. Give a proof of the Lemma 20.2.

6. Give a proof of the Theorem 20.5.

7. Give a proof of the Theorem 20.6.

8. An automobile company produces and sells its cars under 3 different brand names. An autoanalyst wants to see whether different brand of cars have same performance. He tested 20 cars from 3 different brands and recorded the mileage per gallon.

Brand 1	Brand 2	Brand 3
32	31	34
29	28	25
32	30	31
25	34	37
35	39	32
33	36	
34	38	
31		

Do the data suggest a rejection of the null hypothesis at a significance level 0.05 that the mileage per gallon generated by three different brands are same.