Chapter 8

8.1 One Way ANOVA

Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table 13.1.

The model for this situation may be set up as follows. There are 6 observations taken from each of 5 populations with means $\mu_1, \mu_2, \ldots, \mu_5$, respectively. We may wish to test

$H0$: $\mu_1 = \mu_2 = \cdots = \mu_5$,

$H1$: At least two of the means are not equal.

Table 13.1: Absorption of Moisture in Concrete Aggregates

Aggregate	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16854
Mean	553.33	569.33	610.5	465.17	610.67	561.8

Two Sources of Variability in the Data

In the analysis-of-variance procedure, it is assumed that whatever variation exists among the aggregate averages is attributed to (1) variation in absorption among observations *within* aggregate types and (2) variation *among* aggregate types, that is, due to differences in the chemical composition of the aggregates. The **within aggregate variation** is, of course, brought about by various causes. Perhaps humidity and temperature conditions were not kept entirely constant throughout the experiment. It is possible that there was a certain amount of heterogeneity in the batches of raw materials that were used. At any rate, we shall consider the within-sample variation to be **chance or random variation**. Part of the goal of the analysis of variance is to determine if the differences among the 5 sample means are what we would expect due to random variation alone or, rather, due to variation beyond merely random effects, i.e., differences in the chemical composition of the aggregates.

One-Way Analysis of Variance:
Completely Randomized Design (One-Way ANOVA)

Random samples of size n are selected from each of k populations. The k different populations are classified on the basis of a single criterion such as different treatments or groups. Today the term **treatment** is used generally to refer to the various classifications, whether they be different aggregates, different analysts, different fertilizers, or different regions of the country.

Assumptions and Hypotheses in One-Way ANOVA

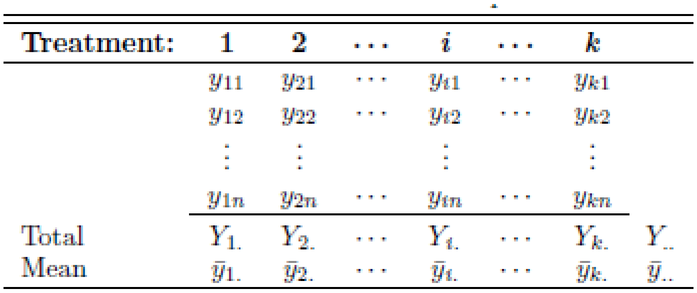
It is assumed that the *k* populations are independent and normally distributed with means $\mu_1, \mu_2, \ldots, \mu_k$ and common variance σ^2.

We wish to derive appropriate methods for testing the hypothesis

$H0$: $\mu_1 = \mu_2 = \cdot\cdot\cdot = \mu_k$,

$H1$: At least two of the means are not equal.

Let yij denote the *j*th observation from the *i*th treatment and arrange the data as in Table 13.2. Here, $Yi.$ is the total of all observations in the sample from the *i*th treatment, $\overline{yi.}$ is the mean of all observations in the sample from the *i*th treatment, $Y..$ is the total of all *nk* observations, and $\overline{y..}$ is the mean of all *nk* observations.



Treatment:	1	2	···	i	···	k	
	y_{11}	y_{21}	···	y_{i1}	···	y_{k1}	
	y_{12}	y_{22}	···	y_{i2}	···	y_{k2}	
	⋮	⋮		⋮		⋮	
	y_{1n}	y_{2n}	···	y_{in}	···	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	···	$Y_{i.}$	···	$Y_{k.}$	$Y_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	···	$\bar{y}_{i.}$	···	$\bar{y}_{k.}$	$\bar{y}_{..}$

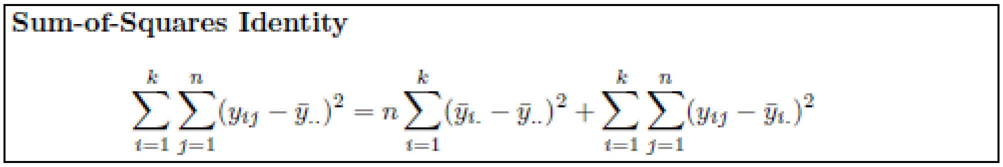
Model for One-Way ANOVA

Each observation may be written in the form

$Yij = \mu i + \varepsilon ij$,

where εij measures the deviation of the *j*th observation of the *i*th sample from the corresponding treatment mean. The εij -term represents random error and plays the same role as the error terms in the regression models.

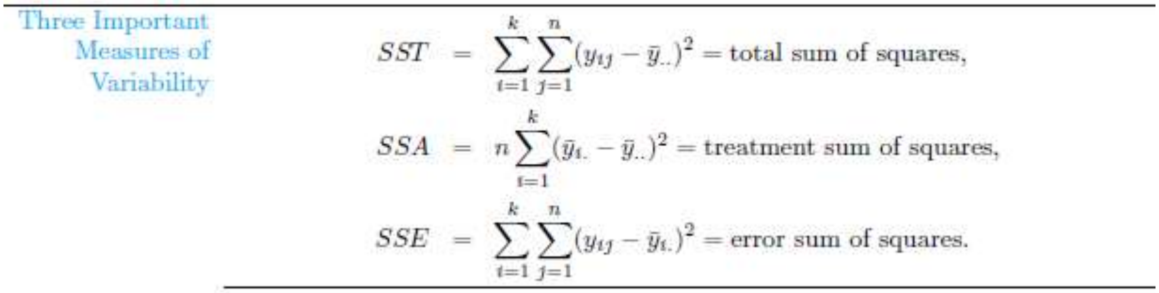
Theorem: Sum-of-Squares Identity



Sum-of-Squares Identity

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

It will be convenient in what follows to identify the terms of the sum-of-squares identity by the following notation:



Three Important Measures of Variability

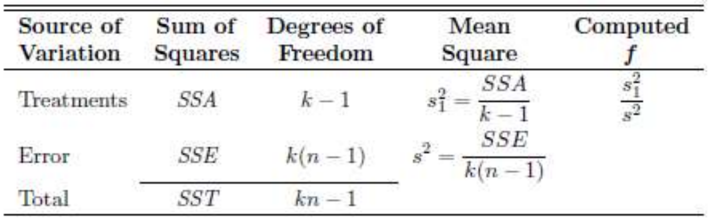
$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = \text{total sum of squares},$$

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 = \text{treatment sum of squares},$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 = \text{error sum of squares}.$$

The sum-of-squares identity can then be represented symbolically by the equation
SST = *SSA* + *SSE*.
The identity above expresses how between-treatment and within-treatment variation add to the total sum of squares.

Analysis of Variance for the One-Way ANOVA



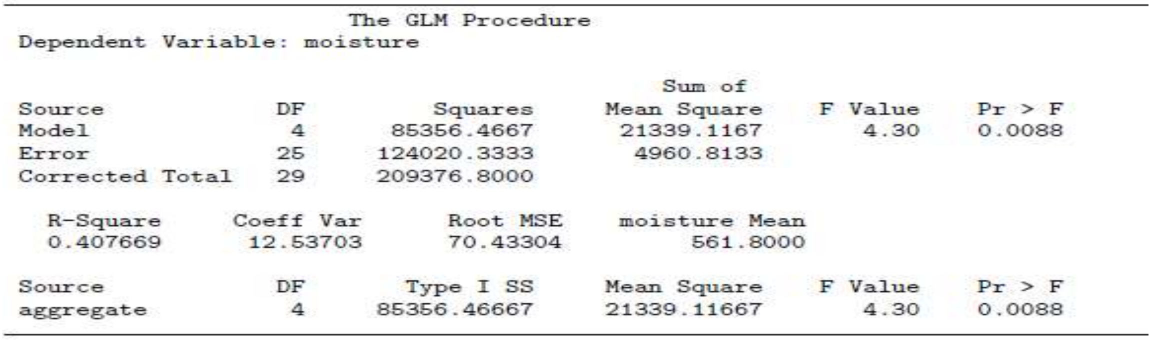
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Treatments	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SSE	$k(n-1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

F-Ratio for Testing Equality of Means
When $H0$ is true, the ratio $f = s_1^2/s^2$ is a value of the random variable F having the F-distribution with $k-1$ and $k(n-1)$ degrees of freedom.
The null hypothesis $H0$ is rejected at the α-level of significance when $f > f\alpha[k - 1, k(n - 1)]$.

Example:

Test the hypothesis $\mu_1 = \mu_2 = \cdot\cdot\cdot = \mu_5$ at the 0.05 level of significance for the data of Table 13.1 on absorption of moisture by various types of cement aggregates.

Solution **:** The hypotheses are
$H0$: $\mu_1 = \mu_2 = \cdot\cdot\cdot = \mu_5$,
$H1$: At least two of the means are not equal.
$\alpha = 0.05$.
Critical region: $f > 2.76$ with $v1 = 4$ and $v2 = 25$ degrees of freedom. The sum-of-squares computations give
SST = 209,377, *SSA*= 85,356,
SSE = 209,377 − 85,356 = 124,021. the ratio $f = s_1^2/s^2$ =4.30 These results and the remaining computations are exhibited in the next figure in the *SAS* ANOVA procedure.



```
                              The GLM Procedure
 Dependent Variable: moisture

                                               Sum of
 Source               DF        Squares     Mean Square    F Value    Pr > F
 Model                 4     85356.4667      21339.1167       4.30    0.0088
 Error                25    124020.3333       4960.8133
 Corrected Total      29    209376.8000

   R-Square      Coeff Var       Root MSE     moisture Mean
   0.407669       12.53703       70.43304          561.8000

 Source               DF      Type I SS     Mean Square    F Value    Pr > F
 aggregate             4    85356.46667     21339.11667       4.30    0.0088
```

Decision: Reject *H*0 and conclude that the aggregates do not have the same mean absorption. The *P*-value for *f* = 4.30 is 0.0088, which is smaller than 0.05.

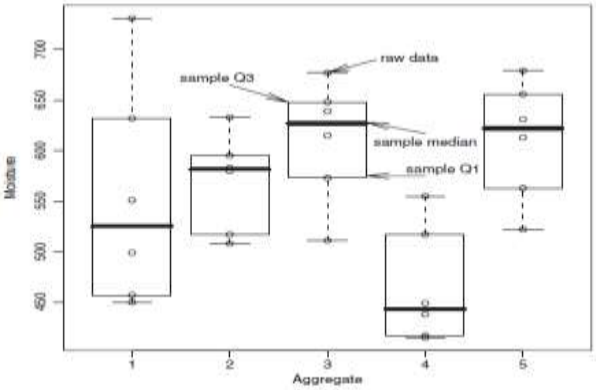
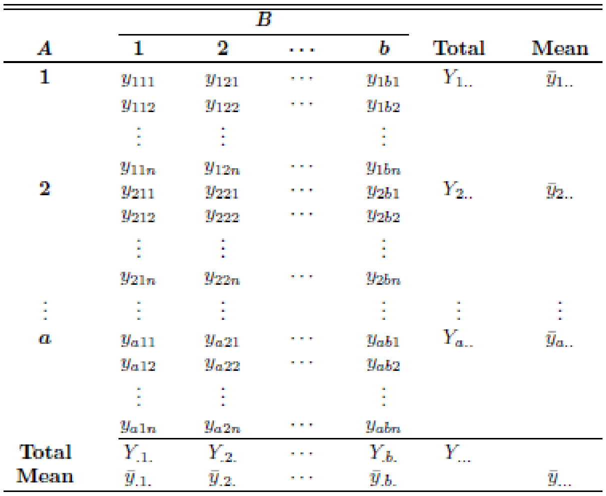


Figure: Box plots for the absorption of moisture in concrete aggregates.

8.2 Two-Factor Analysis of Variance

To present general formulas for the analysis of variance of a two-factor experiment using repeated observations in a completely randomized design, we shall consider the case of *n* replications of the treatment combinations determined by *a* levels of factor *A* and *b* levels of factor *B*. The observations may be classified by means of a rectangular array where the rows represent the levels of factor *A* and the columns represent the levels of factor *B*. Each treatment combination defines a cell in our array. Thus, we have *ab* cells, each cell containing *n* observations. Denoting the *k*th observation taken at the *i*th level of factor *A* and the *j*th level of factor *B* by *yijk*, The following table shows the *abn* observations.



	B					
A	**1**	**2**	**⋯**	**b**	**Total**	**Mean**
1	y_{111}	y_{121}	⋯	y_{1b1}	$Y_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	⋯	y_{1b2}		
	⋮	⋮		⋮		
	y_{11n}	y_{12n}	⋯	y_{1bn}		
2	y_{211}	y_{221}	⋯	y_{2b1}	$Y_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	⋯	y_{2b2}		
	⋮	⋮		⋮		
	y_{21n}	y_{22n}	⋯	y_{2bn}		
⋮	⋮	⋮		⋮	⋮	⋮
a	y_{a11}	y_{a21}	⋯	y_{ab1}	$Y_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	⋯	y_{ab2}		
	⋮	⋮		⋮		
	y_{a1n}	y_{a2n}	⋯	y_{abn}		
Total	$Y_{.1.}$	$Y_{.2.}$	⋯	$Y_{.b.}$	$Y_{...}$	
Mean	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	⋯	$\bar{y}_{.b.}$		$\bar{y}_{...}$

The observations in the (*ij*)th cell constitute a random sample of size *n* from a population that is assumed to be normally distributed with mean *μij* and variance *σ*2. All *ab* populations are assumed to have the same variance *σ*2. Let us define the following useful symbols:

Yij. = sum of the observations in the (*ij*)th cell,
Yi.. = sum of the observations for the *i*th level of factor *A*,
Y.*j*. = sum of the observations for the *j*th level of factor *B*,
Y... = sum of all *abn* observations,
$\overline{y\imath\jmath.}$= mean of the observations in the (*ij*)th cell,
$\overline{y\imath..}$= mean of the observations for the *i*th level of factor *A*,
$\overline{y.\jmath.}$= mean of the observations for the *j*th level of factor *B*,
$\overline{y...}$= mean of all *abn* observations.

Unlike in the one-factor situation covered at length in Chapter 13, here we are assuming that the **populations**, where *n* independent identically distributed observations are taken, are **combinations** of factors. Also we will assume throughout that an equal number (*n*) of observations are taken at each factor combination.

Model and Hypotheses for the Two-Factor Problem

Each observation in Table 14.1 may be written in the form
yijk = *μij* + *εijk*,

where _ijk measures the deviations of the observed $yijk$ values in the (ij)th cell from the population mean μij . If we let $(\alpha\beta)ij$ denote the interaction effect of the ith level of factor A and the jth level of factor B, αi the effect of the ith level of factor A, βj the effect of the jth level of factor B, and μ the overall mean, we can write
$\mu ij = \mu + \alpha i + \beta j + (\alpha\beta)ij$,
and then
$yijk = \mu + \alpha i + \beta j + (\alpha\beta)ij + \varepsilon ijk$,

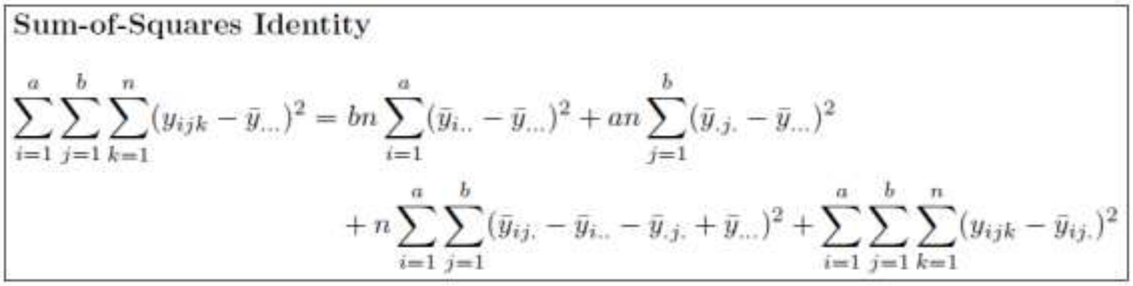
on which we impose the restrictions
$\sum_{i=1}^{a} \alpha_i$=0
$\sum_{j=1}^{b} \beta j$=0
$\sum_{i=1}^{a} (\alpha\beta) ij$=0
$\sum_{j=1}^{b} (\alpha\beta) ij$ =0

The three hypotheses to be tested are as follows:
1. H_0': $\alpha 1 = \alpha 2 = \cdots = \alpha a = 0$,
H_1': At least one of the αi is not equal to zero.
2. H_0'' : $\beta 1 = \beta 2 = \cdots = \beta b = 0$,
H_1'' : At least one of the βj is not equal to zero.
3. H_0''': $(\alpha\beta)11 = (\alpha\beta)12 = \cdots = (\alpha\beta)ab = 0$,
H_1''': At least one of the $(\alpha\beta)ij$ is not equal to zero.

We warned the reader about the problem of masking of main effects when interaction is a heavy contributor in the model. It is recommended that the interaction test result be considered first. The interpretation of the main effect test follows, and the nature of the scientific conclusion depends on whether interaction is found.
If interaction is ruled out, then hypotheses 1 and 2 above can be tested and the interpretation is quite simple. However, if interaction is found to be present the interpretation can be more complicated, as we have seen from the discussion of the drying time and temperature in the previous section. In what follows, the structure of the tests of hypotheses 1, 2, and 3 will be discussed. Interpretation of results will be incorporated in the discussion of the analysis in Example 14.1. The tests of the hypotheses above will be based on a comparison of independent estimates of $\sigma 2$ provided by splitting the total sum of squares of our data into four components by means of the following identity.

Partitioning of Variability in the Two-Factor Case

Theorem:



Sum-of-Squares Identity

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2$$

$$+\,n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

Symbolically, we write the sum-of-squares identity as
SST = *SSA* + *SSB* + *SS*(*AB*) + *SSE*,
where *SSA* and *SSB* are called the sums of squares for the main effects *A* and *B*, respectively, *SS*(*AB*) is called the interaction sum of squares for *A* and *B*, and *SSE* is the error sum of squares. The degrees of freedom are partitioned according to the identity
abn − 1 = (*a* − 1) + (*b* − 1) + (*a* − 1)(*b* − 1) + *ab*(*n* − 1).

Formation of Mean Squares

If we divide each of the sums of squares on the right side of the sum-of-squares identity by its corresponding number of degrees of freedom, we obtain the four statistics

$$S_1^2=\frac{SSA}{a-1},\qquad S_2^2=\frac{SSB}{b-1},\qquad S_3^2=\frac{SS(AB)}{(a-1)(b-1)},\qquad S^2=\frac{SSE}{ab(n-1)}.$$

All of these variance estimates are independent estimates of σ^2 under the condition that there are no effects *αi*, *βj* , and, of course, (*αβ*)*ij* . If we interpret the sums of squares as functions of the independent random variables *y*111, *y*112, . . . , *yabn*, it is not difficult to verify that

$$E(S_1^2)=E\left[\frac{SSA}{a-1}\right]=\sigma^2+\frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2,$$

$$E(S_2^2)=E\left[\frac{SSB}{b-1}\right]=\sigma^2+\frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

$$E(S_3^2)=E\left[\frac{SS(AB)}{(a-1)(b-1)}\right]=\sigma^2+\frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2,$$

$$E(S^2)=E\left[\frac{SSE}{ab(n-1)}\right]=\sigma^2,$$

from which we immediately observe that all four estimates of *σ*2 are unbiased when
H_0', H_0'' *and* H_0''' are true.

To test the hypothesis H_0', that the effects of factors *A* are all equal to zero, we compute the following ratio:

F-Test for Factor *A*

*f*1 =s_1^2 /s^2 ,
which is a value of the random variable *F*1 having the *F*-distribution with *a*−1 and *ab*(*n*−1) degrees of freedom when H_0' is true. The null hypothesis is rejected at the *α*-level of significance when

$f1 > f\alpha[a - 1, ab(n - 1)]$.

Similarly, to test the hypothesis H_0'' that the effects of factor B are all equal to zero, we compute the following ratio:

F-Test for Factor *B*

$f2 = s_2^2 / s^2$,

which is a value of the random variable $F2$ having the F-distribution with $b-1$and $ab(n - 1)$ degrees of freedom when H_0'' is true. This hypothesis is rejected at the α-level of significance when

$f2 > f\alpha[b - 1, ab(n - 1)]$.

Finally, to test the hypothesis H_0''', that the interaction effects are all equal to zero, we compute the following ratio:

F-Test for Interaction

$f3 = s_3^2 / s^2$,

which is a value of the random variable $F3$ having the F-distribution with $(a - 1)(b - 1)$ and $ab(n - 1)$ degrees of freedom when H_0''' is true. We conclude that, at the α-level of significance, interaction is present when

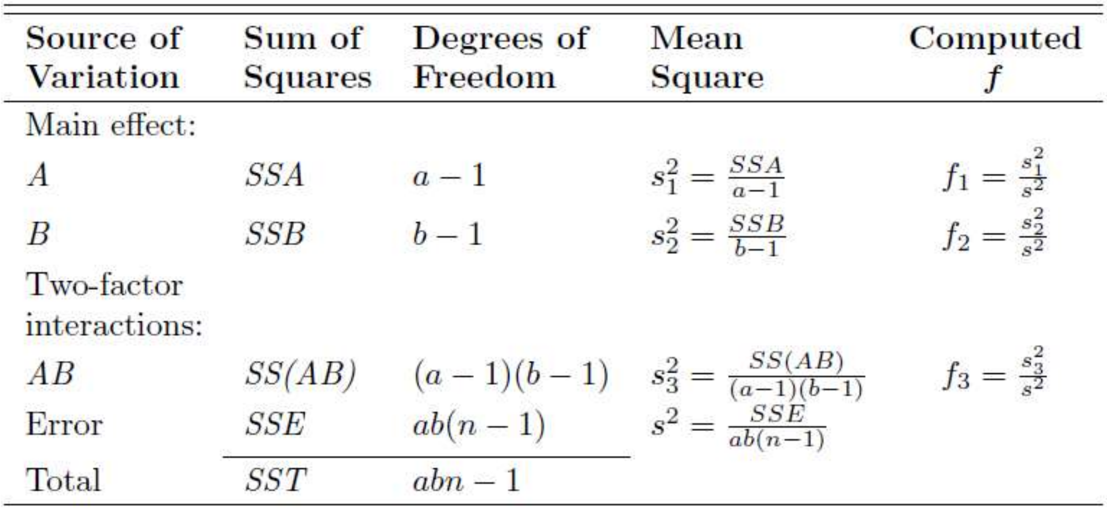
$f3 > f\alpha[(a - 1)(b - 1), ab(n - 1)]$.

As indicated before, it is advisable to interpret the test for interaction before attempting to draw inferences on the main effects. If interaction is not significant, there is certainly evidence that the tests on main effects are interpretable.

Rejection of hypothesis 1 on implies that the response means at the levels of factor *A* are significantly different, while rejection of hypothesis 2 implies a similar condition for the means at levels of factor *B*. However, a significant interaction could very well imply that the data should be analyzed in a somewhat different manner—**perhaps observing the effect of factor *A* at fixed levels of factor *B***, and so forth.

The computations in an analysis-of-variance problem, for a two-factor experiment with *n* replications, are usually summarized as

Analysis of Variance for the Two-Factor Experiment with *n* Replications



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Main effect:				
A	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Two-factor interactions:				
AB	$SS(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$		

In an experiment conducted to determine which of 3 different missile systems is preferable, the propellant burning rate for 24 static firings was measured. Four different propellant types were used. The experiment yielded duplicate observations of burning rates at each combination of the treatments.

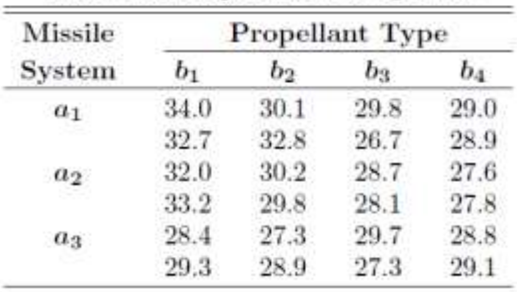
The data, after coding, are given in Table 14.3. Test the following hypotheses:

(a) H_0': there is no difference in the mean propellant burning rates when different missile systems are used,

(b) H_0'': there is no difference in the mean propellant burning rates of the 4 propellant types,

(c) H_0''': there is no interaction between the different missile systems and the different propellant types.

Table 14.3: Propellant Burning Rates



Missile System	Propellant Type			
	b_1	b_2	b_3	b_4
a_1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
a_2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
a_3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1

***Solution* :**

1. (a) H_0': $\alpha 1 = \alpha 2 = \alpha 3 = 0$.

(b) H_0'': $\beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$.

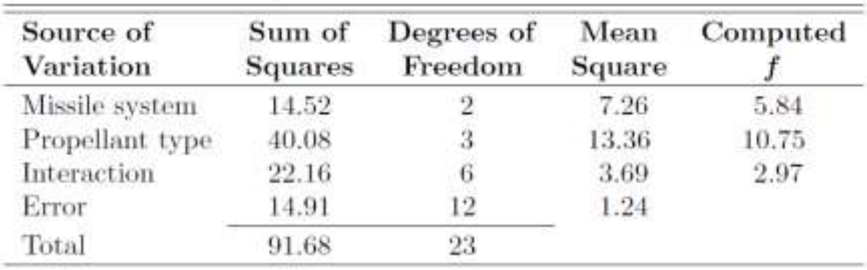
(c) H_0''': $(\alpha\beta)11 = (\alpha\beta)12 = \cdot\cdot\cdot = (\alpha\beta)34 = 0$.

2. (a) H_1': At least one of the αi is not equal to zero.

(b) H_1'': At least one of the βj is not equal to zero.

(c) H_1''': At least one of the $(\alpha\beta)ij$ is not equal to zero.

Analysis of Variance for the Data

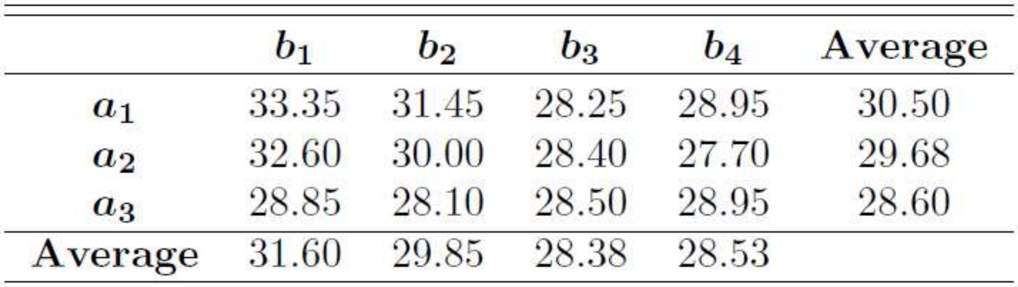


Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Missile system	14.52	2	7.26	5.84
Propellant type	40.08	3	13.36	10.75
Interaction	22.16	6	3.69	2.97
Error	14.91	12	1.24	
Total	91.68	23		

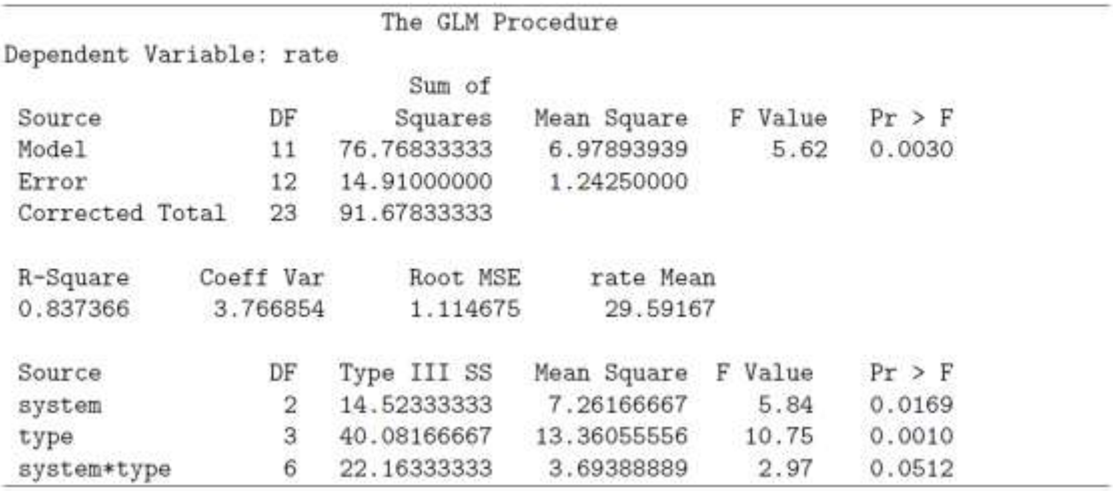
(a) Reject H_0' and conclude that different missile systems result in different mean propellant burning rates. $f1 = s_1^2 / s^2 = 5.84$,
which is a value of the random variable $F1$ having the F-distribution with $a-1=2$ and $ab(n-1)=12$ degrees of freedom when H_0' is true. The null hypothesis is rejected at the α-level of significance when
$f1 > f\alpha[2, 12]= 3.89$

(b) Reject H_0'' and conclude that the mean propellant burning rates are not the same for the four propellant types. $f2 = s_2^2 / s^2 = 10.75$,
which is a value of the random variable $F2$ having the F-distribution with $b-1=3$and $ab(n-1)=12$ degrees of freedom when H_0'' is true. This hypothesis is rejected at the α-level of significance when
$f2 > f\alpha[3, 12]=3.49$.

(c) Fail to Reject H_0''', $f3 = s_3^2 / s^2 =2.97$,
which is a value of the random variable $F3$ having the F-distribution with $(a-1)(b-1)=6$and $ab(n-1)=12$ degrees of freedom when H_0''' is true. We conclude that, at the α-level of significance, interaction is present when
$f3 > f\alpha[(6, 12]=3.00$. Interaction is barely insignificant at the 0.05 level.,
This would indicate that interaction must be taken seriously.



	b_1	b_2	b_3	b_4	Average
a_1	33.35	31.45	28.25	28.95	30.50
a_2	32.60	30.00	28.40	27.70	29.68
a_3	28.85	28.10	28.50	28.95	28.60
Average	31.60	29.85	28.38	28.53	



```
                          The GLM Procedure
Dependent Variable: rate
                                Sum of
 Source              DF        Squares    Mean Square   F Value    Pr > F
 Model               11    76.76833333     6.97893939      5.62    0.0030
 Error               12    14.91000000     1.24250000
 Corrected Total     23    91.67833333

 R-Square     Coeff Var      Root MSE      rate Mean
 0.837366      3.766854      1.114675       29.59167

 Source              DF    Type III SS    Mean Square  F Value    Pr > F
 system               2    14.52333333     7.26166667     5.84    0.0169
 type                 3    40.08166667    13.36055556    10.75    0.0010
 system*type          6    22.16333333     3.69388889     2.97    0.0512
```

SAS printout of the analysis of the propellant rate data