CHAPTER 8: ANALYSIS OF VARIANCE (ANOVA)

STATISTICS (FERM)
Lecturer: Nguyen Minh Quan, PhD
quannm@hcmiu.edu.vn

CONTENTS

1. One-factor analysis of variance

2. Two-Factor Analysis of Variance: Parameter Estimation

3. Two-Factor Analysis of Variance: Testing Hypotheses

Introduction to analysis of variance

- It is a statistical quality control method.

- The emphasis in statistical quality control has shifted from overseeing a manufacturing process to designing that process

- For instance, when producing computer chips, for a given set of choices of these factors (size, shape, etc..), the manufacturer wants to know the mean quality value of the resulting chip.

- The analysis of variance (ANOVA), invented by R. A. Fisher, is a statistical technique used for analyzing the foregoing type of problem. It is a general method for making inferences about a multitude of parameters relating to population means.

One-factor analysis of variance

- Consider m samples, each of size n. Suppose that these samples are independent and that sample i comes from a population that is normally distributed with mean μ_i and variance σ^2, $i = 1, ..., m$.

- We will be interested in testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 = ... = \mu_m$$

against

H_a: not all the means are equal

One-factor analysis of variance

- Let $\overline{X}_i$ and S_i^2 denote the sample mean and sample variance, respectively, for the data of the ith sample, $i = 1, ..., m$.
- Our test of our null hypothesis will be carried out by comparing the values of two estimators of the common variance σ^2.
- Our first estimator of σ^2 is given by

$$\sum_{i=1}^{m} S_i^2 / m$$

Note that this estimator was obtained without assuming anything about the truth or falsity of the null hypothesis.

One-factor analysis of variance

- Let

$$\overline{S}^2 = \frac{\sum_{i=1}^{m} \left(\overline{X}_i - \overline{\overline{X}}\right)^2}{m-1}, \text{ where } \overline{\overline{X}} = \frac{1}{m}\sum_{i=1}^{m} \overline{X}_i$$

- $\overline{S}^2$ is an unbiased estimator of σ^2/n when H_0 is true, it follows in this case that $n\overline{S}^2$ is an estimator of σ^2.

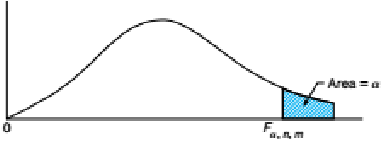
- Since it can be shown that $n\overline{S}^2$ will tend to be larger when H_0 is not true, it is reasonable to let the test statistic TS be given

$$TS = \frac{n\overline{S}^2}{\sum_{i=1}^{m} S_i^2/m}$$

and to reject H_0 when TS is sufficiently large.

One-factor analysis of variance

- To determine how large TS needs to be to justify rejecting H_0, we use the fact that when H_0 is true, TS will have what is known as an F distribution with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.



Density function of $F_{n,m}$.

- Let $F_{\alpha,m-1,m(n-1)}$ denote the α critical value of this distribution. The significance-level-α test of H_0 is as follows:
  - Reject H_0 if $TS = \frac{n\overline{S}^2}{\sum_{i=1}^{m} S_i^2/m} \geq F_{\alpha,m-1,m(n-1)}$.
  - Do not reject H_0 otherwise ("accept").

A Remark on the Degrees of Freedom

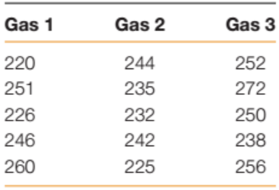
- The numerator degrees of freedom of the F random variable are determined by the numerator estimator $n\overline{S}^2$. Since $n\overline{S}^2$ is the sample variance from a sample of size m, it follows that it has m - 1 degrees of freedom.

- Similarly, the denominator estimator is based on the statistic $\sum\limits_{i=1}^{m} S_i^2$. Since each of the sample variances S_i is based on a sample of size n, it follows that they each have n - 1 degrees of freedom. Summing the m sample variances then results in a statistic with m(n -1) degrees of freedom.

One-factor analysis of variance

Example

An investigator for a consumer cooperative organized a study of the mileages obtainable from three different brands of gasoline. Using 15 identical motors set to run at the same speed, the investigator randomly assigned each brand of gasoline to 5 of the motors. Each of the motors was then run on 10 gallons of gasoline, with the total mileages obtained as follows.



Gas 1	Gas 2	Gas 3
220	244	252
251	235	272
226	232	250
246	242	238
260	225	256

Test the hypothesis that the average mileage obtained is the same for all three types of gasoline. Use the 5 percent level of significance.

One-factor analysis of variance

Solution

We see that m = 3 and n = 5.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

against H_a: not all the means are equal

The sample means are

$$\overline{X_1} = 240.6; \overline{X_2} = 235.6; \overline{X_3} = 253.6$$

The average of the three sample means is

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{m} \overline{X_i}}{m} = \frac{240.6 + 235.6 + 253.6}{3} = 243.2667$$

One-factor analysis of variance

Solution

Therefore, the sample variance of the data $\overline{X_i}, i = 1, 2, 3$, is

$$\overline{S}^2 = \frac{\sum_{i=1}^{m} \left(\overline{X}_i - \overline{\overline{X}}\right)^2}{m-1} = 86.3333$$

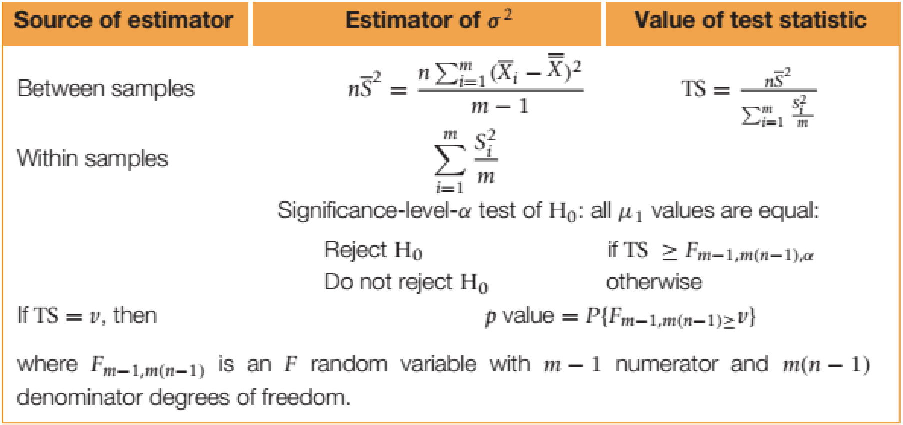
Computing the sample variances from the three samples yields $S_1^2 = 287.8$, $S_2^2 = 59.3$, and $S_3^2 = 150.8$.

$$TS = \frac{5\overline{S}^2}{\frac{1}{3}\sum_{i=1}^{3} S_i^2} = \frac{431.667}{165.967} = 2.60$$

From Appendix Table A4, we see that $F_{0.05,m-1,m(n-1)} = F_{0.05,2,12} = 3.89$. Since $TS < F_{0.05,2,12}$, we cannot reject (i.e. we "accept") the null hypothesis that the gasolines give equal mileage.

Summary: One-Factor ANOVA Table

Variables $\bar{X}_i$ and S_i^2, $i = 1, ..., m$, are the sample means and sample variances, respectively, of independent samples of size n from normal populations having means μ_i and a common variance σ.



Source of estimator	Estimator of σ^2	Value of test statistic
Between samples	$n\overline{S}^2 = \dfrac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{\overline{X}})^2}{m-1}$	$\text{TS} = \dfrac{n\overline{S}^2}{\sum_{i=1}^{m}\frac{S_i^2}{m}}$
Within samples	$\displaystyle\sum_{i=1}^{m}\frac{S_i^2}{m}$	

Significance-level-α test of H_0: all μ_1 values are equal:

Reject H_0 if $\text{TS} \geq F_{m-1,m(n-1),\alpha}$

Do not reject H_0 otherwise

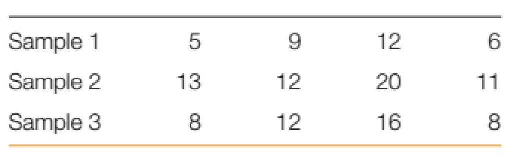
If $\text{TS} = v$, then p value $= P\{F_{m-1,m(n-1)\geq}v\}$

where $F_{m-1,m(n-1)}$ is an F random variable with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.

One-factor analysis of variance

Exercise 1

Consider the data from three samples, each of size 4. (That is, m = 3, n = 4.)



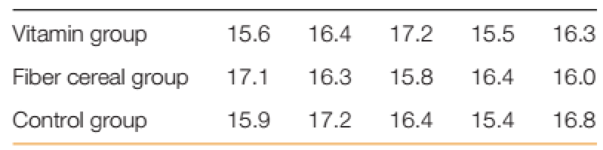
Sample 1	5	9	12	6
Sample 2	13	12	20	11
Sample 3	8	12	16	8

Test the hypothesis that the three population means are equal. Use the 5 percent level of significance.

One-factor analysis of variance

Exercise 2

A nutritionist randomly divided 15 bicyclists into three groups of five each. Members of the first group were given vitamin supplements to take with each of their meals over the next 3 weeks. The second group was instructed to eat a particular type of high-fiber whole-grain cereal. for the next 3 weeks. Members of the third group were instructed to eat as they normally do. After the 3-week period elapsed, the nutritionist had each bicyclist ride 6 miles. The following times were recorded.



Vitamin group	15.6	16.4	17.2	15.5	16.3
Fiber cereal group	17.1	16.3	15.8	16.4	16.0
Control group	15.9	17.2	16.4	15.4	16.8

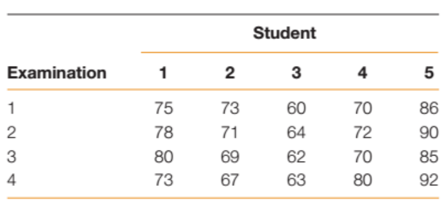
Are these data consistent with the hypothesis that neither the vitamin nor the fiber cereal affects the speed of a bicyclist? Use the 5 percent level of significance.

Two-Factor Analysis of Variance: Introduction

In this section we suppose that each data value is affected by two factors.

Exercise 2

Four different standardized reading achievement tests were administered to each of five students. Their scores were as follows:



	Student				
Examination	**1**	**2**	**3**	**4**	**5**
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

Each value in this set of 20 data points is affected by two factors: the examination and the student whose score on that examination is being recorded. The examination factor has four possible values, or levels, and the student factor has five possible levels.

Two-Factor Analysis of Variance: Introduction

- In general, let us suppose that there are m possible levels of the first factor and n possible levels of the second. Let X_{ij} denote the value of the data obtained when the first factor is at level i and the second factor is at level j.

$$\begin{array}{cccc} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & X_{ij} & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{array}$$

- We refer to the first factor as the row factor and the second factor as the column factor.

Two-Factor Analysis of Variance: Parameter Estimation

- We suppose that all the data values X_{ij}, $i = 1, ..., m, j = 1, ..., n$, are independent normal random variables with common variance σ^2.We will suppose that the mean value of the data point depends on both its row and its column.

- If we let X_{ij} represent the value of the jth member of sample i, then this model supposes that $E[X_{ij}] = \mu_i$.

- If we now let μ denote the average value of the μ_i, that is,

$$\mu = \frac{\sum\limits_{i=1}^{m} \mu_i}{m}$$

Two-Factor Analysis of Variance: Parameter Estimation

- In the case of two factors, we suppose that the expected value of variable X_{ij} can be expressed as follows:

  $$E\left[X_{ij}\right] = \mu + \alpha_i + \beta_j$$

  The value μ is referred to as the grand mean, α_i is the deviation from the grand mean due to row i, and β_j is the deviation from the grand mean due to column j.

- It is easy to see that

  $$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{n} \beta_j = 0$$

Two-Factor Analysis of Variance: Parameter Estimation

Let us start by determining estimators for parameters μ, α_i, and β_j, $i = 1, ..., m, j = 1, ..., n$. To do so, we will find it convenient to introduce the following dot notation. Let

$$X_{i.} = \frac{\sum\limits_{j=1}^{n} X_{ij}}{n} = average\ of\ all\ values\ in\ row\ i$$

$$X_{.j} = \frac{\sum\limits_{i=1}^{m} X_{ij}}{m} = average\ of\ all\ values\ in\ column\ j$$

$$X_{..} = \frac{\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} X_{ij}}{nm} average\ of\ all\ nm\ data\ values$$

Two-Factor Analysis of Variance: Parameter Estimation

- It is not difficult to show that

$$E\left[X_{i.}\right]=\mu+\alpha_i; E\left[X_{.j}\right]=\mu+\beta_i; E\left[X_{..}\right]=\mu$$

- Therefore,

$$E\left[X_{..}\right]=\mu; E\left[X_{i.}-X_{..}\right]=\alpha_i; E\left[X_{.j}-X_{..}\right]=\beta_j$$

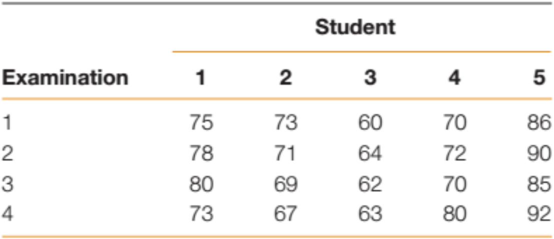
- We see that unbiased estimators of μ, α_i, and β_j-call them $\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$-are given by

$$\hat{\mu}=X_{..}, \hat{\alpha}_i=X_{i.}-X_{..}, \hat{\beta}_j=X_{.j}-X_{..}$$

Two-Factor Analysis of Variance: Parameter Estimation

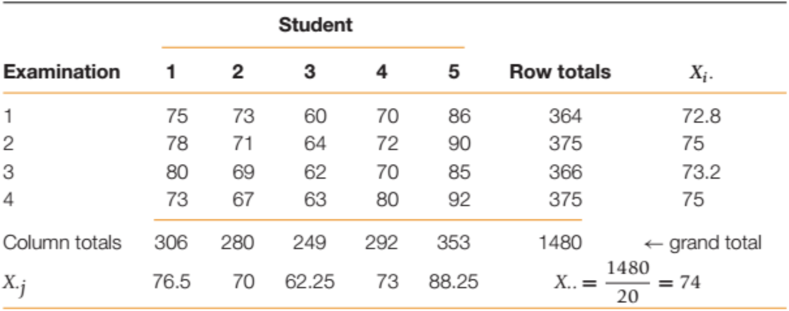
Example

The following data give the scores obtained when four different reading tests were given to each of five students. Use it to estimate the parameters of the model.



	Student				
Examination	**1**	**2**	**3**	**4**	**5**
1	75	73	60	70	86
2	78	71	64	72	90
3	80	69	62	70	85
4	73	67	63	80	92

Solution



	Student						
Examination	1	2	3	4	5	Row totals	$X_{i\cdot}$
1	75	73	60	70	86	364	72.8
2	78	71	64	72	90	375	75
3	80	69	62	70	85	366	73.2
4	73	67	63	80	92	375	75
Column totals	306	280	249	292	353	1480	← grand total
$X_{\cdot j}$	76.5	70	62.25	73	88.25	$X_{\cdot\cdot} = \frac{1480}{20} = 74$	

The estimators are

$$\hat{\mu} = 74$$

$$\begin{aligned}
\hat{\alpha}_1 &= 72.8 - 74 = -1.2 & \hat{\beta}_1 &= 76.5 - 74 = 2.5 \\
\hat{\alpha}_2 &= 75 - 74 = 1 & \hat{\beta}_2 &= 70 - 74 = -4 \\
\hat{\alpha}_3 &= 73.2 - 74 = -0.8 & \hat{\beta}_3 &= 62.25 - 74 = -11.75 \\
\hat{\alpha}_4 &= 75 - 74 = 1 & \hat{\beta}_4 &= 73 - 74 = -1 \\
& & \hat{\beta}_5 &= 88.25 - 74 = 14.25
\end{aligned}$$

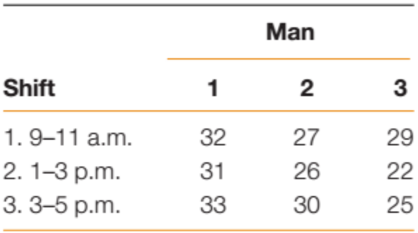
Solution

Comments on the use of the estimators:

- For instance, if one of the students is randomly chosen and then given a randomly chosen examination, then our estimate of the mean score that will be obtained is $\hat{\mu} = 74$.
- If we were told that examination i was taken, then this would increase our estimate of the mean score by the amount $\hat{\alpha}_i$; if we were told that the student chosen was number j, then this would increase our estimate of the mean score by the amount $\hat{\beta}_j$.
- Thus, for instance, we would estimate that the score obtained on examination 1 by student 2 is the value of a random variable whose mean is $\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2 = 741.24 = 68$.

Exercise

The following data refer to the numbers of boxes packed by each of three men during three different shifts.



	Man		
Shift	**1**	**2**	**3**
1. 9–11 a.m.	32	27	29
2. 1–3 p.m.	31	26	22
3. 3–5 p.m.	33	30	25

Assuming the model of this section, estimate the unknown parameters.

Two-Factor Analysis of Variance: Testing Hypotheses

Consider the two-factor model in which one has data values X_{ij}, $i = 1, ..., m$ and $j = 1, ..., n$. These data are assumed to be independent normal random variables with a common variance σ^2 and with mean values satisfying

$$E\left[X_{ij}\right] = \mu + \alpha_i + \beta_j$$

where

$$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{n} \beta_j = 0$$

Two-Factor Analysis of Variance: Testing Hypotheses

In this section we will test the hypothesis for row
H_0: all $\alpha_i = 0$ against H_a: not all α_i are 0.

Note: This null hypothesis states that there is no row effect, in that the value of a datum is not affected by its row factor level.

We will also test the analogous hypothesis for columns, namely,
H_0: all $\beta_j = 0$ against H_a: not all β_j are 0.

Two-Factor Analysis of Variance: Testing Hypotheses

Key method: To obtain tests for the foregoing null hypotheses, we will apply the analysis of variance approach in which two different estimators are derived for the variance σ^2: They are SS_e/N and $SS_r/m-1$, where

$$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2$$

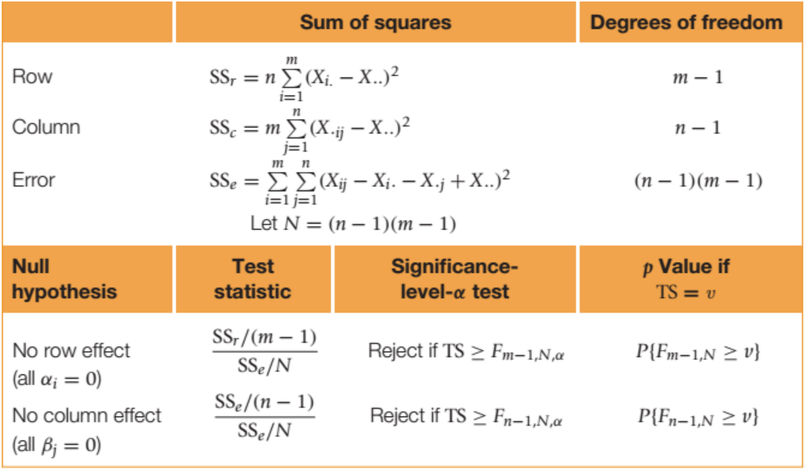
is called the error sum of squares and

$$SS_r = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2$$

is called the row sum of squares.

We then define the test statistic $TS(row) := \frac{SS_r/(m-1)}{SS_e/N}$ and compare the TS(row) with $F_{\alpha,m-1,N}$.

Two-Factor Analysis of Variance: Testing Hypotheses



	Sum of squares	Degrees of freedom
Row	$SS_r = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2$	$m-1$
Column	$SS_c = m\sum_{j=1}^{n}(X_{.ij} - X_{..})^2$	$n-1$
Error	$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2$ Let $N = (n-1)(m-1)$	$(n-1)(m-1)$

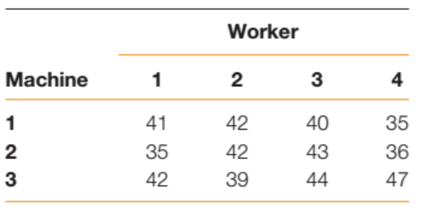
Null hypothesis	Test statistic	Significance-level-α test	p Value if $\mathrm{TS} = v$
No row effect (all $\alpha_i = 0$)	$\dfrac{SS_r/(m-1)}{SS_e/N}$	Reject if $\mathrm{TS} \geq F_{m-1,N,\alpha}$	$P\{F_{m-1,N} \geq v\}$
No column effect (all $\beta_j = 0$)	$\dfrac{SS_e/(n-1)}{SS_e/N}$	Reject if $\mathrm{TS} \geq F_{n-1,N,\alpha}$	$P\{F_{n-1,N} \geq v\}$

Remark: Notation $F_{m-1,N,\alpha} \equiv F_{\alpha,m-1,N}$.

Two-Factor Analysis of Variance: Testing Hypotheses

Example

The following are the numbers of defective items produced by four workers using, in turn, three different machines.

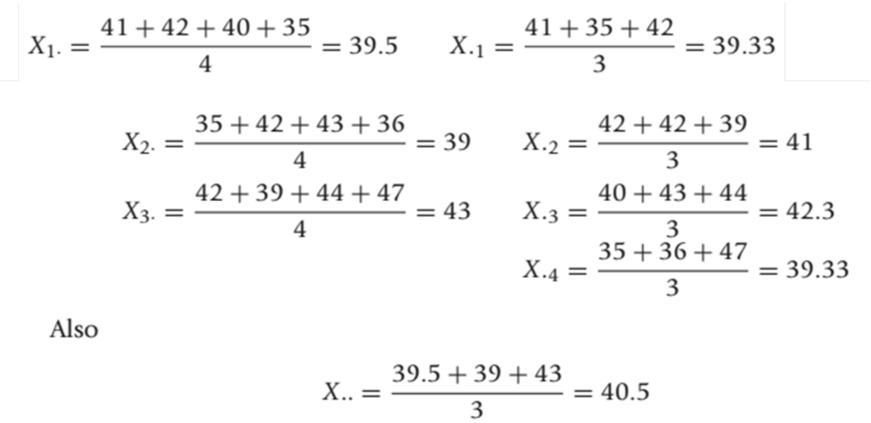


	Worker			
Machine	**1**	**2**	**3**	**4**
1	41	42	40	35
2	35	42	43	36
3	42	39	44	47

Test whether there are significant differences between the machines.

Two-Factor Analysis of Variance: Testing Hypotheses

Solution



$$X_{1\cdot} = \frac{41 + 42 + 40 + 35}{4} = 39.5 \qquad X_{\cdot 1} = \frac{41 + 35 + 42}{3} = 39.33$$

$$X_{2\cdot} = \frac{35 + 42 + 43 + 36}{4} = 39 \qquad X_{\cdot 2} = \frac{42 + 42 + 39}{3} = 41$$

$$X_{3\cdot} = \frac{42 + 39 + 44 + 47}{4} = 43 \qquad X_{\cdot 3} = \frac{40 + 43 + 44}{3} = 42.3$$

$$X_{\cdot 4} = \frac{35 + 36 + 47}{3} = 39.33$$

Also

$$X_{\cdot\cdot} = \frac{39.5 + 39 + 43}{3} = 40.5$$

Solution

We compute

$$SS_r = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2 = 4\left[1^2 + 1.5^2 + 2.5^2\right] = 38$$

$$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2 = 94.05$$

[by summing $nm = 12$ terms, for instance, term for $i = 1, j = 1$ is $(41 - 39.5 - 39.33 + 40.5)^2$]

$$TS(row) = \frac{38/2}{94.05/6} = 1.2$$

We see that $F_{0.05,2,6} = 5.14 > TS(row)$, and so the hypothesis that the mean number of defective items is unaffected by which machine is used is not rejected at the 5 percent level of significance.

Remark

Test whether there are significant differences between the workers:

$$SS_c = m \sum_{i=1}^{n} (X_{.j} - X_{..})^2 = 19.01$$

The test statistic for the hypothesis that there is no column effect is

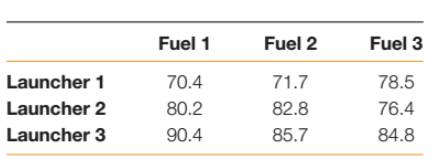
$TS(col.) = \frac{19.010/3}{94.05/6} = 0.40$

We see that $F_{0.05,3,6} = 4.76$, and so the hypothesis that the mean number of defective items is unaffected by which worker is used is also not rejected at the 5 percent level of significance.

Two-Factor Analysis of Variance: Testing Hypotheses

Exercise

An experiment was performed to determine the effect of three different fuels and three different types of launchers on the range of a certain missile. The following data, in the number of miles traveled by the missile, resulted.



	Fuel 1	**Fuel 2**	**Fuel 3**
Launcher 1	70.4	71.7	78.5
Launcher 2	80.2	82.8	76.4
Launcher 3	90.4	85.7	84.8

Find out whether these data imply, at the 5 percent level of significance, that there are differences in the mean mileages obtained by using

(a) Different launchers

(b) Different fuels

Selected exercises

Chapter 10: 1,4,6,11,16,19,20(a)-(b).

THE END! THANK YOU AND GOOD LUCK!