PROBABILITY AND STATISTICS

CHAPTER 7: ANALYSIS OF VARIANCE (ANOVA)

Dr. Phan Thi Huong

HoChiMinh City University of Technology
Faculty of Applied Science, Department of Applied Mathematics
Email: huongphan@hcmut.edu.vn

HCM city — 2021.

Introduction

Example 1

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens arc tested on a laboratory tensile tester, in random order. The data from this experiment are shown in the Table below.

Introduction

Hardwood Concentration (%)	Observations 1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

TABLE 1: Tensile Strength of Paper (psi)

INTRODUCTION

- Question: does the hardwood concentration affect the tensile strength of the bags?

Introduction

- Question: does the hardwood concentration affect the tensile strength of the bags?
- Statistical problem: comparing the tensile strength means between 4 groups of hardwood concentration.

INTRODUCTION

- Question: does the hardwood concentration affect the tensile strength of the bags?
- Statistical problem: comparing the tensile strength means between 4 groups of hardwood concentration.
- The experiment is carried out in random order (a completely randomized design).

INTRODUCTION

- Question: does the hardwood concentration affect the tensile strength of the bags?
- Statistical problem: comparing the tensile strength means between 4 groups of hardwood concentration.
- The experiment is carried out in random order (a completely randomized design).
- $\Rightarrow$ we need a statistical technique called ANOVA.

The Analysis of Variance - The definitions

- The levels of the factor are sometimes called treatments.
- The response for each of the a treatments is a random variable.
- The observed data would appear as shown in the Table below.

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}	$y_{k\cdot}$	$\bar{y}_{k\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

Where

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{i\cdot} = y_{i\cdot}/n, \qquad i = 1, 2, \ldots, k$$

$$y_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}, \quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N, \quad N = kn$$

The Analysis of Variance - The models

Considering the model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{1}$$

where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. In the formula,

- μ is a parameter common to all treatments called the overall mean,
- τ_i is a parameter associated with the ith treatment called the ith treatment effect,
- and ϵ_{ij} a random error component.

The Analysis of Variance - The models

The model is also written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots n \end{cases} \tag{2}$$

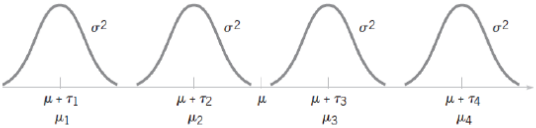
where $\mu_i = \mu + \tau_i$ is the mean of the ith treatment.

- $\Rightarrow$ each treatment defines a population that has mean μ_i.
- $\Rightarrow$ if $\varepsilon_{ij} \sim N(0, \sigma^2)$, each treatment can be thought of as a normal population with mean μ_i and variance σ^2.

The Analysis of Variance - The Assumptions

The Assumptions of the ANOVA for the fixed-effects and single factor model:

- $\sum_{i=1}^{k} \tau_i = 0$
- The populations are normally distributed.
- The population has equal variances, e.i. $\varepsilon_{ij} \sim N(0, \sigma^2)$.
- The samples are random and independent.



The Analysis of Variance - The hypotheses

- The null hypothesis:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0$$

Changing the levels of the factor has no effect on the mean. response.

- The alternative hypothesis:

$$H_1 : \tau_i \neq 0 \quad \text{for at least one } i$$

There exists the difference between the levels of the factor.

THE ANALYSIS OF VARIANCE - THE VARIATION

The ANOVA partitions the total variability in the sample data into two component parts.

The sum of squares identity is

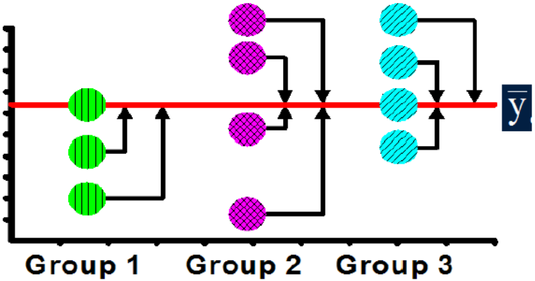
$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

or

$$SST = SSB + SSE$$

The Analysis of Variance - The total variation

- SST describes the total variability in the data: $SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$.

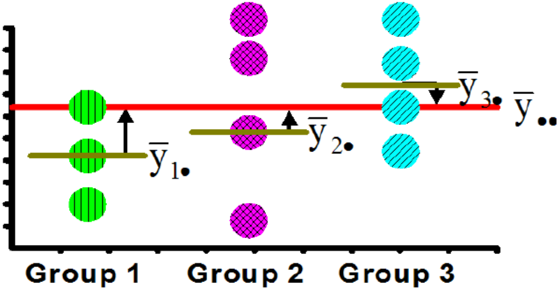


- A computational formula:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N}$$

The Analysis of Variance - The variation between treatments means

- SSB describes the total variability between treatment means: $SSB = n\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

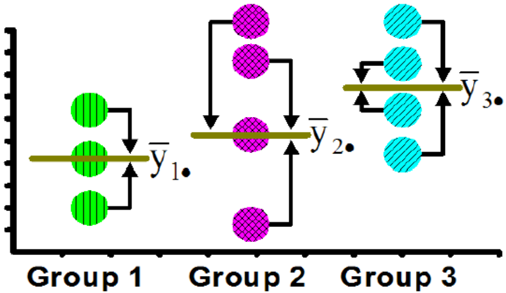


- A computational formula:

$$SSB = \sum_{i=1}^{k}\frac{y_{i\cdot}^2}{n} - \frac{y_{\cdot\cdot}^2}{N}$$

THE ANALYSIS OF VARIANCE - THE VARIATION WITHIN TREATMENTS

- SSE describes the total variability of observation within treatments: $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2$.



- A computational formula:

$$SSE = SST - SSB.$$

The Analysis of Variance - The mean squares

- The mean square for treatment: $MSB = \frac{SSB}{k-1}$
- The mean square for errors: $MSE = \frac{SSE}{k(n-1)}$
- The expected value of the treatment sum of squares is

$$E(SSB) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\tau_i^2$$

- and the expected value of the error sum of squares is

$$E(SSE) = k(n-1)\sigma^2$$

$\Rightarrow$ if H0 is true, MSB is an unbiased estimator of σ^2.

$\Rightarrow$ MSE is an unbiased estimator of σ^2 regardless of whether or not H0 is true.

THE ANALYSIS OF VARIANCE - THE ANOVA F-TEST

$$\begin{cases} H_0 : H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \\ H_1 : \tau_i \neq 0 \quad \text{with at least one i} \end{cases}$$

- The test statistic:

$$F_0 = \frac{MSB}{MSE} = \frac{SSB/(k-1)}{SSE/[k(n-1)]} \tag{3}$$

- F_0 has a Fisher distribution with $(k-1)$ and $k(n-1)$ degrees of freedom, $F_0 \sim f_{k-1,k(n-1)}$.
- Given α, we would reject H0 if $f_0 > f_{k-1,k(n-1),\alpha}$.

The Analysis of Variance - The ANOVA F-test

Source of variation	SS	df	MS	F
Treatments	SSB	$k-1$	MSB	$f_0 = \frac{MSB}{MSE}$
Error	SSE	$k(n-1)$	MSE	
Total	SST	$kn-1$		

Table 2: Analysis of Variance for a Single-Factor Experiment, Fixed-Effects Model

Multiple comparisons methods following ANOVA.

- When the null hypothesis $H_0 : \tau_1 = \tau_2 = \ldots = \tau_k$ is rejected in the ANOVA, we know that some of the treatment or factor-level means are different. However, the ANOVA does not identify which means are different.
- To identify which pairs of treatment means are different, we use multiple comparisons methods. Here we describe a very simple one, Fisher's least significant difference (LSD) method.
- The Fisher LSD method compares all pairs of means with the null hypothesest $H_0 : \mu_i = \mu_j$ (for all $i \neq j$).

Multiple comparisons methods following ANOVA.

Theorem 2.1

If the assumption of ANOVA is adapted, then

$$T = \frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{2MSE}{n}}}$$

follows Student distribution with $k(n-1)$ degrees of freedom.

The Fisher LSD method for confidence intervals.

$100(1-\alpha)\%$ CI for $\mu_i - \mu_j$ is given by

$$\bar{y}_i - \bar{y}_j - t_{k(n-1),\alpha/2}\sqrt{\frac{2MSE}{n}} \le \mu_i - \mu_j \le \bar{y}_i - \bar{y}_j + t_{k(n-1),\alpha/2}\sqrt{\frac{2MSE}{n}}$$

MULTIPLE COMPARISONS METHODS FOLLOWING ANOVA.

THE FISHER LSD METHOD FOR HYPOTHESIS TESTS.

Consider the hypotheses:

$$H0: \mu_i - \mu_j = 0$$
$$H1: \mu_i - \mu_j \neq 0$$

The test statistic value: $t_0 = \dfrac{\bar{y}_i - \bar{y}_j}{\sqrt{\dfrac{2MSE}{n}}} \longrightarrow t - test$

Particularly, $H0$ is rejected when

$$|\bar{y}_i - \bar{y}_j| > t_{\alpha/2}^{k(n-1)} \sqrt{\frac{2MSE}{n}}$$

where $LSD = t_{\alpha/2}^{k(n-1)} \sqrt{\dfrac{2MSE}{n}}$ is called the least significant difference.

EXAMPLE

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens arc tested on a laboratory tensile tester, in random order. The data from this experiment are shown in the Table below.

Example

Hardwood	Observations							
Concentration (%)	1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

TABLE 3: Tensile Strength of Paper (psi)

EXAMPLE

(A) Does the hardwood concentration affect the tensile strength of the bags?
(B) Find the confidence interval for the different means of tensile strength of the bags between two hardwood concentration levels 10 and 15.
(C) Interpret the multiple comparison result.