HCMC University of Technology

Dung Nguyen

Probability and Statistics

Anova

Outline I

The analysis of variance (ANOVA): the analysis of quantitative responses from experimental units.

1. The effects of (five) different brands of gasoline on automobile engine operating efficiency (mpg).
2. The effects of the presence of (four) different sugar solutions (glucose, sucrose, fructose, and a mixture of the three) on bacterial growth.
3. Whether hardwood concentration in pulp (%) has an effect on tensile strength of bags made from the pulp.
4. Whether the color density of fabric specimens depends on the amount of dye used.

Example

A manufacturer of paper used for making grocery bags is interested in improving the product's tensile strength. Product engineering believes 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 by using a pilot plant. All 24 specimens are tested on a laboratory tensile tester in random order.

Hardwood concentration	**Tensile strength**					
5%	7	8	15	11	9	10
10%	12	17	13	18	19	15
15%	14	18	19	17	16	18
20%	19	25	22	23	18	20

Hardwood concentration	Tensile strength						Sum	Average
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	15.99

- The levels of the factor: treatments.
- Each treatment: observations or replicates.
- The runs: in random order.
- Balanced design vs. Unbalanced design

Group 1	x_{11}	x_{12}	...	x_{1J_1}
Group 1	x_{21}	x_{22}	...	x_{2J_2}
...	...	...	...	...
Group I	x_{I1}	x_{I2}	...	x_{IJ_I}

Let $\overline{X}_1, \ldots, \overline{X}_I$ be the sample means of the subpopulations and $\overline{X}$ be the grand mean

$$X_i = \sum_{j=1}^{J_i} X_{ij}, \quad \overline{X}_i = \frac{\sum\limits_{j=1}^{J_i} X_{ij}}{J_i}, \quad \overline{X} = \frac{\sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J_i} X_{ij}}{N}.$$

					Sum	**Average**
Group 1	x_{11}	x_{12}	...	x_{1J_1}	X_1	$\overline{X}_1$
Group 1	x_{21}	x_{22}	...	x_{2J_2}	X_2	$\overline{X}_2$
...	...	...	...	...	...	...
Group I	x_{I1}	x_{I2}	...	x_{IJ_I}	X_I	$\overline{X}_I$
					X	$\overline{X}$

Assumptions (0.1)

The I population or treatment distributions are all normal with the same variance σ^2:

$$X_{ij} \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mathrm{E}(X_{ij}) = \mu_i, \quad \mathrm{V}(X_{ij}) = \sigma^2.$$

$$\boxed{X_{ij} = \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)}$$

Sums of squares

- The total sum of squares

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X})^2.$$

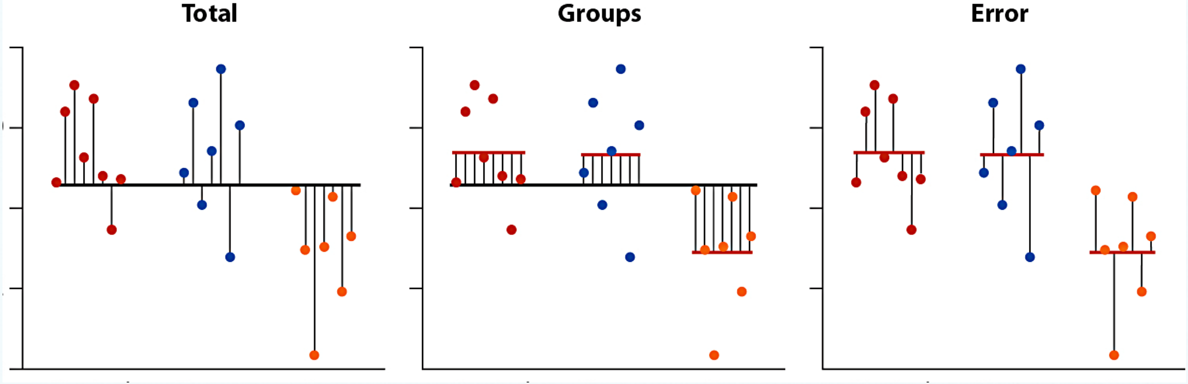
- The treatment sum of squares

$$SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\overline{X}_i - \overline{X})^2 = \sum_{i=1}^{I} J_i(\overline{X}_i - \overline{X})^2.$$

- The error sum of squares

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_i)^2$$

Hardwood concentration	Tensile strength						Sum	Average
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	15.99



The fundamental Anova identity

$$SST = SSTr + SSE \quad \text{and} \quad \mathrm{df}(SST) = \mathrm{df}(SSTr) + \mathrm{df}(SSE).$$

Sum of squares	df	**Definition**	**Computation**
Total (SST)	$N-1$	$\sum_{i,j}(X_{ij}-\overline{X})^2$	$\sum_{i,j} X_{ij}^2 - \frac{X^2}{N}$
Treatment $(SSTr)$	$I-1$	$\sum_{i,j}(\overline{X}_i-\overline{X})^2$	$\sum_{i} \frac{X_i^2}{J_i} - \frac{X^2}{N}$
Error (SSE)	$N-I$	$\sum_{i,j}(X_{ij}-\overline{X}_i)^2$	$SST - SSTr$

Hardwood concentration	Tensile strength						Sum	Average
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	15.99

$$SST = \left(7^2 + 8^2 + \cdots + 20^2\right) - \frac{383^2}{(4)(6)} = 512.9583$$

$$SSTr = \frac{1}{6}\left(60^2 + 94^2 + 102^2 + 127^2\right) - \frac{383^2}{(4)(6)} = 382.7917$$

$$SSE = 512.9583 - 382.7917 = 130.1667.$$

- The mean square for treatment: $MSTr = SSTr/\,\mathrm{df}(SSTr)$.
- The mean square for error: $MSE = SSE/\,\mathrm{df}(SSE)$.

Consider the following statistic

$$F = \frac{MSTr}{MSE} = \frac{\dfrac{SSTr}{I-1}}{\dfrac{SSE}{N-I}}.$$

If H_0 is true then

$$F \sim \mathsf{F}(I-1, N-I).$$

Source of variation	Df	Sum of squares	Mean square	F
Treatment	$I-1$	$SSTr$	$MSTr = \dfrac{SSTr}{I-1}$	$\dfrac{MSTr}{MSE}$
Error	$N-I$	SSE	$MSE = \dfrac{SSE}{N-I}$	
Total	$N-1$	SST		

Rejection region	$F \geq F_{\alpha,I-1,N-I}$

Source of variation	**Df**	**Sum of squares**	**Mean square**	F
Treatment	3	382.79	127.60	19.60
Error	20	130.17	6.51	
Total	23	512.96		

Rejection region	$F \geq 3.01$

Confidence Intervals

$$\boxed{\sigma^2 \approx MSE}$$

Confidence Interval on a Treatment Mean:

$$\mu_i = \overline{X}_i \pm t_{\alpha/2}\,\text{se}, \qquad \text{se} = \sqrt{\frac{MSE}{J_i}}$$

Hardwood concentration	Tensile strength						Sum	Average
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	15.99

$$\text{MOE} = t_{0.025}\,\text{se} = 2.086 * \sqrt{6.51/6} = 2.1728$$

$$\mu_1 = 10.00 \pm 2.1728, \quad \mu_2 = 15.67 \pm 2.1728,$$

$$\mu_3 = 17.00 \pm 2.1728, \quad \mu_4 = 21.17 \pm 2.1728,$$

Multiple Comparisons

$$\mu_i - \mu_k = (\overline{X}_i - \overline{X}_k) \pm LSD, \qquad LSD = t_{\alpha/2}\sqrt{\frac{MSE}{J_i} + \frac{MSE}{J_k}}.$$

Hardwood concentration	Tensile strength						Sum	Average
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	15.99

$$LSD = t_{0.025}\sqrt{\frac{MSE}{6} + \frac{MSE}{6}} = 2.086\sqrt{2(6.51)/6} = 3.07.$$

Therefore, any pair of treatment averages that differs by more than 3.07 implies that the corresponding pair of treatment means are different.

Hardwood concentration	Tensile strength						Sum	Average
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	15.99

The comparisons among the observed treatment averages are as follows (LSD=3.07):

- 4 vs. 1 = 21.17 − 10.00 = 11.17 > 3.07
- 4 vs. 2 = 21.17 − 15.67 = 5.50 > 3.07
- 4 vs. 3 = 21.17 − 17.00 = 4.17 > 3.07
- 3 vs. 1 = 17.00 − 10.00 = 7.00 > 3.07
- 3 vs. 2 = 17.00 − 15.67 = 1.33 < 3.07
- 2 vs. 1 = 15.67 − 10.00 = 5.67 > 3.07

The Random-Effects Model

In Montgomery's book, he describes a single-factor experiment involving the random-effects model in which a textile manufacturing company weaves a fabric on a large number of looms. The company is interested in loom-to-loom variability in tensile strength. To investigate this variability, a manufacturing engineer selects four looms at random and makes four strength determinations on fabric samples chosen

Loom	**Tensile strength**			
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95
4	95	96	99	98

Loom	Tensile strength				Sum	Average
1	98	97	99	96	390	97.5
2	91	90	93	92	366	91.5
3	96	95	97	95	383	95.8
4	95	96	99	98	388	97.0
					1527	95.45

Source of variation	Df	Sum of squares	Mean square	F
Loom	3	89.188	29.729	16.183
Error	12	22.045	1.837	(> 5.953)
Total	15	111.938		