Analysis of Variance (ANOVA)

Compare several means

Radu Trîmbiţaş

1 Analysis of Variance for a One-Way Layout

1.1 One-way ANOVA

Analysis of Variance for a One-Way Layout

- procedure for one-way layout
- Suppose k samples from normal populations with mean μ_1, μ_2, $\ldots$, μ_k, and common variance σ^2. Sample sizes n_i for population i, for $i = 1, 2, \ldots, k$, could be different. The total number of observations in the experiment is $n = n_1 + n_2 + \cdots + n_k$.
- Y_{ij} the response for the jth experimental unit in the ith sample and let $Y_{i\bullet}$ and $\overline{Y}_{i\bullet}$ be the total and the average for the n_i responses in the ith sample. The dot in the second position in the subscript of $Y_{i\bullet}$ means summation over all values of missing subscript — j, in this case. Similarly, subscripts of $\overline{Y}_{i\bullet}$ indicate the mean for the ith sample. Hence, for $i = 1, 2, \ldots, k$,

  $$Y_{i\bullet} = \sum_{j=1}^{n_i} Y_{ij} \text{ şi } \overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} Y_{i\cdot}.$$

  These notations will simplify the description of SSs.
- We have

  $$TotalSS = SST + SSE$$

  (proof later), where

  $$TotalSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - CM$$

  $$CM = \frac{1}{n}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_j} Y_{ij}\right)^2 = n\overline{Y}^2,$$

(CM denotes *correction for the mean*),

$$\begin{aligned} SST &= \sum_{i=1}^{k} n_i \left(\overline{Y}_{i\bullet} - \overline{Y}\right)^2 = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{n_i} - CM, \\ SSE &= TotalSS - SST. \end{aligned}$$

- Although SSE coul be computed by subtraction, it is interesting to see that SSE is the pooled sum of squares for all k samples and is

$$\begin{aligned} SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 = \\ = \sum_{i=1}^{k} \left(n_i - 1\right) S_i^2, \end{aligned}$$

where

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2.$$

- SSE depends only on sample variances S_i^2, for $i = 1, 2, \dots, k$. Since S_i^2 are unbiased estimators for $\sigma_i^2 = \sigma^2$ with $n_i - 1$ dfs, an unbiased estimator for σ^2 with $n_1 + n_2 + \cdots + n_k - k = n - k$ dfs is given by

$$S^2 = MSE = \frac{SSE}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)} = \frac{SSE}{n - k}. \qquad (1)$$

- Since

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{Y}_{i\cdot},$$

it follows SST is a function of only sample means $\overline{Y}_{i\cdot}$, for $i = 1, 2, \dots, k$. MST has $(k-1)$ dfs—i.e. #means minus 1 and

$$MST = \frac{SST}{k - 1}. \qquad (2)$$

- To test the null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

against the alternative that at least one of the equalities does not hold, we compare MST with MSE, using the F statistic based on $\nu_1 = k - 1$ and $\nu_2 = n - k$ numerator and denominator dfs, respectively.

- The null hypothesis will be rejected if

$$F = \frac{MST}{MSE} > F_{\nu_1,\nu_2,\alpha},$$

  where $F_{\nu_1,\nu_2,\alpha}$ is the critical value for F test at level α. Under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, F posesses a F distribution with $k-1$ dfs at numerator and $n-k$ dfs at denominator, respectively.

Assumptions underlying ANOVA F test

- The assumptions underlying the ANOVA F tests deserve particular attention.
- Independent random samples are assumed to have been selected from the k populations.
- The k populations are assumed to be normally distributed with variances $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ and means $\mu_1, \mu_2, \ldots, \mu_k$.
- Moderate departures from these assumptions will not seriously affect the properties of the test. This is particularly true of the normality assumption.
- The assumption of equal population variances is less critical if the sizes of the samples from the respective populations are all equale ($n_1 = n_2 = \cdots = n_k$).
- A one-way layout with equal numbers of observations per treatment is said to be *balanced*.

Example

Example 1. Four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. As a result of dropouts from the experimental groups (due to sickness, transfer, etc.), the number of students varied from group to group. Do the data shown in Table 1 present sufficient evidence to indicate a difference in mean achievement for the four teaching techniques?

Solution. See `anova/studentianova.pdf` □

1.2 ANOVA Table

ANOVA Table

- The calculations for an ANOVA are usually displayed in an ANOVA (or AOV) table.

	1	2	3	4
	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	
	69	79	76	
		90		
$y_{i\cdot}$	454	549	425	351
n_i	6	7	6	4
$\overline{y}_{i\cdot}$	75.67	78.43	70.83	87.75

Table 1: Data for Example 1

Source	df	SS	MS	F
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}\right)^2$		

Table 2: A one-way ANOVA table

- The table for RBD design for comparing k treatment means is shown in Table 2. The first column shows the source associated with each sum of squares; the second column gives the respective degrees of freedom; the third and fourth columns give the sums of squares and mean squares, respectively. A calculated value of F, comparing MST and MSE, is usually shown in the fifth column.

- Notice that $SST + SSE = TotalSS$ and that the sum of the degrees of freedom for treatments and error equals the total number of degrees of freedom.

The ANOVA table for Example 1, shown in Table 3, gives a compact presentation of the appropriate computed quantities for the analysis of variance.

1.3 A Statistical Model for the One-Way Layout

A Statistical Model for the One-Way Layout

Source	df	SS	MS	F
Treatments	3	712.6	237.5	3.77
Error	19	1196.6	63.0	
Total	22	1909.2		

Table 3: ANOVA table for Example 1

- Y_{ij} RVs with values y_{ij}, for $i = 1, 2, \ldots, k$ and $j = 1, 2, ..., n_i$. $Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$ independen, for $i = 1, 2, \ldots, k$ and $j = 1, 2, ..., n_i$. Consider a random sample from population i and write

$$Y_{ij} = \mu_i + \varepsilon_{ij} \Longleftrightarrow \varepsilon_{ij} = Y_{ij} - \mu_i, \qquad j = 1, 2, \ldots, n_i. \tag{3}$$

  where $\varepsilon_{ij} \sim N(0, \sigma^2)$, independent

- The error terms simply represent the difference between the observations in each sample and the corresponding population means.

- Consider means μ_i, for $i = 1, 2, \ldots, k$,

$$\mu_i = \mu + \tau_i \qquad \text{where } \tau_1 + \tau_2 + \cdots + \tau_k = 0.$$

- Notice that $\sum_{i=1}^{k} \mu_i = k\mu + \sum_{i=1}^{k} \tau_i = k\mu$, so $\mu = k^{-1} \sum_{i=1}^{k} \mu_i$ is the average of the k population means (the μ_i-values). For these reason μ is called *global mean*.

- For $i = 1, 2, \ldots, k$, $\tau_i = \mu_i - \mu$ quantifies the difference between the mean for population i and the overall mean, τ_i *effect of treatment* (or population) i.

- **The model** for one-way ANOVA

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, k, \; j = 1, 2, ..., n_i$$

  where

$Y_{ij} =$ the jth observation from population (treatment) i,

$\mu =$ the overall mean,

$\tau_i =$ the nonrandom effect of treatment i, where $\sum_{i=1}^{k} \tau_i = 0$,

$\varepsilon_{ij} =$ random error terms such that $\varepsilon_{ij} \sim N(0, \sigma^2)$, independent

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ can be restated as

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

  and $H_a : \exists (i, i'), i \neq i', \mu_i \neq \mu_{i'} \Longleftrightarrow H_a : \exists i, 1 \leq i \leq k, \tau_i \neq 0$.

- Test statistic

$$F = \frac{MST}{MSE}$$

  MST and MSE given by (2), (1)

- Rejection region

$$F > F_{k-1, n-k, \alpha}$$

1.4 Proof of Additivity of the Sums of Squares and E (MST) for a One-Way Layout

Proof of Additivity of the Sums of Squares and E (MST) for a One-Way Layout

- For one-way layout we have

$$TotalSS = SST + SSE$$

- Thus

$$\begin{aligned} TotalSS &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}+\overline{Y}_{i\bullet}-\overline{Y}\right)^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(Y_{ij}-\overline{Y}_{i\bullet}\right)+\left(\overline{Y}_{i\bullet}-\overline{Y}\right)\right]^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2+2\left(Y_{ij}-\overline{Y}_{i\bullet}\right)\left(\overline{Y}_{i\bullet}-\overline{Y}\right)\right. \\ &\left. +\ \left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2\right] \end{aligned}$$

- Summing first over j, we obtain

$$\begin{aligned} TotalSS = \sum_{i=1}^{k}&\left[\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2+2\left(\overline{Y}_{i\bullet}-\overline{Y}\right)\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)\right. \\ &\left.+n_i\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2\right], \end{aligned}$$

where

$$\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right) = \overline{Y}_{i\bullet} - n_i\overline{Y}_{i\bullet} = 0.$$

- Then, summing over i, one obtains

$$\begin{aligned} TotalSS &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i\bullet}\right)^2+\sum_{i=1}^{k}n_i\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2 \\ &= SSE + SST. \end{aligned}$$

Proof of the additivity of the ANOVA sums of squares for other experimental designs can be obtained in a similar manner although the procedure is often tedious.

- We now proceed with the derivation of the expected value of MST for a one-way layout (including a completely randomized design). Using the statistical model for the one-way layout presented in Section 3, it follows that

$$\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mu + \tau_i + \varepsilon_{ij}\right) = \mu + \tau_i + \overline{\varepsilon}_i,$$

where $\overline{\varepsilon}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}$.

- Since ε_{ij} are independent RVs with $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma^2$, we have $E(\overline{\varepsilon}_i) = 0$ and $V(\overline{\varepsilon}_i) = \sigma^2/n_i$.
- Analogously, $\overline{Y}$ is given by

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mu + \tau_i + \varepsilon_{ij}\right) = \mu + \overline{\tau} + \overline{\varepsilon}$$

where

$$\overline{\tau} = \frac{1}{n}\sum_{i=1}^{k} n_i\tau_i \qquad \text{and } \overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}.$$

- τ_i constants, for $i = 1, 2, \ldots, k \Longrightarrow \overline{\tau}$ constant $\Longrightarrow E(\overline{\varepsilon}) = 0$ and $V(\overline{\varepsilon}) = \sigma^2/n$.

$$\begin{aligned} MST &= \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\overline{Y}_{i\bullet} - \overline{Y}\right)^2 \\ &= \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\tau_i + \overline{\varepsilon}_i - \overline{\tau} - \overline{\varepsilon}\right)^2 \\ &= \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\tau_i - \overline{\tau}\right)^2 + \frac{1}{k-1}\sum_{i=1}^{k} 2n_i\left(\tau_i - \overline{\tau}\right)\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right) \\ &+ \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right)^2. \end{aligned}$$

Again, τ_i constants, for $i = 1, 2, \ldots, k \Longrightarrow \overline{\tau}$ constant $\Longrightarrow E(\varepsilon_{ij}) = E(\overline{\varepsilon}_i) = E(\overline{\varepsilon}) = 0$; it results

$$E(MST) = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\tau_i - \overline{\tau}\right)^2 + \frac{1}{k-1}E\left[\sum_{i=1}^{k} n_i\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right)^2\right].$$

- Notice that

$$\begin{aligned} \sum_{i=1}^{k} n_i\left(\overline{\varepsilon}_i - \overline{\varepsilon}\right)^2 &= \sum_{i=1}^{k}\left(n_i\overline{\varepsilon}_i^2 - 2n_i\overline{\varepsilon}_i\overline{\varepsilon} + n_i\overline{\varepsilon}^2\right) \\ &= \sum_{i=1}^{k} n_i\overline{\varepsilon}_i^2 - 2n\overline{\varepsilon}^2 + n\overline{\varepsilon}^2 = \sum_{i=1}^{k} n_i\overline{\varepsilon}_i^2 - n\overline{\varepsilon}^2 \end{aligned}$$

- Since $E(\bar{\varepsilon}_i) = 0$ and $V(\bar{\varepsilon}_i) = \sigma^2/n_i$, it follows $E(\bar{\varepsilon}_i^2) = \sigma^2/n_i$, for $i = 1, 2, \ldots, k$. Similarly, $E(\bar{\varepsilon}^2) = \sigma^2/n$, and hence,

$$\begin{aligned} E\left[\sum_{i=1}^{k} n_i \left(\bar{\varepsilon}_i - \bar{\varepsilon}\right)^2\right] &= \sum_{i=1}^{k} n_i E\left(\bar{\varepsilon}_i^2\right) - nE\left(\bar{\varepsilon}^2\right) = k\sigma^2 - \sigma^2 \\ &= (k-1)\sigma^2. \end{aligned}$$

  Summarizing, we obtain

$$E(MST) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\tau_i - \bar{\tau}\right)^2, \text{ where } \bar{\tau} = \frac{1}{n}\sum_{i=1}^{k} \tau_i.$$

- Under hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$, it follows that $\bar{\tau} = 0$, and hence, $E(MST) = \sigma^2$. Thus, when H_0 is true, MST/MSE is the ratio of two unbiased estimators for σ^2. If there exists an i, $1 \leq i \leq k$, such that $H_a : \tau_i \neq 0$ is true, the quantity

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\tau_i - \bar{\tau})^2$$

  is strictly positive and MST is a positively biased estimator for σ^2.

1.5 Estimation in the One-Way Layout

Estimation in the One-Way Layout

- Confidence intervals for a single treatment mean and for the difference between a pair of treatment means based on data obtained in a one-way layout are analogous to classical estimations, with the difference that one-way ANOVA estimations uses MSE for σ^2.

- CIs for for the mean of treatment i or the difference between the means for treatments i and i' are, respectively:

$$\bar{Y}_{i\bullet} \pm t_{\alpha/2, n-k} \frac{S}{\sqrt{n_i}}$$

  and

$$\left(\bar{Y}_{i\bullet} - \bar{Y}_{i'\bullet}\right) \pm t_{\alpha/2, n-k} S \frac{1}{\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}}$$

  where

$$S = \sqrt{S^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n_1 + n_2 + \cdots + n_k - k}}.$$

- The confidence intervals just stated are appropriate for a single treatment mean or a comparison of a pair of means selected prior to observation of the data. These intervals are likely to be shorter than the corresponding classical intervals, because the values of $t_{\alpha/2}$ are based on $n - k$ dfs instead of $n_i - 1$ or $n_i + n_{i'} - 2$, respectively. The stated confidence coefficients are appropriate for a single mean or difference in two means identified prior to observing the actual data. If we were to look at the data and always compare the populations that produced the largest and smallest sample means, we would expect the difference between these sample means to be larger than for a pair of means specified to be of interest before observing the data.

Examples

Examples 2. Find a 95% CI for the mean score for teaching technique 1, Example 1. Find a 95% confidence interval for the difference in mean score for teaching techniques 1 and 4, Example 1.

Solution. The 95% CI for technique 1 is

$$\overline{Y}_{1\bullet} \pm t_{0.025,19} \frac{S}{\sqrt{n_i}}$$

where $t_{0.025,19}$ is the t-quantile for $\alpha = 0.025$ and $n - k = 19$ dfs;

$$75.67 \pm (2.093)\sqrt{\frac{63}{6}} = 75.67 \pm 6.7821$$

The 95% CI for $(\mu_1 - \mu_4)$ is

$$(\overline{Y}_{1\bullet} - \overline{Y}_{4\bullet}) \pm (2.093)(7.94)\sqrt{1/6 + 1/4} = -12.08 \pm 10.727,$$

that is $(-22.81, -1.35)$. At a confidence level of 95% we concludes that $\mu_4 > \mu_1$. See *anova/ anovaex13_ 3. pdf* □

2 ANOVA for the Randomized Block Design

2.1 A Statistical Model for the Randomized Block Design

A Statistical Model for the Randomized Block Design

- The randomized block design is a design for comparing k treatments using b blocks.

- The blocks are selected so that, hopefully, the experimental units within each block are essentially homogeneous. The treatments are randomly assigned to the experimental units in each block in such a way that each treatment appears exactly once in each of the b blocks.

- Thus, the total number of observations obtained in a randomized block design is $n = bk$.
- Implicit in the consideration of a randomized block design is the presence of two qualitative independent variables, “blocks” and “treatments.”
- **Statistical Model for a Randomized Block Design**

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, \; i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, b$$

- where

$Y_{ij} =$ the observation on treatment i in block j,

$\mu =$ the overall mean,

$\tau_i =$ the nonrandom effect of treatment i, where $\sum_{i=1}^{k} \tau_i = 0$,

$\beta_j =$ the nonrandom effect of block j, where $\sum_{j=1}^{b} \beta_j = 0$,

$\varepsilon_{ij} =$ random error terms such that ε_{ij} are independent normally distributed random variables with $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma^2$.

- Notice that μ, τ_1, τ_2, ..., τ_k and β_1, β_2, ..., β_b are unknown constants.
- *fixed block effects model* (there exists random block effects models, we don’t consider them here)
- For observation Y_{ij} in treatment i, block j, $E(Y_{ij}) = \mu + \tau_i + \beta_j$ and $V(Y_{ij}) = \sigma^2$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, b$.
- Two observations which received treatment i have means that differ only by the difference of the block effects. If $j \neq j'$,

$$E(Y_{ij}) - E(Y_{ij'}) = \mu + \tau_i + \beta_j - (\mu + \tau_i + \beta_{j'}) = \beta_j - \beta_{j'}.$$

- Similarly, two observations that are taken from the same block have means that differ only by the difference of the treatment effects. If $i \neq i'$,

$$E(Y_{ij}) - E(Y_{i'j}) = \mu + \tau_i + \beta_j - (\mu + \tau_{i'} + \beta_j) = \tau_i - \tau_{i'}.$$

- Observations that are taken on different treatments and in different blocks have means that differ by the difference in the treatment effects plus the difference in the block effects because, if $i \neq i'$ and $j \neq j'$,

$$\begin{aligned} E(Y_{ij}) - E(Y_{i'j'}) &= \mu + \tau_i + \beta_j - (\mu + \tau_{i'} + \beta_{j'}) \\ &= (\tau_i - \tau_{i'}) + (\beta_j - \beta_{j'}). \end{aligned}$$

The Analysis of Variance for a Randomized Block Design

- For a randomized block design involving b blocks and k treatments, we have the following sums of squares:

$$\begin{aligned} TotalSS &= \sum_{i=1}^{k}\sum_{j=1}^{b}\left(Y_{ij}-\overline{Y}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}^2 - CM \\ &= SSB + SST + SSE, \end{aligned}$$

  where

$$\begin{aligned} SSB &= k\sum_{j=1}^{b}\left(\overline{Y}_{\bullet j}-\overline{Y}\right)^2 = \sum_{j=1}^{b}\frac{\overline{Y}_{\bullet j}^2}{k} - CM \\ SST &= b\sum_{i=1}^{k}\left(\overline{Y}_{i\bullet}-\overline{Y}\right)^2 = \sum_{i=1}^{k}\frac{\overline{Y}_{i\bullet}^2}{b} - CM \\ SSE &= TotalSS - SSB - SST. \end{aligned}$$

- In the preceding formulas,

$$\overline{Y} = (\text{average of all } n = bk \text{ observations}) = \frac{1}{bk}\sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}$$

  and

$$CM = \frac{(\text{total of all observations})^2}{n} = \frac{1}{bk}\left(\sum_{i=1}^{k}\sum_{j=1}^{b} Y_{ij}\right)^2.$$

- Table 4 is an ANOVA table for the randomized block design.

- To test the null hypothesis that there is no difference in treatment means, we use the F statistic

$$F = \frac{MST}{MSE}$$

  and reject the null hypothesis if $F > F_{\alpha,\nu_1,\nu_2}$, where F_{α,ν_1,ν_2} is the α-quantile of an F distribution with $\nu_1 = (k-1)$ dfs at numerator and $\nu_2 = (n-b-k+1)$ dfs at denominator, respectively.

- Blocking can be used to control for an extraneous source of variation (the variation between blocks). In addition, with blocking, we have the opportunity to see whether evidence exists to indicate a difference in the mean response for blocks.

- Under the null hypothesis that there is no difference in mean response for blocks (that is, $\beta_j = 0$, for $j = 1, 2, \ldots, b$), the mean square for blocks (MSB) provides an unbiased estimator for σ^2 based on $(b-1)$ dfs.

Source	df	SS	MS
Blocks	$b-1$	SSB	$\frac{SSB}{b-1}$
Treatments	$k-1$	SST	$\frac{SST}{k-1}$
Error	$n-b-k+1$	SSE	MSE
Total	$n-1$	$TotalSS$	

Table 4: ANOVA table for the randomized block design

- Where real differences exist among block means, MSB will tend to be inflated in comparison with MSE, and

  $$F = \frac{MSB}{MSE}$$

  provides a test statistic. As in the test for treatments, the rejection region for the test is

  $$F > F_{\alpha,\nu_1,\nu_2},$$

  where F_{α,ν_1,ν_2} is the α-quantile of an F distribution with $\nu_1 = b-1$ and $\nu_2 = n-b-k+1$ numerator and denominator degrees of freedom, respectively.

Example

Example 3. A stimulus–response experiment involving three treatments was laid out in a randomized block design using four subjects. The response was the length of time until reaction, measured in seconds. The data, arranged in blocks, are shown in Figure 1. The treatment number is circled and shown above each observation. Do the data present sufficient evidence to indicate a difference in the mean responses for stimuli (treatments)? Subjects? Use $\alpha = .05$ for each test and give the associated p-values.

Solution

$$CM = \frac{total^2}{n} = \frac{21.2^2}{12} = 37.45$$

$$TotalSS = \sum_{j=1}^{4}\sum_{i=1}^{3}\left(y_{ij}-\overline{y}\right)^2 = \sum_{j=1}^{4}\sum_{i=1}^{3} y_{ij}^2 - CM = 46.86 - 37.45 = 9.41,$$

$$SSB = \sum_{j=1}^{4}\frac{Y_{\bullet j}^2}{3} - -CM = 40.93 - 37.45 = 3.48,$$

$$SST = \sum_{i=1}^{3}\frac{Y_{i\bullet}^2}{4} - CM = 42.93 - 37.45 = 5.48,$$

$$SSE = TotalSS - SSB - SST = 9.41 - 3.48 - 5.48 = .45.$$

See *anova/ exanovardb. pdf*

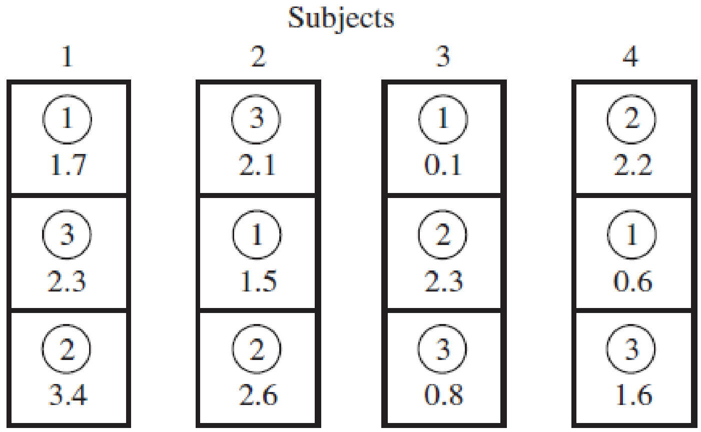

Figure 1: Randomized block design for Example 3

2.2 Estimation in the Randomized Block Design

Estimation in the Randomized Block Design

- The confidence interval for the difference between a pair of treatment means in a randomized block design is completely analogous to that associated with the completely randomized design.
- The $100(1-\alpha)\%$ CI for $\tau_i - \tau'_i$ is

$$(\overline{Y}_{i\bullet} - \overline{Y}_{i'\bullet}) \pm \ t_{\alpha/2,\nu} S \sqrt{\frac{2}{b}},$$

  where $n_i = n_{i'} = b$ and $S = \sqrt{MSE}$. The difference consists of #dfs, which is $\nu = n - b - k + 1 = (b-1)(k-1)$ and S is from ANOVA table for RBD.

Example

Example 4. Construct a 95% confidence interval for the difference between the mean responses for treatments 1 and 2, Example 3.

Solution. The confidence interval for the difference in mean responses for a pair of treatments is

$$(\overline{Y}_{i\bullet} - \overline{Y}_{i'\bullet}) \pm \ t_{\alpha/2,\nu} S \sqrt{\frac{2}{b}},$$

where $t_{\alpha/2,\nu}$ is the quantile of a T distribution for $\alpha = 0.05$ and $\nu = 6$ dfs. For treatments 1 and 2, we have

$$\begin{aligned}(.98 - 2.63) \pm (2.447)(.27)\sqrt{\frac{2}{b}} &= -1.65 \pm .47 \\ &= (-2.12, -1.18).\end{aligned}$$

□

2.3 Sample Size

Selecting the Sample Size

- The method for selecting the sample size for the one-way layout (including the completely randomized) or the randomized block design is an extension of the procedures for two samples.
- restrict to $n_1 = n_2 = \cdots = n_k$, for the treatments of the one-way layout. The number of observations per treatment is equal to the number of blocks b for the randomized block design.
- The problem is to determine n_1 or b
- The determination of sample sizes follows a similar procedure for both designs; we outline a general method.
- First, the experimenter must decide on the parameter (or parameters) of major interest. Usually, this involves comparing a pair of treatment means.
- Second, the experimenter must specify a bound on the error of estimation that can be tolerated.
- Once this has been determined, the next task is to select n_i (the size of the sample from population or treatment i) or, correspondingly, b (the number of blocks for a randomized block design) that will reduce the half-width of the confidence interval for the parameter so that, at a prescribed confidence level, it is less than or equal to the specified bound on the error of estimation.
- It should be emphasized that the sample size solution always will be an approximation because σ is unknown and an estimate for σ is unknown until the sample is acquired.
- The best available estimate for σ will be used to produce an approximate solution.

Example for One-way Layout

Example 5. A completely randomized design is to be conducted to compare five teaching techniques in classes of equal size. Estimation of the differences in mean response on an achievement test is desired correct to within 30 test-score points, with probability equal to .95. It is expected that the test scores for a given teaching technique will possess a range approximately equal to 240. Find the approximate number of observations required for each sample in order to acquire the specified information.

Solution

The confidence interval for the difference between a pair of treatment means is

$$(\overline{Y}_{i\bullet} - \overline{Y}_{i'\bullet}) \pm \; t_{\alpha/2,\nu} S \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

Therefore, we wish to select n_i and $n_{i'}$ so that

$$t_{\alpha/2,\nu} S \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} \leq 30$$

The value of σ is unknown, S is a RV. However, an approximate solution for $n_i = n_{i'}$ can be obtained by conjecturing that the observed value of s will be roughly equal to one-fourth of the range. Thus, $s \approx 240/4 = 60$. The value of $t_{\alpha/2,\nu}$ will be based on $(n_1 + n_2 + \cdots + n_5 - 5)$ dfs and, and for even moderate values of n_i, $t_{.025,\nu}$ will be approximately equal 2.

Then,

$$t_{\alpha/2,\nu} S \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} \approx 2 \cdot 60 \cdot \sqrt{\frac{2}{n_i}} = 30,$$

or

$$n_i = 32, \qquad i = 1, \ldots, 5.$$

Example for Randomized Block Design

Example 6. An experiment is to be conducted to compare the toxic effects of three chemicals on the skin of rats. The resistance to the chemicals was expected to vary substantially from rat to rat. Therefore, all three chemicals were to be tested on each rat, thereby blocking out rat-to-rat differences. The standard deviation of the experimental error was unknown, but prior experimentation involving several applications of a similar chemical on the same type of rat suggested a range of response measurements equal to 5 units. Find a value for b such that the error of estimating the difference between a pair of treatment means is less than 1 unit, with probability equal to .95.

Solution

A very approximate value for s is one-fourth of the range, or $s \approx 1.25$. Then, we wish to select b so that

$$t_{.025,\nu} S \sqrt{\frac{1}{b} + \frac{1}{b}} \leq t_{.025,\nu} S \sqrt{\frac{2}{b}} \leq 1.$$

Since $t_{.025,\nu}$ will depend on the degrees of freedom associated with s^2, which will be $(n - b - k + 1)$, we will use the approximation $t_{.025,\nu} \approx 2$. Then,

$$2 \cdot 1.25 \sqrt{\frac{2}{b}} \leq 1 \Longrightarrow b \approx 13.$$

Approximately thirteen rats will be required to obtain the desired information. Since we will make three observations ($k = 3$) per rat, our experiment will require that a total of $n = bk = 13(3) = 39$ measurements be made. The degrees of freedom associated with the resulting estimate s^2, will be $(n - b - k + 1) = 39 - 13 - 3 + 1 = 24$, based on this solution. Therefore, the guessed value of t would seem to be adequate for this approximate solution.

Comments

- The sample size solutions for Examples 5 and 6 are very approximate and are intended to provide only a rough estimate of sample size and consequent costs of the experiment.

- The actual lengths of the resulting confidence intervals will depend on the data actually observed. These intervals may not have the exact lengths specified by the experimenter but will have the required confidence coefficient.

- If the resulting intervals are still too long, the experimenter can obtain information on σ as the data are being collected and can recalculate a better approximation to the number of observations per treatment (n_i or b) as the experiment proceeds.

3 Simultaneous Confidence Intervals for More Than One Parameter

Simultaneous Confidence Intervals for More Than One Parameter

- The methods devoted to estimations in one-way layout can be used to construct $100(1 - \alpha)\%$ confidence intervals for a single treatment mean or for the difference between a pair of treatment means.

- Suppose that in the course of an analysis we wish to construct several of these confidence intervals. Although it is true that each interval will enclose the estimated parameter with probability $1-\alpha$, what is the probability that all the intervals will enclose their respective parameters?

- We will present a procedure for forming sets of confidence intervals so that the simultaneous confidence coefficient is no smaller than $1-\alpha$ for any specified value of α.

- Suppose that we want to find confidence intervals $I_1, I_2, \dots, Im$ for parameters $\theta_1, \theta_2, \dots, \theta_m$ so that
$$P(\theta_j \in I_j \;\forall\; j = 1,2,...,m) \geq 1-\alpha.$$

- This goal can be achieved by using a simple probability inequality, known as the Bonferroni (or Boole) inequality. For any events $A_1, A_2, \dots, A_m$, we have
$$P(A_1 \cap A_2 \cap \cdots \cap A_m) \geq 1 - \sum_{j=1}^{m} P\left(\overline{A}_j\right).$$
Suppose that $P(\theta_j \in I_j) = 1-\alpha_j$ and let A_j denote the event $\{\theta_j \in I_j\}$. Then,
$$P(\theta_1 \in I_1, \dots, \theta_m \in I_m) \geq 1 - \sum_{j=1}^{m} P\left(\theta_j \notin I_j\right) = 1 - \sum_{j=1}^{m} \alpha_j.$$

- If all α_j's, for $j = 1,2,\dots,m$, are chosen equal to α, we can see that the simultaneous confidence coefficient of the intervals I_j, for $j = 1,2,\dots,m$, could be as small as $(1-m\alpha)$, which is smaller than $(1-\alpha)$ if $m > 1$.

- A simultaneous confidence coefficient of at least $(1-\alpha)$ can be ensured by choosing the confidence intervals I_j, for $j = 1,2,\dots,m$, so that $\sum_{j=1}^{m} \alpha_j = \alpha$. One way to achieve this objective is if each interval is constructed to have confidence coefficient $1-(\alpha/m)$. We apply this technique in the following example.

Example 7. For the four treatments given in Example 1, construct confidence intervals for all comparisons of the form $\mu_i - \mu_{i'}$, with simultaneous confidence coefficient no smaller than .95.

Solution

The appropriate $100(1-\alpha)\%$ confidence interval for a single comparison (say, $\mu_1 - \mu_2$) is
$$(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}) \pm\; t_{\alpha/2,\nu} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Because there are six such differences to consider, each interval should have confidence coefficient $1 - (\alpha/6)$. Thus, the corresponding t-value is $t_{\alpha/2(6)} = t_{\alpha/12}$. Because we want simultaneous confidence coefficient at least .95, the appropriate t-value is $t_{.05/12} = t_{.00417}$. The MSE for the data in Example 1 is based on 19 df, so t-value is 2.9435.

Because $s = \sqrt{MSE} = \sqrt{63} = 7.937$, the interval for $\mu_1 - \mu_2$ among the six with simultaneous confidence coefficient at least .95 is

$$\begin{aligned} \mu_1 - \mu_2 &: (75.67 - 78.43) \pm 2.9435(7.937)\sqrt{\frac{1}{6} + \frac{1}{7}} \\ &= -2.76 \pm 12.996 = (-15.756, 10.236). \end{aligned}$$

Analogously, the entire set of six realized intervals are

$\mu_1 - \mu_2 : -2.76 \pm 12.996 = (-15.756, 10.236)$
$\mu_1 - \mu_3 : 4.84 \pm 13.11 = (-8.27, 17.95)$
$\mu_1 - \mu_4 : -12.08 \pm 14.66 = (-26.74, 2.58)$
$\mu_2 - \mu_3 : 7.60 \pm 12.63 = (-5.03, 20.23)$
$\mu_2 - \mu_4 : -9.32 \pm 14.23 = (-23.55, 4.91)$
$\mu_3 - \mu_4 : -16.92 \pm 14.66. = (-31.58, -2.26)$.
See *anova/studentianovasimCI.pdf*

We emphasize that the technique presented in this section guarantees simultaneous coverage probabilities of at least $1 - \alpha$. The actual simultaneous coverage probability can be much larger than the nominal value $1 - \alpha$. Other methods for constructing simultaneous confidence intervals can be found in the books listed in the bibliography.

4 ANOVA Using Linear Models

ANOVA Using Linear Models

- Linear models can be adapted for use in the ANOVA.

- We illustrate the method by formulating a linear model for data obtained through a completely randomized design involving $k = 2$ treatments.

- Let Y_{ij} denote the random variable to be observed on the jth observation from treatment i, for $i = 1, 2$. Let us define a *dummy*, or *indicator*, variable x as follows:

$$x = \begin{cases} 1, & \text{if the observation is from population 1,} \\ 0, & \text{otherwise.} \end{cases}$$

- Although such dummy variables can be defined in many ways, this definition is consistent with the coding used in SAS and other statistical analysis computer programs.

Bolt of material	Treatments			
	A	B	C	D
I	10.1	11.4	9.9	12.1
II	12.2	12.9	12.3	13.4
III	11.9	12.7	11.4	12.9

Table 5: Data for Example 8

- If we use x as an independent variable in a linear model, we can model Y_{ij} as
$$Y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij},$$
where the error $\varepsilon_{ij} \sim N(0, \sigma^2)$. In this model,
$$\mu_1 = E(Y_{1j}) = \beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1,$$
şi
$$\mu_2 = E(Y_{2j}) = \beta_0 + \beta_1 \cdot 0 = \beta_0.$$
- Thus, it follows that $\beta_1 = \mu_1 - \mu_2$ and a test of the hypothesis $\mu_1 - \mu_2 = 0$ is equivalent to the test that $\beta_1 = 0$.
- The intuition suggests $\widehat{\beta}_0 = \overline{Y}_{2\cdot}$ and $\widehat{\beta}_1 = \overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}$ are good estimators for β_0 and β_1; indeed, it can be shown (proof - homework) that these are the least-squares estimators obtained by fitting the preceding linear model.

Example

Example 8. An experiment was conducted to compare the effects of four chemicals A, B, C, and D on water resistance in textiles. Three different bolts of material I, II, and III were used, with each chemical treatment being applied to one piece of material cut from each of the bolts. The data are given in Table 13.7. Write a linear model for this experiment and test the hypothesis that there are no differences among mean water resistances for the four chemicals. Use $\alpha = .05$.

Solution

In formulating the model, we define β_0 as the mean response for treatment D on material from bolt III, and then we introduce a distinct indicator variable for each treatment and for each bolt of material (block). The model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon,$$

where

$$x_1 = \begin{cases} 1, & \text{if material from bolt I is used,} \\ 0, & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{if material from bolt II is used,} \\ 0, & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{if treatment A is used,} \\ 0, & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1, & \text{if treatment B is used,} \\ 0, & \text{otherwise} \end{cases}$$

$$x_5 = \begin{cases} 1, & \text{if treatment C is used,} \\ 0, & \text{otherwise} \end{cases}$$

We want to test the hypothesis that there are no differences among treatment means, which is equivalent to $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. Thus, we must fit a complete and a reduced model.

For the complete model, we have

$$\mathbf{Y} = \begin{bmatrix} 10.1 \\ 12.2 \\ 11.9 \\ 11.4 \\ 12.9 \\ 12.7 \\ 9.9 \\ 12.3 \\ 11.4 \\ 12.1 \\ 13.4 \\ 12.9 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A little matrix algebra yields, for this complete model,

$$SSE_C = \mathbf{Y}^T\mathbf{Y} - \widehat{\beta}'\mathbf{X}^T\mathbf{Y} = 0.530$$

The relevant reduced model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

and the corresponding **X** matrix consists of only the first three columns of the **X** matrix given for the complete model. We then obtain

$$\widehat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 12.2250 \\ -1.3500 \\ 0.4750 \end{bmatrix}$$

and

$$SSE_R = Y^T Y - \widehat{\beta} X^T Y = 5.7350$$

It follows that the F ratio appropriate to compare these complete and reduced models is

$$F^* = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/\left[n-(k+1)\right]} = \frac{(5.7350 - 0.530)\,/(5-2)}{0.530/(12-(5+1))} = 19.642.$$

We have $\nu_1 = 3$ numerator dfs and $\nu_2 = 6$ denominator dfs, respectively. The associated p-value is $p = P(F_{3,6} > F^*) = 0.02$, hence for $\alpha = 0.05$ we reject the null hypothesis and conclude that the data present sufficient evidence to indicate that differences exist among the treatment means.

See *anova/ anovaexlm. pdf*

5 References

References

References

[1] Box, G. E. P., W. G. Hunter, and J. S. Hunter. 2005. *Statistics for Experimenters*, 2d ed. New York: Wiley Interscience.

[2] Cochran, W. G., and G. Cox. 1992. *Experimental Designs*, 2d ed. New York: Wiley.

[3] Graybill, F. 2000. *Theory and Application of the Linear Model*. Belmont Calif.: Duxbury.

[4] Hicks, C. R., and K. V. Turner. 1999. *Fundamental Concepts in the Design of Experiments*, 5th ed. New York: Oxford University Press.

[5] Hocking, R. R. 2003. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, 5th ed. New York: Wiley Interscience.

[6] Montgomery, D. C. 2006. *Design and Analysis of Experiments*, 6th ed. New York: Wiley.

[7] Scheaffer, R. L., W. Mendenhall, and L. Ott. 2006. *Elementary Survey Sampling*, 6th ed. Belmont Calif.: Duxbury.

[8] Scheffé, H. 2005. *The Analysis of Variance*. New York: Wiley Interscience.