12-Analysis of Variance

ENSY SILVER[1]

Saturday 12th September, 2020

[1]Thanks to my family, my friend and freedom.

1 Variables of different treatment level

Suppose that data from a completely randomized one-factor design will consist of k independent random samples of sizes $n_1, n_2, \cdots, n_k$, the total sample size being denoted $n = \sum_{i=1}^{n} n_i$. We will let Y_{ij} represent the i^{th} observation recorded for the j^{th} level. We define two symbols,

$$T_{.j} = \sum_{i=1}^{n_j} Y_{ij}$$

and

$$T_{..} = \sum_{j=1}^{k} T_{.j}$$

In the rest of this notes, we assume that for all Y_{ij}, it has a normal distribution with μ_j and σ^2.

2 Sum of squares

We define the **treatment sum of squares** (SSTR) by

$$SSTR = \sum_{j=1}^{k} n_j(\bar{Y}_{.j} - \bar{Y}_{..})^2$$

Theorem 2.1. *Let SSTR be the treatment sum of squares defined for k independent random samples of sizes $n_1, n_2, \cdots, n_k$. Then*

$$E(SSTR) = (k-1)\sigma^2 + \sum_{j=1}^{k} n_j(\mu_j - \mu)^2$$

Similar to the previous notes, we can deduce the type of distribution for $SSTR/\sigma^2$.

Theorem 2.2. *When $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true, $SSTR/\sigma^2$ has a χ^2 distribution with $k-1$ degrees of freedom.*

However, the theorem 2.2 requires the property that σ is known. For the case of unknown σ, we define the **error sum of squares**, or SSE:

$$SSE = \sum_{j=1}^{k}(n_j - 1)S_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{.j})^2$$

ans we have the theorem.

Theorem 2.3. *Whether or not $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true,*

1. *SSE/σ^2 has a χ^2 distribution with $n-k$ degrees of freedom.*
2. *SSE and SSTR are independent.*

If we ignore the treatments and consider the data as one sample, then the variation about the parameter μ can be estimated by the double sum

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_{..})^2$$

which is known as the **total sum of squares** and denoted SSTOT.

Theorem 2.4. *If n observations are divided into k samples of sizes $n_1, n_2, \cdots, n_k$,*

$$SSTOT = SSTR + SSE$$

Since $SSTR/\sigma^2$ and SSE/σ^2 are both χ^2 distribution, we can build F distribution and eliminate σ^2.

Theorem 2.5. *Suppose that each observation in a set of k independent random samples is normally distributed with the same variance σ^2. Let $\mu_1, \mu_2, \cdots, \mu_k$ be the true means associated with the k samples. Then*

1. *If $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is true,*

$$F = \frac{SSTR/(k-1)}{SSE/(n-k)}$$

   has an F distribution with $k-1$ and $n-k$ degrees of freedom.

2. *At the α level of significance, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ should be rejected if $F \geq F_{1-\alpha,k-1,n-k}$.*

We define the **mean square for treatments** by

$$MSTR = \frac{SSTR}{k-1}$$

and the **mean square for error** by

$$MSE = \frac{SSE}{n-k}$$

Let $C = T_{..}^2/n$, SSTOT and SSTR has another representation that

$$SSTOT = \sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ij}^2 - C$$

and

$$SSTR = \sum_{j=1}^{k}\frac{T_{.j}^2}{n_j} - C$$

Recall the test of comparing μ_X and μ_Y, we find that when $k = 2$,

$$F = \frac{\frac{nm}{n+m}(\bar{X} - \bar{Y})^2}{\frac{(n+m-2)S_p^2}{n+m-2}} = \frac{(\bar{X} - \bar{Y})^2}{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$$

which indicates that the F test is an extension of the previous two-sample test.

3 Tukey's method

Suppose, for example, we did ten independent tests of the form $H_0 : \mu_i = \mu_j$ versus $H_1 : \mu_i \neq \mu_j$, each at level $\alpha = 0.05$, on a large set of population means. Even though the probability of making a Type I error on any given test is only 0.05, the chances of incorrectly rejecting a true H_0 with at least one of the ten t tests increases dramatically to 0.40.

Addressing that concern, mathematical statisticians have paid a good deal of attention to the so-called multiple comparison problem. Many different procedures, operating under various sets of assumptions, have been developed. All have the objective of keeping the probability of committing at least one Type I error small, even when the number of tests performed is large (or even infinite). In this section, we develop one of the earliest of these techniques, a still widely used method due to John Tukey.

Theorem 3.1. *Let $W_1, \cdots, W_k$ be a set of k independent, normally distributed random variables with mean μ and variance σ^2, let R denote their range:*

$$R = \max W_i - \min W_j$$

Suppose S^2 is based on a χ^2 random variable with v degrees of freedom, independent of the W_i, where $E(S^2) = \sigma^2$. The ***studentized range****, $Q_{k,v}$ is the ratio*

$$Q_{k,v} = \frac{R}{S}$$

If we define $W_t = \bar{Y}_{.t} - \mu_t$, the next theorem is obvious.

Theorem 3.2. *Let $\bar{Y}_{.j}, j = 1, 2, \cdots, k$ be the k sample means in a completely randomized one-factor design. Let $n_j = r$ be the common sample size, and let μ_j be the true means, $j = 1, 2, \cdots, k$. The probability is $1 - \alpha$ that all $\binom{k}{2}$ differences $\mu_i - \mu_j$ will simultaneously satisfies the inequalities*

$$\bar{Y}_{.i} - \bar{Y}_{.j} - D\sqrt{MSE} < \mu_i - \mu_j < \bar{Y}_{.i} - \bar{Y}_{.j} + D\sqrt{MSE}$$

where $D = Q_{\alpha,k,rk-k}/\sqrt{R}$. If, for a given i ans j, zero is not contained in the preceding inequality, $H_0 : \mu_i = \mu_j$ can be rejected in favor of $H_1 : \mu_i \neq \mu_j$, at the α level of significance.

4 Contrast

Theorem 4.1. *Let $\mu_1, \mu_2, \cdots, \mu_k$ denote the true means of k factor levels being sampled. A linear combination, C, of the μ_j is said to be a contrast if the sum of its coefficients is 0. That is, C is a contrast if $C = \sum_{j=1}^{k} c_j \mu_j$, where the c_j are constants such that $\sum_{j=1}^{k} c_j = 0$.*

Tow contrasts

$$C_1 = \sum_{j=1}^{k} c_{1j}\mu_j \quad \text{and} \quad C_2 = \sum_{j=1}^{k} c_{2j}\mu_j$$

are said to be *orthogonal* if

$$\sum_{j=1}^{k} \frac{c_{1j}c_{2j}}{n_j} = 0$$

A set of q contrasts, $\{C_i\}_{i=1}^{q}$ are said to be *mutually orthogonal* if

$$\sum_{j=1}^{k} \frac{c_{sj}c_{tj}}{n_j} = 0 \quad \text{for all} \;\; s \neq t$$

The estimator of contrast C is

$$\hat{C} = \sum_{j=1}^{k} c_j \bar{Y}_{.j}$$

which has the mean

$$E(\hat{C}) = C$$

and the varince

$$\mathrm{Var}(\hat{C}) = \sigma^2 \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Theorem 4.2. *Let $C_i = \sum_{j=1}^{k} x_{ij}\mu_j$ be any contrast. The sum of squares associated with C_i is given by*

$$SS_{C_i} = \frac{\hat{C}_i^2}{\sum_{j=1}^{k} \frac{c_{ij}^2}{n_j}}$$

Theorem 4.3. *Let $\left\{C_i = \sum_{j=1}^{k} c_{ij}\mu_j\right\}_{i=1}^{k-1}$ be a set of $k-1$ mutually orthogonal contrasts. Let $\left\{C_i = \sum_{j=1}^{k} c_{ij}\bar{Y}_{.j}\right\}_{i=1}^{k-1}$ be their estimators. Then*

$$SSTR = SS_{C_1} + SS_{C_2} + \cdots + {}_{C_{k-1}}$$

Theorem 4.4. *Let C be a contrast having the same coefficients as the subhypothesis $H_0 : \sum_{j=1}^{k} c_j \mu_j = 0$, where $\sum_{j=1}^{k} c_j = 0$. Let $n = \sum_{j=1}^{k} n_j$ be the total sample size. Then*

1. *$F = \frac{SS_C/1}{SSE/(n-k)}$ has an F distribution with 1 and $n-k$ degrees of freedom.*

2. *$H_0 : \sum_{j=1}^{k} c_j \mu_j = 0$ shouls be rejected at the α level of significance if $F \geq F_{1-\alpha,1,n-k}$.*

5 Data transformation

Sometimes, a group of data Y_{ij} have different variances, and we need to transform them to W_{ij}, whose variance are same. We build a function A : $A(Y_{ij}) = W_{ij}$, where $\mathrm{Var}(W_{ij}) = c_1^2$. By Taylor's theorem,

$$W_{ij} = A(\mu_j) + (Y_{ij} - \mu_j)A'(\mu_j)$$

and the variance is

$$\mathrm{Var}(W_{ij}) = [A'(\mu_j)]^2 g(\mu_j)$$

For Y_{ij} in the neighborhood of μ_j, it follows that

$$A(Y_{ij}) = c_1 \int \frac{1}{\sqrt{g(y_{ij})}}\, dy_{ij} + c2$$