Topic 22
Analysis of Variance
Comparing Multiple Populations

Outline

Overview

- Two-sample t procedures are designed to compare the means of two populations.
- Our next step is to compare the means of several populations.
- To explain the methodology, we consider the data set gathered from the forests in Borneo.

Figure: Danum Valley Conservation Area

Overview

Example. The data on 30 forest plots in Borneo are the number of trees per plot.

	never logged	logged 1 year ago	logged 8 years ago
n_j	12	12	9
$\bar{y}_j$	23.750	14.083	15.778
s_j	5.065	4.981	5.761

We compute these statistics from the data $y_{11}, \ldots y_{n_1 1}$, $y_{12}, \ldots y_{n_2 2}$ and $y_{13}, \ldots y_{n_3 3}$,

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2.$$

Overview

- The basic question is: Are these means the same (the null hypothesis) or not (the alternative hypothesis)?
- The basic idea of the test is to examine the ratio of $s^2_{between}$, the variance between groups (indicated by the variation in the center lines of the boxes) and $s^2_{residual}$, a statistic that measures the variances within groups.
- If the resulting ratio test statistic is sufficiently large, then we say, based on the data, that the means of these groups are distinct and we reject H_0.

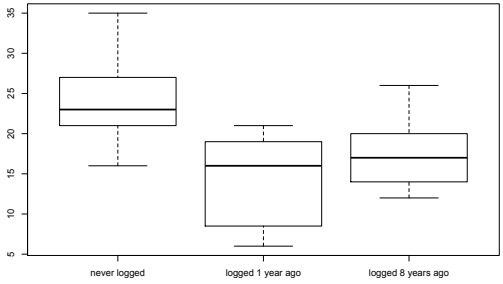

Figure: Side-by-side boxplots of the number of trees per plot.

One Way Analysis of Variance

The hypothesis for one way analysis of variance is

$$H_0 : \mu_j = \mu_k \text{ for all } j, k \quad \text{and} \quad H_1 : \mu_j \neq \mu_k \text{ for some } j, k.$$

The data $\{y_{ij}, 1 \leq i \leq n_j, 1 \leq j \leq q\}$ represents that we have n_j observation for the j-th group and that we have q groups. The total number of observations is denoted by $n = n_1 + \cdots + n_q$. The model is

$$y_{ij} = \mu_j + \epsilon_{ij}$$

where ϵ_{ij} are independent $N(0, \sigma)$ random variables with σ^2 unknown. This allows us to define the likelihood and to use that to determine the analysis of variance F test as a likelihood ratio test.

NB The model for analysis requires a common value σ for all of the observations.

Sample Means

To set up the F statistic, we introduce two types of sample means:

- The within group means is the sample mean inside each of the groups,

$$\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \quad j = 1 \ldots, q.$$

- The mean of the data taken as a whole, known as the grand mean,

$$\overline{\overline{y}} = \frac{1}{n} \sum_{j=1}^{q} \sum_{i=1}^{n_j} y_{ij} = \frac{1}{n} \sum_{j=1}^{q} n_j \bar{y}_j,$$

the weighted average of the $\bar{y}_j$ with weights n_j, the sample size in each group.

Exercise. For the Borneo rain forest example, show that the grand mean is 18.06055.

Sums of Squares

Analysis of variance uses the total sums of squares

$$SS_{total} = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \overline{\overline{y}})^2,$$

the total square variation of individual observations from their grand mean. The test statistic is determined by decomposing SS_{total}. We first rewrite the interior sum as

$$\sum_{i=1}^{n_j} (y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2 = (n_j - 1)s_j^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2.$$

Here, s_j^2 is the unbiased sample variance based on the observations in the j-th group.

Exercise. Show the first equality above. (Hint: Begin with the difference in the two sums.)

Sums of Squares

Summing the expression above over the groups j to yield the decomposition of the variation

$$SS_{total} = SS_{residual} + SS_{between}$$

with

$$SS_{residual} = \sum_{j=1}^{q} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2 = \sum_{j=1}^{q} (n_j - 1) s_j^2 \quad \text{and} \quad SS_{between} = \sum_{j=1}^{q} n_j (\overline{y}_j - \overline{\overline{y}})^2.$$

For the rain forest example, we find that

$$SS_{residual} = (12-1) \cdot 5.065^2 + (12-1) \cdot 4.981^2 + (9-1) \cdot 5.761^2 = 820.6234$$

and

$$SS_{between} = 12 \cdot (23.750 - \overline{\overline{y}})^2 + 12 \cdot (14.083 - \overline{\overline{y}})^2 + 9 \cdot (15.778 - \overline{\overline{y}})^2) = 625.1793$$

Sums of Squares

source of variation	degrees of freedom	sums of squares	mean square
between groups	$q-1$	$SS_{between}$	$s^2_{between} = SS_{between}/(q-1)$
residuals	$n-q$	$SS_{residual}$	$s^2_{residual} = SS_{residual}/(n-q)$
total	$n-1$	SS_{total}	

- The $q-1$ degrees of freedom between groups is derived from the q groups minus 1 degree of freedom used to compute $\overline{\overline{y}}$.
- The $n-q$ degrees of freedom within the groups is derived from the n_j-1 degree of freedom used to compute the variances s_j^2.

Sums of Squares

The analysis of variance information for the Borneo rain forest data is summarized in the table below.

source of variation	degrees of freedom	sums of squares	mean square
between groups	2	625.2	312.6
residuals	30	820.6	27.4
total	32	1445.8	

The F Statistic

The test statistic is

$$F = \frac{s^2_{between}}{s^2_{residual}} = \frac{SS_{between}/(q-1)}{SS_{residual}/(n-q)}.$$

- Under the null hypothesis, F is a constant multiple of the ratio of two independent χ^2 random variables, namely $SS_{between}$ and $SS_{residual}$.
- This ratio is called an F random variable with $q-1$ numerator degrees of freedom and $n-q$ denominator degrees of freedom and written $F_{q-1,n-q}$

F Statistic

For the rain forest data,

$$F = \frac{s^2_{between}}{s^2_{residual}} = \frac{312.6}{27.4} = 11.43.$$

The critical value for an 0.01 level test is 5.390. So, we reject H_0 stating mean number of trees does not depend on logging history.

```
> 1-pf(11.43,2,30)
[1] 0.0002041322
> qf(0.99,2,30)
[1] 5.390346
```

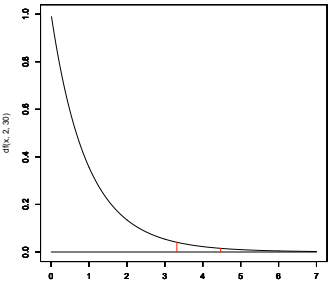

Figure: Upper tail critical values. The density for an $F_{2,30}$ random variable. The indicated values 3.316, 4.470, and 5.390 are critical values for significance levels $\alpha = 0.05$, 0.02, and 0.01, respectively.

Exercise. Use R to determine these critical values.

Confidence Intervals

Confidence intervals are determined using the data from all of the groups as an unbiased estimate $s^2_{residuals} = SS_{residuals}/(n-q)$ for the variance, σ^2. This allows us to increase the degrees of freedom in the t distribution and reduce the margin of error. Thus, the γ-level confidence interval for μ_j is

$$\bar{y}_j \pm t_{(1-\gamma)/2,n-q} s_{residual}/\sqrt{n_j}.$$

The interval for the difference in $\mu_j - \mu_k$ is similar to that for a pooled two-sample t confidence interval,

$$\bar{y}_j - \bar{y}_k \pm t_{(1-\gamma)/2,n-q} s_{residual} \sqrt{\frac{1}{n_j} + \frac{1}{n_k}}.$$

The 95% confidence interval for mean number of trees on a lot logged 1 year ago

$$14.083 \pm 2.042 \frac{\sqrt{27.4}}{\sqrt{12}} = 14.083 \pm 4.714 = (9.369, 18.979).$$

Exercise. Give the 95% confidence interval for the difference in trees between plots never logged plots versus logged 8 years ago.