Topic 20: Analysis of Variance

December 8, 2009

Our next step is to compare the means of several populations. Consider the data set gathered from the forests in Borneo

Example 1 (Rain forest logging)**.** *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	12	12	9
$\bar{x}$	23.750	14.083	15.778
s	5.065	4.981	5.761

One way analysis of variance is a statistical procedure that allows us to test for the differences in two or more independent groups. In the situation above, we have set our design so that the data in each of the three groups is a random sample from within the groups. The basic question is: Are these groups the same (the null hypothesis) or not (the alternative hypothesis)?

The analysis of variance test is a likelihood ratio test. Because all of the basic ideas can be seen in the case of two groups, we begin with a development of this case.

1 Two Sample Procedures

Consider the **two-sided hypothesis**

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2.$$

The data are n_j independent $N(\mu_j, \sigma^2)$ random variables $Y_{j1} \ldots, Y_{jn_j}$ with unknown **common** variance σ^2, $j = 1$ and 2. Thus, the parameter spaces are

$$\Theta = \{(\mu_1, \mu_2, \sigma^2); \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}, \quad \Theta_0 = \{(\mu_1, \mu_2, \sigma^2); \mu_1 = \mu_2, \sigma^2 > 0\}$$

To find the test statistic derived from a likelihood ratio test, we first write the log-likelihood.

$$\ln L(\mu_1.\mu_2, \sigma^2|\mathbf{y}) = -\frac{(n_1 + n_2)}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{1i} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \mu_2)^2\right)$$

By taking partial derivatives with respect to μ_1 and μ_2 we see that with two independent samples, the maximum likelihood estimator for the mean μ_i for each of the samples is the sample mean $\bar{y}_i$.

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}, \quad \hat{\mu}_2 = \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i}.$$

Now differentiate with respect to σ^2

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu_1, \mu_2, \sigma^2|\mathbf{x}) = -\frac{n_1 + n_2}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n_1}(y_{1i} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \mu_2)^2\right).$$

Thus, the maximum likelihood estimator of the variance is the **weighted** average, weighted according to the sample size, of the maximum likelihood estimator of the variance for each of the respective samples.

$$\hat{\sigma}^2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2\right).$$

Now, substitute these values into the likelihood to see that the maximum value for the likelihood is

$$L(\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}^2|\mathbf{x}) = \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}}\exp -\frac{n_1+n_2}{2}$$

Next, for the likelihood ratio test, we find the maximum likelihood under the null hypothesis. In this case the two means have a common value which we shall denote by μ.

$$\ln L(\mu,\sigma^2|\mathbf{x}) = -\frac{(n_1+n_2)}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{1i}-\mu)^2 + \sum_{i=1}^{n_2}(y_{2i}-\mu)^2\right)$$

The μ derivative

$$\frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n_1}(y_{1i}-\mu) + \sum_{i=1}^{n_2}(y_{2i}-\mu)\right)$$

and the maximum likelihood is just the sample mean obtained by considering all of the data being derived from one large sample

$$\hat{\mu} = \overline{\overline{y}} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}y_{1i} + \sum_{i=1}^{n_2}y_{2i}\right).$$

The maximum value for σ^2 in Θ_0, is also the maximum likelihood estimator for the variance obtained by considering all of the data being derived from one large sample.

$$\hat{\sigma}_0^2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(y_{1i}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{\overline{y}})^2\right).$$

We can find that the the maximum value on Θ_0 for the likelihood is

$$L(\hat{\mu},\hat{\sigma}_0^2|\mathbf{x}) = \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}}\exp -\frac{n_1+n_2}{2}.$$

This give a likelihood ratio of

$$\Lambda(\mathbf{y}) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{-(n_1+n_2)/2} = \left(\frac{\sum_{i=1}^{n_1}(y_{1i}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{\overline{y}})^2}{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2}\right)^{-(n_1+n_2)/2}.$$

Traditionally, this is simplified by noticing that

$$\sum_{j=1}^{n_i}(y_{ij}-\overline{\overline{y}})^2 - \sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 = \sum_{j=1}^{n_i}(2y_{ij}-\overline{\overline{y}}-\bar{y}_i)(-\overline{\overline{y}}+\bar{y}_i) = n_i(\overline{\overline{y}}-\bar{y}_i)^2.$$

and

$$\sum_{i=1}^{n_1}(y_{1i}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{\overline{y}})^2 = n_1(\overline{\overline{y}}-\bar{y}_1)^2 + n_2(\overline{\overline{y}}-\bar{y}_2)^2 + \sum_{i=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2j}-\bar{y}_2)^2.$$

In words, the the sums of squares of the difference of an observation from the overall mean $\overline{\overline{y}}$ is the sum of two sources. The first is the sums of squares the difference of the average of the group mean and the overall mean, SS^2_{between}. The second is the he sums of squares the difference of the individual differences and the group mean, $SS^2_{\text{residuals}}$.

$$SS^2_{\text{total}} = SS^2_{\text{residual}} + SS^2_{\text{between}}$$

Now, the likelihood ratio reads

$$\Lambda(\mathbf{y}) = \left(1 + \frac{n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2}{\sum_{j=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2i} - \bar{y}_2)^2}\right)^{-(n_1+n_2)/2} = \left(1 + \frac{SS^2_{\text{between}}}{SS^2_{\text{residuals}}}\right)^{-(n_1+n_2)/2}$$

The critical region $\Lambda(\mathbf{y}) \leq \lambda_0$ is equivalent to

$$\frac{SS^2_{\text{between}}}{SS^2_{\text{residuals}}} = \frac{n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2}{\sum_{j=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2i} - \bar{y}_2)^2} \geq c$$

Thus, we reject if the variation between thr groups is large compared to the variation within the groups.

Exercise 2. *For an α level test, show that the test above is equivalent to*

$$|T(\mathbf{y})| > t_{\alpha/2, n_1+n+2}.$$

where

$$T(\mathbf{y}) = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

and s_p is the standard deviation of the data pooled into one sample.

2 One Way Analysis of Variance

For one way analysis of variance, we expand to more than three groups and ask whether or not all the groups are the same. The hypothesis in this case 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 $\{y_{ij}, 1 \leq i \leq q, 1 \leq j \leq n_j\}$ represents that we have n_i observation for the i-th group and that we have q groups. The model is

$$y_{ij} = \mu_i + \epsilon_{ij}.$$

where ϵ_{ij} are independent $N(0, \sigma^2)$ random variables with σ^2 unknown. The total number of observations $n = n_1 + \cdots + n_q$. The within group mean

$$\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_j} y_{ij}.$$

is the maximum likelihood estimator $\widehat{\mu}_i$ for μ_i.

The overall mean

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

is the maximum likelihood estimator $\widehat{\mu}$

Write the total **sums of squares**

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

then

$$\widehat{\sigma^2} = \frac{1}{n}SS^2_{\text{total}}.$$

The interior sum can be written

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

yielding as we saw in the two sample case

$$SS^2_{\text{total}}=SS^2_{\text{residual}}+SS^2_{\text{between}}$$

with

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

This gives the general form for one-way analysis of variance.

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

The test statistic

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

has, under the null hypothesis, an F distribution with $q-1$ numerator degrees of freedom and $n-q$ denominator degrees of freedom.

The analysis of variance for the Borneo rain forest example has

source of variation	degrees of freedom	sums of squares	mean square
between samples	2	625.2	312.6
residuals	30	820.7	27.4
total	32	1445.9	

The value of the F statistic is 11.43 and the P-value is 0.0002. The critical value for a 0.05 level test is 5.390. So, we do reject the null hypothesis that mean number of tress does not depend on the history of logging.

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

The confidence intervals for the difference in $\mu_j-\mu_k$ and is given by

$$\bar{y}_j-\bar{y}_k\pm t^*_{\alpha/2,n-q}s_{\text{residual}}\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}.$$

In this case the 95% confidence interval for the difference in never logged versus logged 8 years ago has $t^*_{0.025,30}=2.042$ and the confidence interval is

$$7.972\pm 4.714=(3.528,12.686)$$