Topic 22: Analysis of Variance*

December 5, 2011

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_i	12	12	9
$\bar{y}_i$	23.750	14.083	15.778
s_i	5.065	4.981	5.761

We compute these statistics from the data $y_{11}, \dots y_{1n_1}$, $y_{21}, \dots y_{2n_2}$ and $y_{31}, \dots y_{2n_2}$

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

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 basic idea of the test is to compare the variance between the groups 1, 2, and 3 with the variance s_1^2, s_2^2, and s_3^2 within the groups. If the resulting ratio test statistic is sufficiently large, then we say, based on the data, that the groups are distinct and we are able to reject the null hypothesis. As we have seen before, this statement will be the consequence of the value of a test statistics - in this case the F statistic. A sufficiently large value of this statistic is evidence against the null hypothesis. As we shall see, the distribution of this test statistic will depend on the number of groups (3 in the example above) and the number of total observations (33 in the example above).

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 in this case that will lead to the F statistic.

1 Two Sample Procedures

Our hypothesis test is based on two independent samples of normal random variables. The data are n_j independent $N(\mu_j, \sigma^2)$ random variables $Y_{j1} \dots, Y_{jn_j}$ with unknown **common** variance σ^2, $j = 1$ and 2. The assumption of a common hypothesis is critical to the ability to compute the test statistics.

Consider the **two-sided hypothesis**

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

Thus, the parameter space is

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

For the null hypothesis, the possible parameter values are

$$\Theta_0 = \{(\mu_1, \mu_2, \sigma^2); \mu_1 = \mu_2, \sigma^2 > 0\}$$

To find the test statistic derived from a likelihood ratio test, we first write the likelihood and its logarithm based on observations $\mathbf{y} = (y_{i1}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})$.

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

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

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 (2) 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 (1) to see that the maximum value for the likelihood is

$$\begin{aligned} L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2|\mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp -\frac{1}{2\hat{\sigma}^2}\left(\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2\right) \\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp -\frac{n_1+n_2}{2} \end{aligned}$$

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.

$$L(\mu, \sigma^2|\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} \exp -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{1i}-\mu)^2 + \sum_{i=1}^{n_2}(y_{2i}-\mu)^2\right) \tag{3}$$

$$\ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{(n_1+n_2)}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{1i}-\mu)^2 + \sum_{i=1}^{n_2}(y_{2i}-\mu)^2\right) \tag{4}$$

The μ derivative of (4) is

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

Set this to 0 and solve to realize that the maximum likelihood estimator under the null hypothesis is the **grand sample mean** $\overline{\overline{y}}$ obtained by considering all of the data being derived from one large sample

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

Intuitively, if the null hypothesis is true, then all of our observations are independent and have the same distribution and thus, we should use all of the data to estimate the common mean of this distribution.

The value for σ^2 that maximizes (3) on 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 maximum value on Θ_0 for the likelihood is

$$\begin{aligned} L(\hat{\mu}, \hat{\sigma}_0^2|\mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp - \frac{1}{2\hat{\sigma}_0^2}\left(\sum_{i=1}^{n_1} (y_{1i} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2} (y_{2i} - \overline{\overline{y}})^2\right) \\ &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp - \frac{n_1 + n_2}{2} \end{aligned}$$

This give a likelihood ratio of

$$\Lambda(\mathbf{y}) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{-(n_1+n_2)/2} = \left(\frac{\sum_{i=1}^{n_1}(y_{1i} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{2i} - \overline{\overline{y}})^2}{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}\right)^{-(n_1+n_2)/2}. \tag{5}$$

This is the ratio, SS_{total}, of the variation of individuals observations from the grand mean and $SS_{residuals}$. the variation of these observations from the mean of its own groups. Traditionally, this is simplified by looking at the differences of these two types of variation, the numerator in (5)

$$SS_{\text{total}} = \sum_{i=1}^{n_1} (y_{1i} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2} (y_{2i} - \overline{\overline{y}})^2$$

and the denominator in (5)

$$SS_{\text{residuals}} = \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2$$

Exercise 2. *Show that* $SS_{\text{total}} - SS_{\text{residuals}} = n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2$.

In words, SS_{total} 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,

$$SS_{\text{between}} = n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2.$$

The second is the sums of squares the difference of the individual differences and the group mean, $SS_{\text{residuals}}$. Thus, we can write

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

Now, the likelihood ratio (5) reads

$$\Lambda(\mathbf{y}) = \left(1 + \frac{n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}\right)^{-(n_1+n_2)/2} = \left(1 + \frac{SS_{\text{between}}}{SS_{\text{residuals}}}\right)^{-(n_1+n_2)/2}$$

Due to the negative power in the exponent, the critical region $\Lambda(\mathbf{y}) \leq \lambda_0$ is equivalent to

$$\frac{SS_{\text{between}}}{SS_{\text{residuals}}} = \frac{n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2} \geq c \tag{6}$$

for an appropriate value c. We will soon see that the ratio in (6) is, under the null hypothesis, a multiple of an F-distribution.

As promised, we reject if the variation between the groups is large compared to the variation within the groups.

Exercise 3 (pooled two-sample t-test)**.** *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.

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)$$

Thus, we can use the two-sample procedure to compare any two of the three groups. For example, to compared the never logged forest plots to those logged 8 years ago., we find the pooled variance

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) = \frac{1}{19}(11 \cdot 5.065^2 + 8 \cdot 5.761^2) = 28.827$$

and $s_p = 5.37$. Thus, the t-statistic

$$t = \frac{23.750 - 15.778}{5.37\sqrt{\frac{1}{12} + \frac{1}{9}}} = 7.644.$$

```
> 1-pt(7.644,19)
[1] 1.636569e-07
```

Thus, the P-value at 1.64×10^7 is strong evidence against the null hypothesis.

We will entend these ideas so that we can compare more than two groups.

2 One Way Analysis of Variance

For one way analysis of variance, we expand to more than two 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

$$\overline{y}_j = \frac{1}{n_j}\sum_{ji=1}^{n_j} y_{ij}.$$

is the maximum likelihood estimator $\widehat{\mu}_i$ for the hypothesized common mean for the μ_i under the null hpothesis.

The overall mean

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

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

Write the total **sums of squares**

$$SS_{\text{total}} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2 \tag{7}$$

then, under the null hypothesis, the maximum likelihood estimator of the variance is

$$\widehat{\sigma^2}_0 = \frac{1}{n} SS^2_{\text{total}}.$$

repeating the computation above, we see that the interior sum in (7) can be written

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

Here, s_j^2 is the unbiased estimator of the variance based on the observations in the j-th group. This yields as we saw in the two sample case

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

with

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

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_{between}	$s^2_{\text{between}} = SS_{\text{between}}/(q-1)$
residuals	$n-q$	SS_{residual}	$s^2_{\text{residual}} = SS_{\text{residual}}/(n-q)$
total	$n-1$	SS_{total}	

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

The test statistic

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

is, under the null hypothesis, a constant multiple of the ratio of two indepndent χ^2 random variables with parameter $q-1$ for the numerator and $n-q$ for the denominator. This ratio is called an F **random variable** 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 an overall mean

$$\overline{\overline{y}} = \frac{1}{n}\sum_{j=1}^{3} n_j \bar{y}_j = \frac{1}{12+12+9}(12 \cdot 23.750 + 12 \cdot 14.083 + 9 \cdot 15.778) = 18.06055.$$

$$SS_{\text{between}} = \sum_{j=1}^{3} n_j(\overline{y}_j - \overline{\overline{y}})^2 = 12 \cdot (23.750 - \overline{\overline{y}})^2 + 12 \cdot (14.083 - \overline{\overline{y}})^2 + 9 \cdot (15.778 - \overline{\overline{y}})^2) = 625.1793$$

$$SS_{\text{residual}} = \sum_{j=1}^{3} (n_j - 1)s_j^2 = (12-1) \cdot 5.065^2 + (12-1) \cdot 4.981^2 + (9-1) \cdot 5.761^2 = 820.6234$$

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

The value of the F-statistic is $312.6/27.4 = 11.43$ and the P-value (calculated below) is 0.0002. The critical value for a $\alpha = 0.01$ level test is 5.390. So, we do reject the null hypothesis that mean number of trees 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 can use the data from all of the groups as an unbiased estimate for the standard deviation. In particular, the variance $s^2_{\text{residuals}}$ is given by $SS_{\text{residuals}}/(n-q)$, shown in the table in the "mean square" column and the "residuals" row. The standard deviation s_{residual} is the square root of this number. For example, the γ-level confidence interval for μ_j is

$$\overline{y}_j \pm t^*_{(1-\gamma)/2,n-q} \frac{s_{\text{residual}}}{\sqrt{n_j}}.$$

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

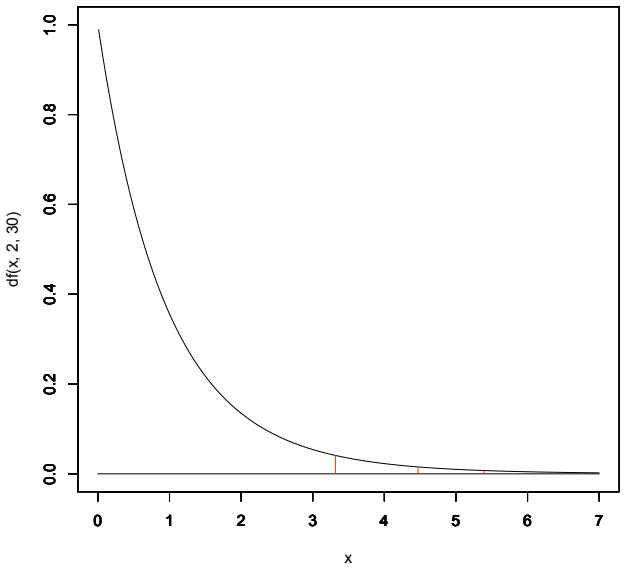

Figure 1: Upper tail critical values. The density for an F random variable with numerator degrees of freedom, 2, and denominator degrees of freedom, 30. The indicated values 3.316, 4.470, and 5.390 are critical values for an significance levels $\alpha = 0.05$, 0.02, and 0.01, respectively.

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

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

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

Example 4. *The development time for a European queen in a honey bee hive is suspected to depend on the temperature of the hive. To examine this, queens are reared in a low temperature hive (*$31.1°$*), a medium temperature hive (*$32.8°$*) and a high temperature hive (*$34.4°$*). The hypothesis is that higher temperatures increase metabolism rate and thus*

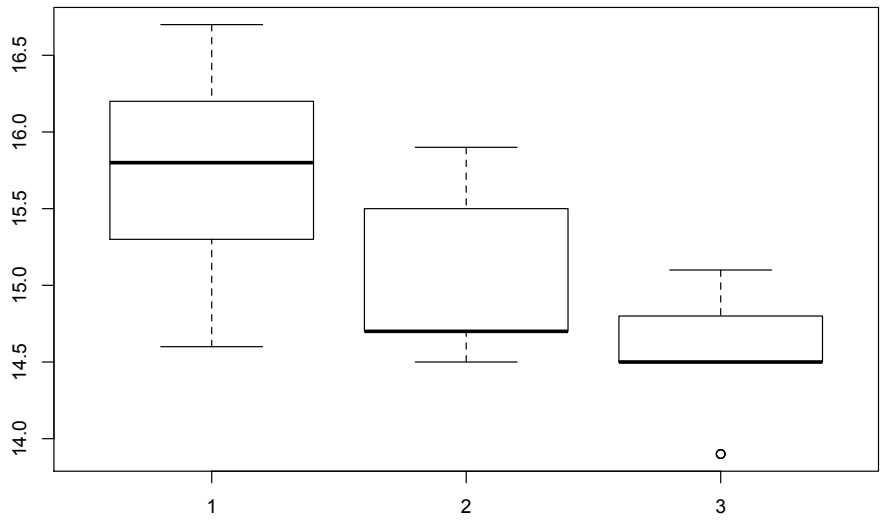

Figure 2: Side-by-side boxplot of queen development times. The time is measured in days. the plots show cool (1) medium (2) and warm (3) hive temperatures.

reduce the time needed from the time the egg is laid until the newly emerged adult queen honey emerges from the cell. The hypothesis is

$$H_0 : \mu_{\text{low}} = \mu_{\text{med}} = \mu_{\text{high}} \quad \text{versus} \quad H_1 : \mu_{\text{low}}, \mu_{\text{med}}, \mu_{\text{high}} \textit{ differ}$$

where $\mu_{\text{low}}, \mu_{\text{med}}$, and μ_{high} are, respectively, the mean queen development time in days for queen eggs reared in a low, a medium, and a high temperature hive.

Here are the data and a boxplot:

```
> ehblow<-c(16.2,14.6,15.8,15.8,15.8,15.8,16.2,16.7,15.8,16.7,15.3,14.6,15.3,15.8)
> ehbmed<-c(14.5,14.7,15.9,15.5,14.7,14.7,14.7,15.5,14.7,15.2,15.2,15.9,14.7,14.7)
> ehbhigh<-c(13.9,15.1,14.8,15.1,14.5,14.5,14.5,14.5,13.9,14.5,14.8,14.8,13.9,14.8,
  14.5,14.5,14.8,14.5,14.8)
> boxplot(ehblow,ehbmed,ehbhigh)
```

The commands in R *to perform analysis and the output are shown below.*

```
> ehb<-c(ehblow,ehbmed,ehbhigh)
> temp<-c(rep(1,length(ehblow)),rep(2,length(ehbmed)),rep(3,length(ehbhigh)))
> ftemp<-factor(temp,c(1:3))
> anova(lm(ehb~ftemp))
Analysis of Variance Table

Response: ehb
          Df Sum Sq Mean Sq F value    Pr(>F)
ftemp      2 11.222  5.6111  23.307 1.252e-07 ***
Residuals 44 10.593  0.2407
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The first line put all of the data in a single vector, `ehb`*. We then put labels for the groups in the variable* `temp`*. This variable is considered as a numerical vector, so we tell* R *that it should be thought of as a factor and list the factors*

in the vector `ftemp`*. Without this, the command* `anova(lm(ehb temp))` *would attempt to do linear regression with* `temp` *as the explanatory variable.*

The anova output shows strong evidence against the null hypothesis. The p-value is 1.252×10^{-7}*. The values in the table can be computed directly*

```
> sum((ehblow-mean(ehblow))^2)+sum((ehbmed-mean(ehbmed))^2)
  + sum((ehbhigh-mean(ehbhigh))^2)
[1] 10.59278
> length(ehblow)*(mean(ehblow)-mean(ehb))^2
  + length(ehbmed)*(mean(ehbmed)-mean(ehb))^2
  + length(ehbhigh)*(mean(ehbhigh)-mean(ehb))^2
[1] 11.22211
```

For confidence intervals we use $s^2_{resid} = 0.2407$*,* $s_{resid} = 0.4906$ *and the* t*-distribution with 44 degrees of freedom.*

For the medium temperature hive, the 95% confidence interval for μ_{med} *can be computed*

```
> mean(ehblow)
[1] 15.74286
> qt(0.975,44)
[1] 2.015368
> length(ehblow)
[1] 14
```

Thus, the intverval is

$$\bar{y}_{med} \pm t_{44,0.025}\frac{s_{resid}}{\sqrt{n_{med}}} = 15.742 \pm 2.0154\frac{0.4906}{\sqrt{14}} = (15.478, 16.006)$$

3 Contrasts

After completing a one way analysis of variance, resulting in rejecting the null hypotheses, a typical follow-up procedure is the use of **contrasts**. Contrasts use as a null hypothesis that some linear combination of the means equals to zero.

Example 5. *If we want to see if the rain forest has seen recovery in logged areas over the past 8 years. This can be written as*

$$H_0 : \mu_2 - \mu_3 = 0 \quad \text{versus} \quad H_1 : \mu_2 - \mu_3 \neq 0$$

or

$$H_0 : \mu_2 = \mu_3 \quad \text{versus} \quad H_1 : \mu_2 \neq \mu_3.$$

Under the null hypothesis, the test statistic

$$t = \frac{\bar{y}_2 - \bar{y}_3}{s_{residual}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}},$$

has a t*-distribution with* $n - q$ *degrees of freedom. Here*

$$t = \frac{14.083 - 15.778}{5.234\sqrt{\frac{1}{12} + \frac{1}{9}}} = -0.7344,$$

with $n - q = 33 - 3$ *degrees of freedom, the p-value*

```
> 2*pt(-0.7344094,30)
[1] 0.4684011
```

is considerably too high to reject the null hypothesis.

Example 6. *To see if the mean queen development medium hive temperature is midway between the time for the high and low temperature hives, we have the contrast,*

$$H_0 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0 \quad \text{versus} \quad H_1 : \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} \neq 0$$

or

$$H_0 : \frac{1}{2}(\mu_{low} + \mu_{high}) = \mu_{med} \quad \text{versus} \quad H_1 : \frac{1}{2}(\mu_{low} + \mu_{high}) \neq \mu_{med}$$

Notice that,

$$E\left[\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right] = \frac{1}{2}\mu_{low} - \mu_{med} + \frac{1}{2}\mu_{high} = 0$$

and

$$\text{Var}\left(\frac{1}{2}\bar{Y}_{low} - \bar{Y}_{med} + \frac{1}{2}\bar{Y}_{high}\right) = \frac{1}{4}\frac{\sigma^2}{n_{low}} + \frac{\sigma^2}{n_{med}} + \frac{1}{4}\frac{\sigma^2}{n_{high}}.$$

This leads to the test statistic

$$t = \frac{\frac{1}{2}\bar{y}_{low} - \bar{y}_{med} + \frac{1}{2}\bar{y}_{high}}{s_{residual}\sqrt{\frac{1}{4n_{low}} + \frac{1}{n_{med}} + \frac{1}{4n_{high}}}} = \frac{\frac{1}{2}15.743 - 15.043 + \frac{1}{2}14.563}{0.4906\sqrt{\frac{1}{4\cdot 14} + \frac{1}{14} + \frac{1}{4\cdot 19}}} = 0.7005.$$

The p-value,

```
> 2*(1-pt(0.7005,44))
[1] 0.487303
```

again, is considerably too high to reject the null hypothesis.

In general, a contrast is a linear combination of the means

$$\psi = c_1\mu_1 + \cdots + c_q\mu_q.$$

The hypothesis is

$$H_0 : \psi = 0 \quad \text{versus} \quad H_1 : \psi \neq 0$$

For sample means, $\bar{y}_1, \ldots, \bar{y}_q$, the test statistic is

$$t = \frac{c_1\bar{y}_1 + \cdots + c_q\bar{y}_q}{s_{residual}\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_q^2}{n_q}}}.$$

Under the null hypothesis the t statistic has a t distribution with $n - q$ degrees of freedom.

4 Answer to Selected Exercises

2. Let's look at this difference for each of the groups.

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

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

Now the numerator in (5) can be written to show the decomposition of the variation into two sources - the within group variation and the between group variation.

$$\begin{aligned}\sum_{i=1}^{n_1}(y_{1i}-\overline{\overline{y}})^2+\sum_{i=1}^{n_2}(y_{2i}-\overline{\overline{y}})^2 &= \sum_{i=1}^{n_1}(y_{1j}-\bar{y}_1)^2+\sum_{i=1}^{n_2}(y_{2j}-\bar{y}_2)^2+n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2.\\ &= (n_1-1)s_1^2+(n_2-1)s_2^2+n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2.\end{aligned}$$

3. We will multiply the numerator in (6) by $(n_1+n_2)^2$ and note that $(n_1+n_2)\overline{\overline{y}}=n_1\bar{y}_1+n_2\bar{y}_2$. Then,

$$\begin{aligned}(n_1+n_2)^2(n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2 &= n_1((n_1+n_2)\overline{\overline{y}}-(n_1+n_2)\bar{y}_1)^2+n_2((n_1+n_2)\overline{\overline{y}}-(n_1+n_2)\bar{y}_2)^2\\ &= n_1(n_1\bar{y}_1+n_2\bar{y}_2-(n_1+n_2)\bar{y}_1)^2+n_2(n_1\bar{y}_1+n_2\bar{y}_2-(n_1+n_2)\bar{y}_2)^2\\ &= n_1(n_2(\bar{y}_2-\bar{y}_1))^2+n_2(n_1(\bar{y}_1-\bar{y}_2))^2\\ &= (n_1n_2^2+n_2n_1^2)(\bar{y}_1-\bar{y}_2)^2=n_1n_2(n_1+n_2)(\bar{y}_1-\bar{y}_2)^2\end{aligned}$$

Consequently

$$(n_1(\overline{\overline{y}}-\bar{y}_1)^2+n_2(\overline{\overline{y}}-\bar{y}_2)^2=\frac{n_1n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)^2=(\bar{y}_1-\bar{y}_2)^2/\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

The denominator

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

The ratio

$$\frac{SS_{\text{between}}}{SS_{\text{residuals}}}=\frac{(\bar{y}_1-\bar{y}_2)^2}{(n_1+n_2-2)s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}=\frac{T(\mathbf{y})^2}{n_1+n_2-2}.$$

Thus, the test is a constant multiple of the square of the t-statistic. Take the square root of both sides to create a test using a threshold value for $|T(\mathbf{y})|$ for the critical region.