Topic 22

Analysis of Variance

22.1 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. We shall explain the methodology through an example. Consider the data set gathered from the forests in Borneo.

Example 22.1 (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 $y_{11}, \dots y_{1n_1}, \quad y_{21}, \dots y_{2n_2}$ 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 (ANOVA) is a statistical procedure that allows us to test for the differences in means for 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 means the same (the null hypothesis) or not (the alternative hypothesis)?

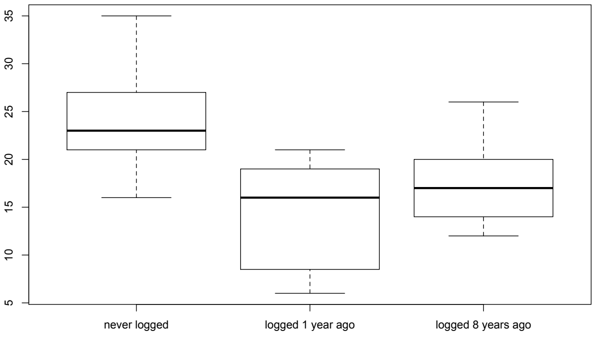

Figure 22.1: Side-by-side boxplots of the number of trees per plot. The groups will be considered different if the differences between the groups (indicated by the variation in the center lines of the boxes) is large compared to the width of the boxes in the boxplot.

As the case with the t procedures, the appropriateness of one way analysis of variance is based on the applicability of the central limit theorem. As with t procedures, ANOVA has an alternative, the Kruskal-Wallis test, based on the ranks of the data for circumstances in which the central limit theorem does not apply.

The basic idea of the test is to examine the ratio of s^2_{between}, the variance between the groups 1, 2, and 3. and s^2_{residual}, a statistic that measures the variances within the 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 are able to reject the null hypothesis. Even though the boxplots use different measures of center (median vs. mean) and spread (quartiles vs. standard deviation), this idea can be expressed by examining the fluctuation in the centers of boxes in Figure 22.1 compared to the width of the boxes.

As we have seen before, this decision to reject H_0 will be the consequence a sufficiently high value of a test statistic - in this case the F statistic. 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). Consequently, variances between groups that are not that are not statistically significant for small sample sizes can become significant as the sample sizes and, with it, the power increase.

22.2 One Way Analysis of Variance

For one way analysis of variance, we expand to more than the two groups seen for t procedures and ask whether or not the means of 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 $\{y_{ij}, 1 \le i \le n_j, 1 \le j \le q\}$ represents that we have n_i observation for the i-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

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

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. Notice that the model for analysis requires a common value σ for all of the observations.

In order to develop the F statistic at the test statistic, we will need to introduce two types of sample means:

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

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

  These are given in the table in Example 22.1 for the Borneo rains forest.

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

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

  This is the weighted average of the $\bar{y}_i$ with weights n_i, the sample size in each group. The Borneo rain forest example has an overall mean

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

Analysis of variance uses the **total sums of squares**

$$SS_{\text{total}} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2, \tag{22.1}$$

the total square variation of individual observations from their grand mean. However, the test statistic is determined by decomposing SS_{total}. We start with a bit of algebra to rewrite the interior sum in (22.1) as

$$\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2 = (n_j - 1)s_j^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2. \tag{22.2}$$

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

source of variation	degrees of freedom	sums of squares	mean square
between groups	$q-1$	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}	

Table I: Table for one way analysis of variance

Exercise 22.2. *Show the first equality in (22.2). (Hint: Begin with the difference in the two sums.)*

Together (22.1) and (22.2) yields the decomposition of the variation

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

with

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

For the rain forest example, we find that

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

and

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

From this, we obtain the general form for one-way analysis of variance as shown in Table I.

- The $q-1$ degrees of freedom between groups is derived from the q groups minus one degree of freedom used to compute $\bar{\bar{y}}$.
- 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. Add these q values for the degrees of freedom to obtain $n-q$.

The test statistic

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

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	

Table II: Analysis of variance information for the Borneo rain forest data

is, under the null hypothesis, a constant multiple of the ratio of two independent 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.

Using Table II, we find the value of the test statistic for the rain forest data is

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

and the p-value (calculated below) is 0.0002. The critical value for an $\alpha = 0.01$ 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
```

Confidence intervals are determined using the data from all of the groups as an unbiased estimate for the variance, σ^2. Using all of the data allows us to increase the number of degrees of freedom in the t distribution and thus reduce the upper critical value for the t statistics and with it the margin of error.

The variance $s^2_{\rm residuals}$ is given by the expression $SS_{\rm residuals}/(n-q)$, shown in the table in the "mean square" column and the "residuals" row. The standard deviation $s_{\rm residual}$ is the square root of this number. For example, the γ-level confidence interval for μ_j is

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

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

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

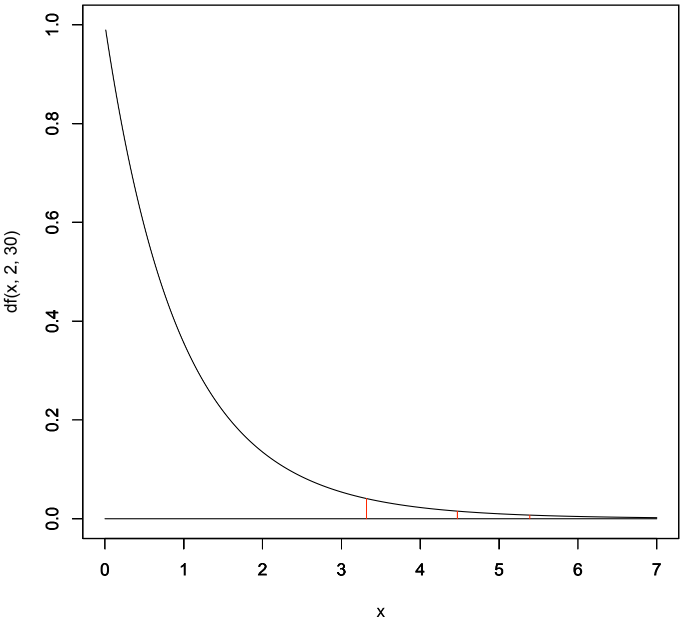

Figure 22.2: Upper tail critical values. The density for an F random variable with numerator degrees of freedom, 2, and denominator degrees of freedom, 30. The indicated values 3.316, 4.470, and 5.390 are critical values for significance levels $\alpha = 0.05$, 0.02, and 0.01, respectively.

In this case, the 95% confidence interval for the mean number of trees on a lot"logged 1 year ago" has $n - q = 33 - 3$, $t_{0.025,30} = 2.042$, $s_{\rm residual} = \sqrt{27.4} = 5.234$ and the confidence interval is

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

Exercise 22.3. *Give the 95% confidence intervals for the difference in trees between plots never logged and plots logged 8 years ago.*

Example 22.4. *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° *C), a medium temperature hive (*32.8° *C) and a high temperature hive (*34.4° *C). The hypothesis is that higher temperatures increase metabolism rate and thus reduce the time needed from the time the egg is laid until an adult queen honey bee emerges from the cell. The hypothesis is*

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

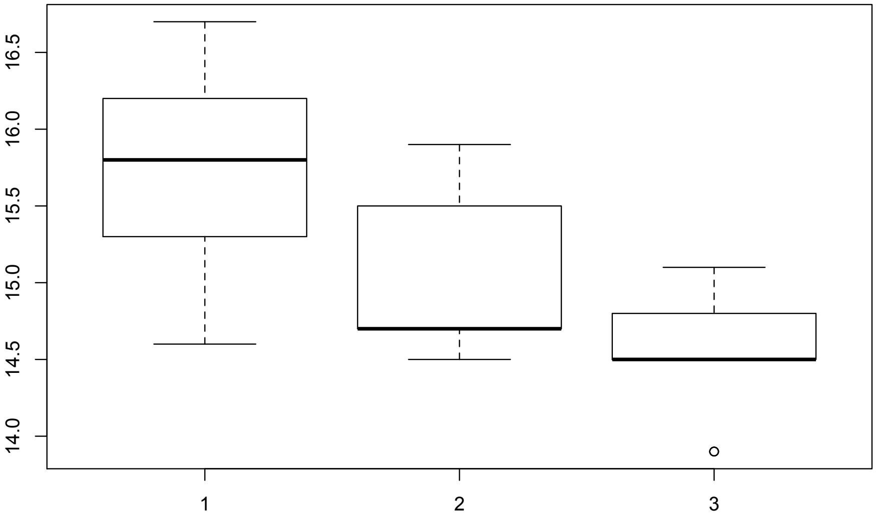

Figure 22.3: Side-by-side boxplot of queen development times. The time is measured in days. the plots show cool (1) medium (2) and warm (3) hive temperatures.

where $\mu_{\rm low}, \mu_{\rm med}$, and $\mu_{\rm high}$ are, respectively, the mean 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. The first line put all of the data in a single vector,* `ehb`. *We then put labels for the groups in the variable or factor* `temp`. *Expressed in this way, this variable is considered by* R *as a numerical vector. To 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.*

```
> 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 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 from the formulas above.*

For the sums of square between groups, SS_{between}*,*

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

and within groups, SS_{residual}*,*

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

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

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

22.3 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 22.5. *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*

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

or

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

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 for this 2-sided test is*

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

is considerably too high to reject the null hypothesis.

Example 22.6. *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,*

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

or

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

Notice that, under the null hypothesis

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

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.

Exercise 22.7. *Under the null hypothesis appropriate for one way analysis of variance, with n_i observations in group $i = 1, \ldots, q$ and $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$,*

$$E[c_1\bar{Y}_1 + \cdots + Y_q\mu_q] = c_1\mu_1 + \cdots + c_q\mu_q, \quad \text{Var}(c_1\bar{Y}_1 + \cdots + c_qY_q) = \frac{c_1^2\sigma^2}{n_1} + \cdots + \frac{c_q^2\sigma^2}{n_q}.$$

In general, a contrast begins with 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.

22.4 Two Sample Procedures

We now show that the t-sample procedure results from a likelihood ratio test. We keep to two groups in the development of the F test. The essential features can be found in this example without the extra notation necessary for an arbitrary number of groups.

Our hypothesis test is based on two independent samples of normal random variables. The data are

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

where ϵ_{ij} are independent $N(0, \sigma)$ random variables with σ unknown. Thus, we have n_j independent $N(\mu_j, \sigma)$ random variables $Y_{1j} \ldots, Y_{n_j j}$ with unknown *common* variance σ^2, $j = 1$ and 2. The assumption of a common variance 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

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

Step 1. Determine the log-likelihood. To find the test statistic derived from a likelihood ratio test, we first write the likelihood and its logarithm based on observations $\mathbf{y} = (y_{11}, \ldots, y_{n_1 1}, y_{12}, \ldots, y_{n_2 2})$.

$$L(\mu_1.\mu_2, \sigma^2|\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} \exp -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right) \tag{22.3}$$

$$\ln L(\mu_1.\mu_2, \sigma^2|\mathbf{y}) = -\frac{(n_1 + n_2)}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right) \tag{22.4}$$

Step 2. Find the maximum likelihood estimates and the maximum value of the likelihood. By taking partial derivatives with respect to μ_1 and μ_2 we see that with two independent samples, the maximum likelihood estimate for the mean μ_j for each of the samples is the sample mean $\bar{y}_j$.

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i1}, \quad \hat{\mu}_2 = \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{i2}.$$

Now differentiate (22.4) with respect to σ^2

$$\frac{\partial}{\partial\sigma^2} \ln L(\mu_1, \mu_2, \sigma^2|\mathbf{x}) = -\frac{n_1 + n_2}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n_1}(y_{i1} - \mu_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \mu_2)^2\right).$$

Thus, the maximum likelihood estimate 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.

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2\right).$$

Now, substitute these values into the likelihood (22.3) to see that the maximum value for the likelihood is

$$\begin{aligned} L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2|\mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp -\frac{1}{2\hat{\sigma}^2}\left(\sum_{i=1}^{n_1}(y_{i1}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2}-\bar{y}_2)^2\right) \\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{(n_1+n_2)/2}} \exp -\frac{n_1+n_2}{2} \end{aligned}$$

Step 3. Find the parameters that maximize the likelihood under the null hypothesis and then find the maximum value of the likelihood on Θ_0. 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.

$$L(\mu, \sigma^2|\mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} \exp -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1}-\mu)^2 + \sum_{i=1}^{n_2}(y_{i2}-\mu)^2\right) \tag{22.5}$$

$$\ln L(\mu, \sigma^2|\mathbf{x}) = -\frac{(n_1+n_2)}{2}(\ln 2\pi + \ln\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1}-\mu)^2 + \sum_{i=1}^{n_2}(y_{i2}-\mu)^2\right) \tag{22.6}$$

The μ derivative of (22.6) is

$$\frac{\partial}{\partial\mu}\ln L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n_1}(y_{i1}-\mu) + \sum_{i=1}^{n_2}(y_{i2}-\mu)\right).$$

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

$$\hat{\mu}_0 = \overline{\overline{y}} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} y_{i1} + \sum_{i=1}^{n_2} y_{i2}\right) = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2}.$$

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 (22.5) 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:

$$\hat{\sigma}_0^2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(y_{i1}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\overline{\overline{y}})^2\right).$$

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

$$\begin{aligned} L(\hat{\mu}_0, \hat{\sigma}_0^2|\mathbf{x}) &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp -\frac{1}{2\hat{\sigma}_0^2}\left(\sum_{i=1}^{n_1}(y_{i1}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\overline{\overline{y}})^2\right) \\ &= \frac{1}{(2\pi\hat{\sigma}_0^2)^{(n_1+n_2)/2}} \exp -\frac{n_1+n_2}{2} \end{aligned}$$

Step 4. Find the likelihood statistic $\Lambda(\mathbf{y})$. From steps 2 and 3, we find a likelihood ratio of

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2|\mathbf{x})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{x})} = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-(n_1+n_2)/2} = \left(\frac{\sum_{i=1}^{n_1}(y_{i1}-\overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\overline{\overline{y}})^2}{\sum_{i=1}^{n_1}(y_{i1}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2}-\bar{y}_2)^2}\right)^{-(n_1+n_2)/2}. \tag{22.7}$$

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.

Step 5. Simplify the likelihood statistic to determine the test statistic F. Traditionally, the likelihood ratio is simplified by looking at the differences of these two types of variation, the numerator in (22.7)

$$SS_{\text{total}} = \sum_{i=1}^{n_1}(y_{i1} - \overline{\overline{y}})^2 + \sum_{i=1}^{n_2}(y_{i2} - \overline{\overline{y}})^2$$

and the denominator in (22.7)

$$SS_{\text{residuals}} = \sum_{i=1}^{n_1}(y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{i2} - \bar{y}_2)^2$$

Exercise 22.8. *Show that* $SS_{\text{total}} - SS_{\text{residuals}} = n_1(\bar{y}_1 - \overline{\overline{y}})^2 + n_2(\bar{y}_2 - \overline{\overline{y}})^2$.

In words, SS_{total} the sums of squares of the differences of an individual observation from the overall mean $\overline{\overline{y}}$, is the sum of two sources. The first is the sums of squares of the difference of the average of each 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 of the difference of the individual observations with its own group mean, $SS_{\text{residuals}}$. Thus, we can write

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

Now, the likelihood ratio (22.7) reads

$$\Lambda(\mathbf{y}) = \left(\frac{SS_{\text{residual}} + SS_{\text{between}}}{SS_{\text{residuals}}}\right) = \left(1 + \frac{SS_{\text{between}}}{SS_{\text{residuals}}}\right)^{-(n_1+n_2)/2}$$

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

$$\frac{SS_{\text{between}}}{SS_{\text{residuals}}} = \frac{n_1(\overline{\overline{y}} - \bar{y}_1)^2 + n_2(\overline{\overline{y}} - \bar{y}_2)^2}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2} \geq c \tag{22.8}$$

for an appropriate value c. The ratio in (22.8) is, under the null hypothesis, a multiple of an F-distribution. The last step to divide both the numerator and denominator by the degrees of freedom. Thus, we see, as promised, we reject if the F-statistics is too large, i.e., the variation between the groups is sufficiently large compared to the variation within the groups.

Exercise 22.9 (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.

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

Exercise 22.10. *Generalize the formulae for $\overline{\overline{y}}$, SS_{between} and $SS_{\text{residuals}}$ from the case $q = 2$ to an arbitrary number of groups.*

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.

22.5 Kruskal-Wallis Rank-Sum Test

The Kruskal-Wallis test is an alternative to one-way analysis of variance in much the same way that the Wilcoxon rank-sum test is a alternative to two-sample t procedures. Like the Wilcoxon test, we replace the actual data with their ranks. This non-parametric alternative obviates the need to use the normal distribution arising from an application of the central limit theorem. The H test statistic has several analogies with the F statistic. To compute this statistic:

- Replace the data $\{y_{ij}, 1 \le i \le n_j, 1 \le j \le q\}$ for n_i observations for the i-th group from each of the q groups with $\{r_{ij}, 1 \le i \le n_j, 1 \le j \le q\}$, the ranks of the data taking all of the groups together. For ties, average the ranks.
- The total number of observations $n = n_1 + \cdots + n_q$.
- The average rank within the groups

$$\bar{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}, \quad i = 1, \ldots, q.$$

- The grand average of the ranks

$$\bar{\bar{r}} = \frac{1}{n}(1 + \cdots + n) = \frac{1}{n}n(n+1) = \frac{n+1}{2}.$$

  (See Exercise 20.6.)

- The Kruskal-Wallis test statistic looks at the sums of squares of ranks between groups and the total sum of squares of ranks

$$H = \frac{SSR_{\text{between}}}{SSR_{\text{total}}/(n-1)} = \frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{\bar{r}})^2}{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(r_{ij} - \bar{\bar{r}})^2/(n-1)},$$

- For larger data sets (each $n_i \ge 5$), the p-value is approximately the probability that a χ^2_{q-1} random variable exceeds the value of the H statistic.
- For smaller data sets, more sophisticated procedures are necessary.
- The test can be followed by using a procedure analogous to contrasts based on the Wilcoxon rank-sum test.

Exercise 22.11. *For the case of no ties, show that*

$$SSR_{\text{total}} = \frac{(n-1)n(n+1)}{12}$$

In this case,

$$H = \frac{12}{n(n+1)}\sum_{i=1}^q n_i\left(\bar{r}_i - \frac{n+1}{2}\right)^2 = \frac{12}{n(n+1)}\sum_{i=1}^q n_i\bar{r}_i^2 - 3(n+1).$$

The Kruskal-Wallis test also gives a very small p-value to the queen development times for Africanized honey bees. Begin with the R commands in Example 22.4 to enter the data and create the temperature factors `ftemp`.

```
> kruskal.test(ehb~ftemp)

Kruskal-Wallis rank sum test

data:  ehb by ftemp
Kruskal-Wallis chi-squared = 20.4946, df = 2, p-value = 3.545e-05
```

22.6 Answer to Selected Exercises

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

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

Now the numerator in (22.7) can be written to show the decomposition of the variation into two sources - the within group variation and the between group variation.

$$\sum_{i=1}^{n_1}(y_{i1}-\bar{\bar{y}})^2 + \sum_{i=1}^{n_2}(y_{i2}-\bar{\bar{y}})^2 = \sum_{i=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2j}-\bar{y}_2)^2 + n_1(\bar{\bar{y}}-\bar{y}_1)^2 + n_2(\bar{\bar{y}}-\bar{y}_2)^2.$$
$$= (n_1-1)s_1^2 + (n_2-1)s_2^2 + n_1(\bar{\bar{y}}-\bar{y}_1)^2 + n_2(\bar{\bar{y}}-\bar{y}_2)^2.$$

22.3. Here, we are looking for a confidence interval for $\mu_1 - \mu_3$. From the summaries, we need

$$n_1 = 12, \quad \bar{y}_1 = 23.750, \quad n_3 = 9, \quad \bar{y}_3 = 17.778.$$

From the computation for the test, we have $s_{\text{residual}} = \sqrt{27.4} = 5.234$ and using the `qt(0.975,30)` command we find $t_{0.025,30} = 2.042$. Thus,

$$\begin{aligned} &(\bar{y}_1-\bar{y}_3) &&\pm t_{(0.975,30)}s_{\text{residual}}\sqrt{\frac{1}{n_1}+\frac{1}{n_3}} \\ =\ &(23.750-17.778) &&\pm 2.042\cdot 5.234\sqrt{\frac{1}{12}+\frac{1}{9}} \\ =\ &\quad 5.972 &&\pm 2.079 = (3.893, 8.051) \end{aligned}$$

22.7. This follows from the fact that expectation is a linear functional and the generalized Pythagorean identity for the variance of a linear combination of independent random variables.

22.8. Look at the solution to Exercise 22.2.

22.9. We will multiply the numerator in (22.8) 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

$$\sum_{j=1}^{n_1}(y_{i1}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{i2}-\bar{y}_2)^2 = (n_1+n_2-2)s_p^2.$$

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.

22.10. For observations, $y_{i1},\ldots y_{in_i}$ in group $i=1,\ldots,q$, let $n = n_1+\cdots+n_q$ be the total number of observations, then the grand mean

$$\overline{\overline{y}} = \frac{1}{n}\left(n_1\bar{y}_1 + \cdots + n_q\bar{y}_q\right)$$

where $\bar{y}_i$ is the sample mean of the observations in group i. The sums of squares are

$$SS_{\text{between}} = \sum_{i=1}^{q} n_i(\bar{y}_i - \overline{\overline{y}})^2 \quad \text{and} \quad SS_{\text{residuals}} = \sum_{i=1}^{q}(n_i-1)s_i^2$$

where s_i^2 is the sample variance of the observations in group i.

22.11. In anticipation of its need, let's begin by showing that

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}.$$

Notice that the formula holds for the case $n=1$ with

$$1^2 = \frac{1(1+1)(2\cdot 1+1)}{6} = \frac{6}{6} = 1.$$

Now assume that the identity holds for $n=k$. We then check that it also holds for $n=k+1$

$$\begin{aligned}
1^2+2^2+\cdots+k^2+(k+1)^2 &= \frac{k(k+1)(2k+1)}{6} + (k+1)^2 \\
&= \frac{k+1}{6}(k(2k+1)+6(k+1)) = \frac{k+1}{6}(2k^2+7k+6) \\
&= \frac{(k+1)(k+2)(2k+3)}{6}
\end{aligned}$$

This is the formula for $n = k + 1$ and so by the mathematical induction, we have the identity for all non-negative integers.

With no ties, each rank appears once and

$$\begin{aligned} SSR_{\text{total}} &= \sum_{j=1}^{n}\left(j - \frac{n+1}{2}\right)^2 = \sum_{j=1}^{n} j^2 - 2\sum_{j=1}^{n} j\frac{n+1}{2} + \sum_{j=1}^{n}\left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)(2n+1)}{6} - 2\frac{n(n+1)}{2}\frac{n+1}{2} + n\left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} \\ &= \frac{n(n+1)}{12}(2(2n+1) - 3(n+1)) = \frac{(n-1)n(n+1)}{12}. \end{aligned}$$