One-Factor Analysis of Variance, chapter 4.8

Grethe Hystad

November 29, 2012

- In this section we will compare the treatment of several populations. One-factor analysis of variance is a statistical method to test for the difference in two or more treatments or groups.
- We will investigate the ratio of the variance between treatments and a statistics that measures the variances within the treatments. If this ratio is sufficiently large, we will reject the null hypothesis and conclude that there is not a significant difference between the treatments. The test statistics is the F-statistics.
- For One-factor analysis of variance, we can compare more than two treatments. For example, we can test three difference teaching methods with mean test result μ_1, μ_2, μ_3 by taking samples of sizes n_1, n_2, n_3 respectively.

Test for difference in mean

- The random variables $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ of size n_i represents a random sample from $N(\mu_i, \sigma^2)$ for the i^{th} treatment for $i = 1, ..., m$ (m treatments).
- We want to test

$$H_0 : \mu_i = \mu_j \quad \text{for all} \quad i, j \in \{1, ..., m\}$$

against

$$H_1 : \mu_i \neq \mu_j \quad \text{for some} \quad i, j.$$

Test for difference in mean

We have the following model:

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, 2, ..., m \quad j = 1, 2, ..., n_i,$$

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

let $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ represent a random sample of size n_i from $N(\mu_i, \sigma^2)$ for $i = 1, ..., m$.

Table: One-Factor Random Samples

					Means
	Y_{11}	Y_{12}	...	Y_{1n_1}	$\overline{Y}_{1.}$
	Y_{21}	Y_{22}	...	Y_{2n_2}	$\overline{Y}_{2.}$
	..	..	..	..	..
	Y_{m1}	Y_{m2}	...	Y_{mn_m}	$\overline{Y}_{m.}$
Grand Mean					$\overline{Y}_{..}$

The within group mean is

$$\overline{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad i = 1, 2, ..., m.$$

The grand mean is

$$\overline{Y}_{..} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}$$

for $n = n_1 + n_2 + \cdots + n_m$. The dot in the notation for the means indicates the index over which the average is taken. $\overline{Y}_{..}$ indicates that the average is taken over both indices while $\overline{Y}_{i.}$ is taken over the index j.

Sum of squares

We can write

$$SS_{Total} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2+\sum_{i=1}^{m} n_i(\overline{Y}_{i.}-\overline{Y}_{..})^2$$

Proof is given in class.

$$\textbf{Total sum of squares}: SS_{Total} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2$$

$$\textbf{The error sum of squares}: SS_{Res} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$$

which is the sum of squares within treatments or groups.
The between treatments sum of squares:

$$SS_{between} = \sum_{i=1}^{m} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$$

which is the sum of squares among the different treatments/groups. So we have

$$SS_{Total} = SS_{Res} + SS_{between}$$

Unbiased estimator

When H_0 is true, we consider Y_{ij} for $i = 1, ..., m$ and $j = 1, ..., n_i$ as a random sample of size $n = n_1 + n_2 + \cdots n_m$ from $N(\mu, \sigma^2)$.

Unbiased estimator of σ^2:

Since $\frac{SS_{Tot}}{\sigma}$ is $\chi^2(n-1)$, we have that

$$\frac{SS_{Tot}}{n-1}$$

is an unbiased estimator of σ^2 when H_0 is true.

Proof is given in class.

Unbiased estimator

Unbiased estimator of σ^2:

An unbiased estimator of σ^2 based only on the sample from the i^{th} distribution is

$$W_i = \frac{\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)}{n_i - 1}$$

for $i = 1, 2, ..., m$, where $\frac{(n_i-1)W_i}{\sigma^2}$ is $\chi^2(n_i - 1)$.

Proof is given in class.

Unbiased estimator of σ^2:

Since

$$\frac{SS_{Res}}{\sigma^2} = \sum_{i=1}^{m} \frac{(n_i - 1)W_i}{\sigma^2}$$

is $\chi^2(n-m)$, we have that

$$\frac{SS_{Res}}{n-m}$$

is an unbiased estimator of σ^2.

Proof is given in class.

We have that

$$\frac{SS_{Tot}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} + \frac{SSbetween}{\sigma^2},$$

where $\frac{SS_{Tot}}{\sigma^2}$ is $\chi^2(n-1)$ and $\frac{SS_{Res}}{\sigma^2}$ is $\chi^2(n-m)$, so it can be proved that SS_{Res} and $SS_{between}$ are independent and $\frac{SSbetween}{\sigma^2}$ is $\chi^2(m-1)$

Unbiased estimator

$$\frac{SS_{Res}}{n-m}$$

is an unbiased estimator of σ^2 regardless of whether H_0 is true or false.

If H_0 is true so that $\mu_1 = \mu_2 = ... = \mu_m$ then

$$\frac{\frac{SS_{between}}{m-1}}{\frac{SS_{Res}}{n-m}}$$

is close to one.
If H_0 is false, the ratio in the expression above becomes larger.
Under H_0, we have that

$$\frac{\frac{SS_{between}}{m-1}}{\frac{SS_{Res}}{n-m}} = \frac{[\frac{SS_{between}/\sigma^2}{m-1}]}{\frac{[SS_{Res}/\sigma^2]}{n-m}} = F$$

has a F-distribution with $(m-1)$ and $(n-m)$ degrees of freedoms because $SS_{between}/\sigma^2$ and SS_{Res}/σ^2 are independent chi- square random variables.

We have that

$$\frac{SS_{Tot}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} + \frac{SSbetween}{\sigma^2},$$

where $\frac{SS_{Tot}}{\sigma^2}$ is $\chi^2(n-1)$ and $\frac{SS_{Res}}{\sigma^2}$ is $\chi^2(n-m)$, so it can be proved that SS_{Res} and $SS_{between}$ are independent and $\frac{SSbetween}{\sigma^2}$ is $\chi^2(m-1)$

Test statistics

In testing

$$H_0 : \mu_i = \mu_j \quad \text{for} \quad i, j \in \{1, ..., m\}$$

$$\text{against} \quad H_1 : \mu_i \neq \mu_j \quad \text{for at least one} \quad i, j \in \{1, ..., m\},$$

if the test statistics

$$F = \frac{\frac{SS_{between}}{m-1}}{\frac{SS_{Res}}{n-m}} \geq F_\alpha(m-1, n-m),$$

we reject H_0 at the $(1-\alpha) \times 100$ % significance level.

Table: Analysis of variance table, ANOVA

Source	(SS)	Df	(MS)	F-ratio
Treatment	SS_{betw}	$m-1$	$MS_{betw} = \frac{SS_{betw}}{m-1}$	MS_{betw}/MS_{Res}
Error	SS_{Res}	$n-m$	$MS_{Res} = \frac{SS_{Res}}{n-m}$	
Total	SS_{Tot}	$n-1$		

SS=Sum of squares
MS=Mean Sum of Squares

Exercise 4.8-7 from textbook

Example

The driver of a diesel-powered automobile decided to test the quality of three types of diesel fuel sold in the area based on mpg. Test the null hypothesis that the three means are equal using the following data. Make the usual assumptions and take $\alpha = 0.05$.

Brand A:	38.7	39.2	40.1	38.9	
Brand B:	41.9	42.3	41.3		
Brand C:	40.8	41.2	39.5	38.9	40.3

Exercise 4.8-7 from textbook Solution

Solution

$H_0 : \mu_1 = \mu_2 = \mu_3$ *against* H_1*: at least one of the means differ.*

						$\overline{Y}_{i.}$
Brand A:	38.7	39.2	40.1	38.9		39.225
Brand B:	41.9	42.3	41.3			41.83
Brand C:	40.8	41.2	39.5	38.9	40.3	40.14
$\overline{Y}_{..}$						40.40

$SS_{Res} = (38.7 - 39.225)^2 + (39.2 - 39.225)^2 + ... + (41.9 - 41.83)^2 + ... + (40.3 - 40.14)^2 = 5.19$

$SS_{betw} = 4(39.225 - 40.40)^2 + 3(41.83 - 40.40)^2 + 5(40.14 - 40.40)^2 = 11.78$

Solution continue

Solution

$m - 1 = 3 - 1 = 2$ *and* $n - m = 12 - 3 = 9$ *and* $n = 4 + 3 + 5 = 12$. *We have*

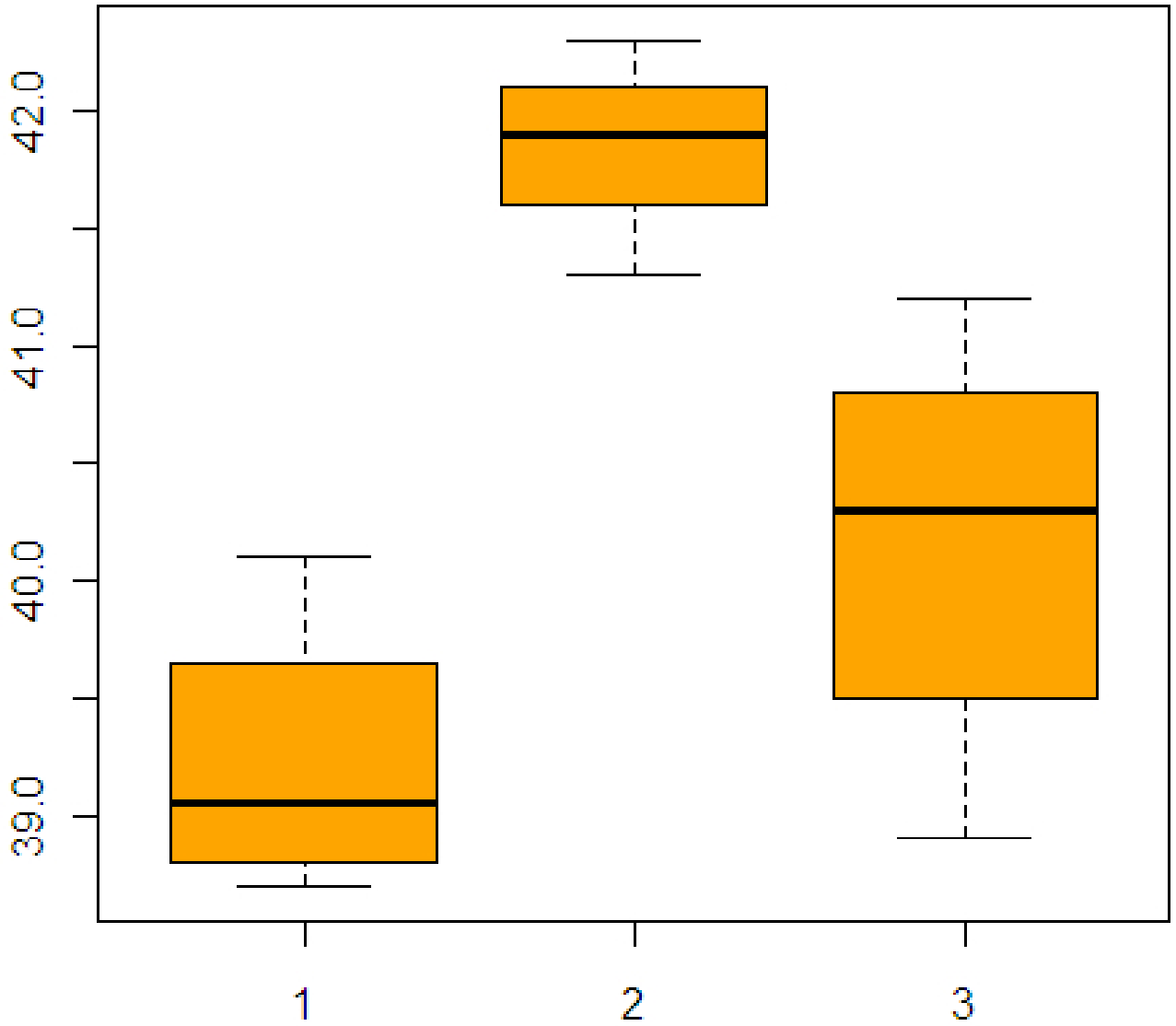
$$\frac{SS_{betw/m-1}}{SS_{Res}/n - m} = \frac{11.78/2}{5.19/9} = 10.224 > F_{0.05}(2, 9) = 4.26.$$

We reject H_0 *so there is a significant difference in the means at the* 5% *level.*

R-code, ANOVA

```
> x=c(38.7,39.2,40.1,38.9)
> y=c(41.9,42.3,41.3)
> z=c(40.8,41.2,39.5,38.9,40.3)
> mpg=c(x,y,z)
> fuel=c(rep(1,length(x)),rep(2,length(y)),rep(3,length(z)))
> factorfuel=factor(fuel,c(1:3))
> anova(lm(mpg~factorfuel))
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value   Pr(>F)
factorfuel  2 11.7830  5.8915  10.224 0.004823 **
Residuals   9  5.1862  0.5762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


42.0
41.0
40.0
39.0
1
2
3