One-Way ANOVA
Comparison of Several Means

Yibi Huang

Textbook 3.1-3.8

Case Study: Grass/Weed Competition

Textbook, Problem 6.1, p.147

To study the competition of big bluestem (from the tall grass prairie) versus quack grass (a weed), we set up an experimental garden with 24 plots. These plots were randomly allocated to the 6 treatments:

Treatment	Nitrogen level	Irrigation
1N	200 mg N/kg soil	No
1Y	200 mg N/kg soil	1 cm/week
2N	400 mg N/kg soil	No
3N	600 mg N/kg soil	No
4N	800 mg N/kg soil	No
4Y	800 mg N/kg soil	1 cm/week

Case Study: Grass/Weed Competition – Data

Big bluestem was first seeded in these plots.
One year later, quack grass was seeded to each plot.

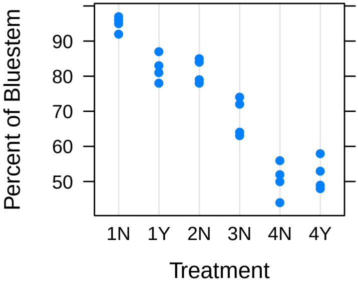
Response: Percentage of living material in each plot that is big bluestem one year after quack grass was seeded.

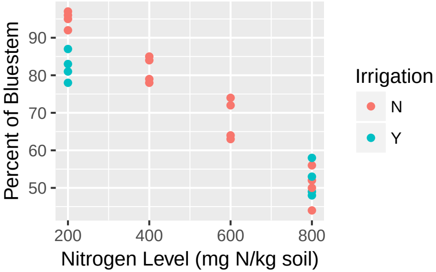
Treatment	1N	1Y	2N	3N	4N	4Y
	97	83	85	64	52	48
	96	87	84	72	56	58
	92	78	78	63	44	49
	95	81	79	74	50	53

Data file: `grassweed.txt`

```
> grass = read.table("grassweed.txt", h=T)
> grass
   percent trt Nlevel Irrigation
1       97  1N    200          N
2       83  1Y    200          Y
3       85  2N    400          N
4       64  3N    600          N
...
```

Case Study: Grass/Weed Competition – Plots






```
grass = read.table("grassweed.txt", h=T)
library(mosaic)
dotplot(percent ~ trt, data=grass,
        ylab = "Percent of Bluestem", xlab ="Treatment")
qplot(Nlevel, percent, color=Irrigation, data=grass,
      ylab="Percent of Bluestem", xlab="Nitrogen Level (mg N/kg soil)")
```

Models for a Completely Randomized Experiment

For an experiment, the N experimental units are randomized to received one of the g treatments, where n_i experimental units received for treatment i, $i = 1, 2, \ldots, g$.

$$
\begin{array}{ll}
\text{Treatment 1}: & y_{11}, y_{12}, \ldots, y_{1n_1} \\
\text{Treatment 2}: & y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2} \\
 & \vdots \\
\text{Treatment g}: & y_{g1}, y_{g2}, \ldots\ldots, y_{gn_g}
\end{array}
$$

$$
\begin{array}{ccccccc}
\text{jth unit for} & & \text{treatment} & & \text{error} & & \\
\text{treatment } i & & \text{effect} & & \text{(or noise)} & & \\
\downarrow & & \downarrow & & \downarrow & & i = 1, 2, \ldots, g \\
y_{ij} & = & \mu_i & + & \varepsilon_{ij} & & j = 1, 2, \ldots, n_i
\end{array}
$$

- μ_i = mean response for the ith treatment
- The error terms ε_{ij} are assumed to be **independent** with mean 0 and **constant variance** σ^2.
  Sometimes we further assume that errors are normal.

Questions of Interest

Unlike a two-sample problem that only compares the two means $\mu_1 - \mu_2$, for a multi-sample problem, there are various comparisons of interest.

For example, the purpose of the Grass/Weed Competition experiment is to see if nitrogen and/or irrigation has any effect on the ability of quack grass to invade big bluestem. The comparisons of interests include

- Irrigation effect: $\mu_{1N} - \mu_{1Y}$, $\mu_{4N} - \mu_{4Y}$ or the combining the two
$$\frac{\mu_{1Y} + \mu_{4Y}}{2} - \frac{\mu_{1N} + \mu_{4N}}{2}$$
- Nitrogen effect: $\mu_{1N} - \mu_{2N}$, $\mu_{2N} - \mu_{3N}$, etc.
- Whether irrigation or nitrogen has any effect

$$\mu_{1N} = \mu_{1Y} = \mu_{2N} = \mu_{3N} = \mu_{4N} = \mu_{4Y}$$

Dot and Bar Notation

A dot ($\bullet$) in subscript means *summing* over that index, for example

$$y_{i\bullet} = \sum_j y_{ij}, \quad y_{\bullet j} = \sum_i y_{ij}, \qquad y_{\bullet\bullet} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$$

A bar over a variable, along with a dot ($\bullet$) in subscript means *averaging* over that index, for example

$$\overline{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \overline{y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$$

Estimation of Means

Recall the least square (LS) estimates $\widehat{\mu}_i$'s are the $\widehat{\mu}_i$'s that *minimize* the *sum of squares* of the observations y_{ij} to their hypothesized means μ_i based on the model,

$$S = \sum_{j=1}^{n_1}(y_{1j} - \widehat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \widehat{\mu}_2)^2 + \cdots + \sum_{j=1}^{n_g}(y_{gj} - \widehat{\mu}_g)^2.$$

To minimize S, we could differentiate it with respect to each μ_i and set the derivative equal to zero.

$$\frac{\partial S}{\partial \widehat{\mu}_i} = -2\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu}_i) = -2n_i(\overline{y}_{i\bullet} - \widehat{\mu}_i) = 0$$

The **least square estimate** for μ_i is thus the **sample mean** of observations in the corresponding treatment group,

$$\widehat{\mu}_i = \overline{y}_{i\bullet} = \frac{1}{n_i}\sum\nolimits_{j=1}^{n_i} y_{ij}.$$

Moreover the LS estimate $\overline{y}_{i\bullet}$ for μ_i is **unbiased**.

Fitted Values and Residuals

- **fitted value** for y_{ij} is $\widehat{y}_{ij} = \widehat{\mu}_i = \overline{y}_{i\bullet}$

- **residual** for y_{ij} is $e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y}_{i\bullet}$

Sum of Squares (1)

$$y_{ij} - \overline{y}_{\bullet\bullet} = (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet})$$

Squaring up both sides, by the identity $(a+b)^2 = a^2+b^2+2ab$, we get

$$(y_{ij} - \overline{y}_{\bullet\bullet})^2 = (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + (y_{ij} - \overline{y}_{i\bullet})^2 + 2(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})(y_{ij} - \overline{y}_{i\bullet})$$

Summing over the indexes i and j, we get

$$\overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\bullet\bullet})^2}^{SST} = \overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}^{SS_{Trt}} + \overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2}^{SSE}$$
$$+ 2\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})(y_{ij} - \overline{y}_{i\bullet})}_{= 0, \text{ see next slide}}$$

Sum of Squares (2)

Observe that

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underbrace{(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})}_{\text{constant in } j}(y_{ij}-\overline{y}_{i\bullet}) = \sum_{i=1}^{g}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})}_{\text{see below}}$$

and

$$\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet}) = y_{i\bullet} - n_i\overline{y}_{i\bullet} = y_{i\bullet} - n_i\left(\frac{y_{i\bullet}}{n_i}\right) = 0$$

and hence

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})(y_{ij}-\overline{y}_{i\bullet}) = 0.$$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{\bullet\bullet})^2}_{SST} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})^2}_{=SS_{Trt}=SSB} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})^2}_{=SSE=SSW}$$

- SST = **total sum of squares**
  - reflects total variability in the response for all the units
- SS_{Trt} = **treatment sum of squares**
  - reflects variability **between** treatments
  - also called **between sum of squares**, denoted as SSB
- SSE = **error sum of squares**
  - Observe that SSE $= \sum_{i=1}^{g}(n_i - 1)s_i^2$, in which

    $$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})^2$$

    is the sample variance **within** treatment group i.
    So SSE reflects the variability **within** treatment groups.
  - also called **within sum of squares**, denoted as SSW

Estimate of the Variance — MSE

Recall in a one sample problem, the population variance σ^2 is estimated by the sample variance

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1} \quad \xrightarrow{\text{estimates}} \sigma^2.$$

In a one way ANOVA problem $y_{ij} = \mu_i + \varepsilon_{ij}$, as all groups have identical variance $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$, the sample variance s_j^2 of any group can estimate σ^2.

$$\begin{aligned}
&\text{Group 1: } s_1^2 \xrightarrow{\text{estimates}} \sigma^2 \\
&\text{Group 2: } s_2^2 \xrightarrow{\text{estimates}} \sigma^2 \\
&\qquad\vdots \\
&\text{Group g: } s_g^2 \xrightarrow{\text{estimates}} \sigma^2
\end{aligned}$$

We can pool all of $s_1^2, s_2^2, \ldots, s_g^2$ to get a better estimate of σ^2.

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_g - 1)s_g^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1)} \\
&= \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2}{N - g} = \frac{\mathrm{SSE}}{N - g} = \mathrm{MSE}
\end{aligned}$$

This estimate is called the **mean square error (MSE)** .

Degrees of Freedom

Under the model $y_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij}'s are i.i.d. $\sim N(0, \sigma^2)$, it can be shown that

$$\frac{\mathsf{SSE}}{\sigma^2} \sim \chi^2_{N-g}.$$

As the mean of a χ^2_k distribution is k, we know that **MSE is an unbiased estimator for σ^2**.

Furthermore if $\mu_1 = \cdots = \mu_g$, then

$$\frac{\mathsf{SST}}{\sigma^2} \sim \chi^2_{N-1}, \quad \frac{\mathsf{SS}_{Trt}}{\sigma^2} \sim \chi^2_{g-1}$$

and SS_{Trt} is independent of SSE.

Note the **degrees of freedom** of the 3 SS

$$dfT = N - 1, \quad df_{Trt} = g - 1, \quad dfE = N - g$$

break down just like $SST = SS_{Trt} + SSE$,

$$dfT = df_{Trt} + dfE$$

One-Way ANOVA Test & ANOVA Table

A one-way ANOVA test is for testing whether the treatments have different effects

$$H_0 : \mu_1 = \cdots = \mu_g \qquad \text{(no difference between treatments)}$$
$$H_a : \mu_i\text{'s not all equal} \qquad \text{(some difference between treatments)}$$

The test statistic is the F-statistic.

$$F = \frac{MS_{Trt}}{\text{MSE}} = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)}$$

which has an F distribution with $g-1$ and $N-g$ degrees of freedom.

Source	Sum of Squares	d.f.	Mean Squares	F
Treatments	SS_{Trt}	$g-1$	$\text{MS}_{Trt} = \frac{SS_{Trt}}{g-1}$	$\frac{MS_{Trt}}{\text{MSE}}$
Errors	SSE	$N-g$	$\text{MSE} = \frac{SSE}{N-g}$	
Total	SST	$N-1$		

Interpretation of the ANOVA F-Statistic

$$H_0 : \mu_1 = \cdots = \mu_g \qquad \text{(no difference between treatments)}$$
$$H_a : \mu_i\text{'s not all equal} \qquad \text{(some difference between treatments)}$$

$$F = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)} = \frac{SSB/(g-1)}{SSW/(N-g)}$$
$$= \frac{\text{Variation Between Groups}}{\text{Variation Within Groups}}$$

The larger the variation between groups relative to variation within each group, the stronger the evidence toward H_a

Case Study: Grass/Weed Competition – SS_{Trt}

Treatment	1N	1Y	2N	3N	4N	4Y
	97	83	85	64	52	48
	96	87	84	72	56	58
	92	78	78	63	44	49
	95	81	79	74	50	53
Mean $\overline{y}_{i\bullet}$	95	82.25	81.5	68.25	50.5	52
SD s_i	2.160	3.775	3.512	5.560	5.000	4.546

$\overline{y}_{\bullet\bullet} = \frac{1}{6}(95 + 82.25 + 81.5 + 68.25 + 50.5 + 52) = \frac{429.5}{6} = 71.583$

The **between** group sum of squares

$$
\begin{aligned}
SS_{Trt} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 = \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \\
&= 4(95 - 71.583)^2 + 4(82.25 - 71.583)^2 + 4(81.5 - 71.583)^2 \\
&\quad + 4(68.25 - 71.583)^2 + 4(50.5 - 71.583)^2 + 4(52 - 71.583)^2 \\
&\approx 6398.333
\end{aligned}
$$

Case Study: Grass/Weed Competition — SSE

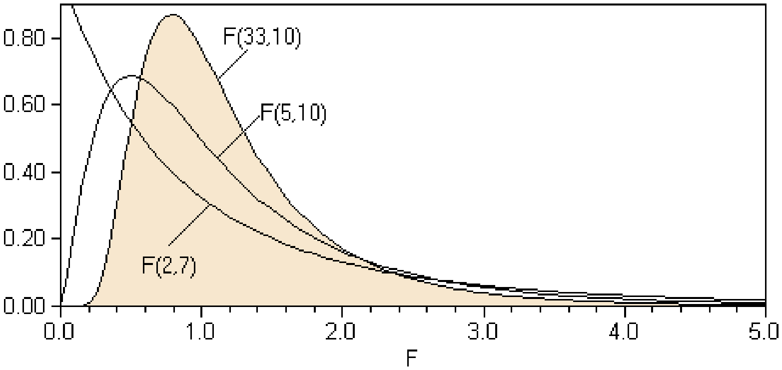
Treatment	1N	1Y	2N	3N	4N	4Y
Mean $\overline{y}_{i\bullet}$	95	82.25	81.5	68.25	50.5	52
SD s_i	2.160	3.775	3.512	5.560	5.000	4.546

$$
\begin{aligned}
SSE &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2 = \sum_{i=1}^{g}(n_i - 1)s_i^2 \\
&= (4-1)(2.160^2 + 3.775^2 + 3.512^2 + 5.560^2 + 5.000^2 + 4.546^2) \\
&\approx 323.4903
\end{aligned}
$$

Case Study: Grass/Weed Competition — ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment	$g-1=$ $6-1=5$	$\text{SS}_{trt}=6398.3$	$\text{MS}_{trt}=\text{SS}_{trt}/df_{trt}$ $=6398.3/5\approx 1279.67$	$F=\text{MS}_{trt}/\text{MSE}$ $=\frac{1279.67}{17.97}\approx 71.2$
Error	$N-g=$ $24-6=18$	$\text{SSE}=323.49$	$\text{MSE}=\text{SSE}/df_E$ $=323.49/18\approx 17.97$	

The F Distributions



An F-distribution has two parameters df1 and df2.
There is one F-density curve with each pair of values of df1 and df2.

P-value of the One-Way ANOVA Test

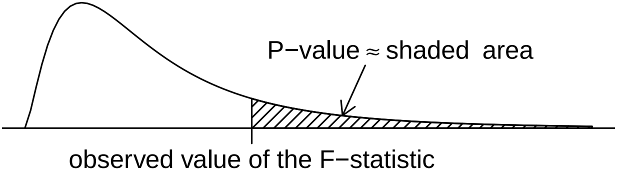
The one-way ANOVA F-statistic

$$F = \frac{MS_{Trt}}{\text{MSE}} = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)}$$

which has an F distribution with $g-1$ and $N-g$ degrees of freedom.

Under H_0: all μ_i's being equal, the P-value is the area of the upper-tail under the F-curve with $g-1$ and $N-g$ degrees of freedom beyond the F statistic.

F-curve with $g-1$ and $N-g$ degrees of freedom



Finding the P-value in R

For the Grass/Weed experiment, the P-value for the F-statistic 71.2 is

$$P\text{-value} = P(F_{5,18} \geq 71.2) = 3.197 \times 10^{-11}.$$


```
> pf(71.2, df1=5, df2=18, lower.tail = F)
[1] 3.198094e-11
```

Conclusion: The data exhibit strong evidence against the H_0 that all means are equal.

Finding the P-value using the F-table (p.627-628)

ANOVA F-Test in R

```
> lm1 = lm(percent ~ trt, data=grass)
> anova(lm1)
Analysis of Variance Table

Response: percent
          Df Sum Sq Mean Sq F value    Pr(>F)
trt        5 6398.3 1279.67  71.203 3.197e-11 ***
Residuals 18  323.5   17.97
```

What Does "ANOVA" Stands For?

"ANOVA" is the shorthand for "<u>AN</u>alysis <u>O</u>f <u>VA</u>riance." Specifically, it is a class of statistical methods that break up the variability of the response into different sources of variations, like

$$\text{SST} = \text{SS}_{trt} + \text{SSE}$$

Throughout STAT 22200, we will introduce several other ANOVA for different models (two-way ANOVA, three-way ANOVA, ANOVA for block designs, and so on.)