One-Way ANOVA
Comparison of Several Means

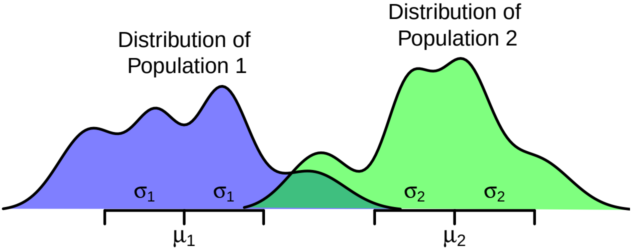
Yibi Huang

Textbook Chapter 3

Two Sample Problems (Review)

Population 1

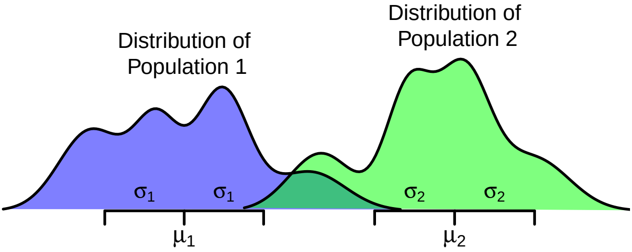
Population 2



Two Sample Problems (Review)

Population 1

Population 2

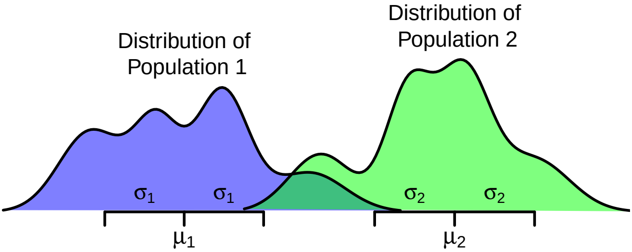


Population distributions may NOT be normal or of the same shape for large samples.

Two Sample Problems (Review)

Population 1

Population 2

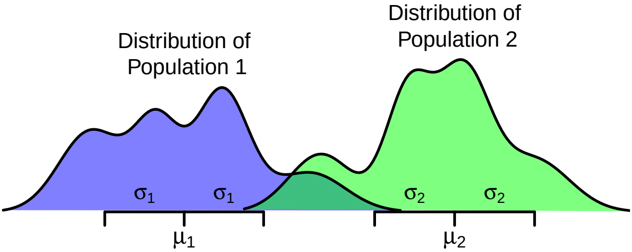


Population SDs σ_1 and σ_2 may not be equal.

Two Sample Problems (Review)

Population 1

Population 2

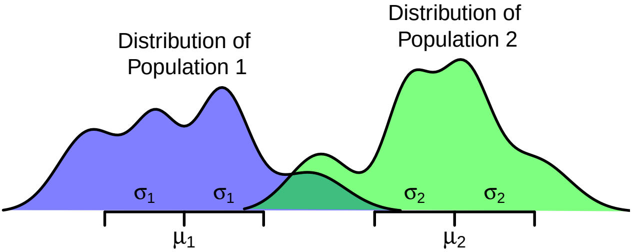


Goal: inference about difference of population means $\mu_1 - \mu_2$.

Two Sample Problems (Review)

Population 1

Population 2



Data may come from experiments or observational studies.

Model for Two-Sample Data (Review)

Data from an Observational Study:

Population 1 $\longrightarrow$ random sample $y_{11}, y_{12}, \ldots, y_{1n_1}$
Population 2 $\longrightarrow$ random sample $y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2}$

Data from a Randomized Experiment:

Treatment 1 $\longrightarrow$ observations $y_{11}, y_{12}, \ldots, y_{1n_1}$
Treatment 2 (Control) $\longrightarrow$ observations $y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2}$

In both cases, we assume

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij}\text{'s are i.i.d.} \sim (0, \sigma_i^2)$$
$$\text{for } i = 1, 2,\ j = 1, \ldots, n_i$$

Two-Sample t-Statistic When $\sigma_1 = \sigma_2$ (Review)

Assuming $\sigma_1 = \sigma_2$, the two-sample t-statistic is

$$t = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$s_p^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2}{n_1 + n_2 - 2},$$

called the "pooled sample variance", is an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

- If the noise ε_{ij} are i.i.d. $N(0, \sigma^2)$, the t-statistic has an exact t-distribution with df $= n_1 + n_2 - 2$, regardless of the sample size n_1 and n_2
- If the noise ε_{ij} are indep. $(0, \sigma^2)$ but **not normal**, the t-statistic has an approx t-distribution with df $= n_1 + n_2 - 2$ when n_1 and n_2 are large

Two-Sample t-Statistic When $\sigma_1 \neq \sigma_2$ (Review)

When $\sigma_1 \neq \sigma_2$, we use the Welch t-statistic

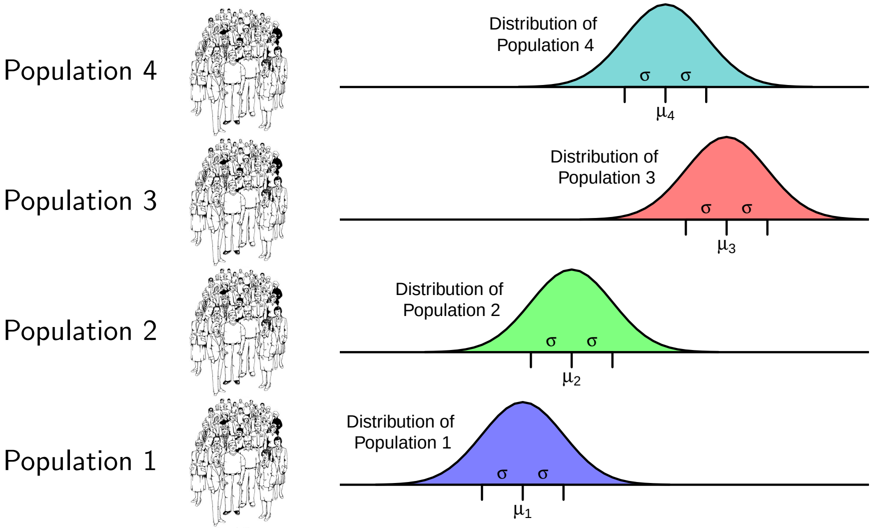
$$t = \frac{\overline{y}_1 - \overline{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}, \quad \text{where} \quad \begin{array}{l} s_1^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2}{n_1 - 1} \\ s_2^2 = \frac{\sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2}{n_2 - 1} \end{array}$$

- Even if the noise ε_{ij}'s are normal, the Welch t-statistic does NOT have a t-distribution, but it can be approximated by a t-distribution with

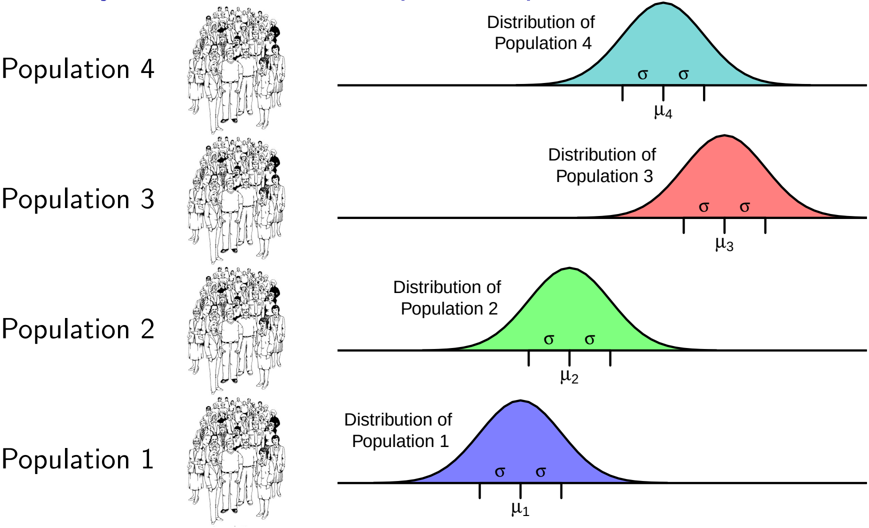
$$df = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}.$$

- When the noise ε_{ij}'s are not normal, the t-approximation above is generally good as long as the sample sizes n_1 and n_2 is large

One Way ANOVA — Multiple-Sample Problems

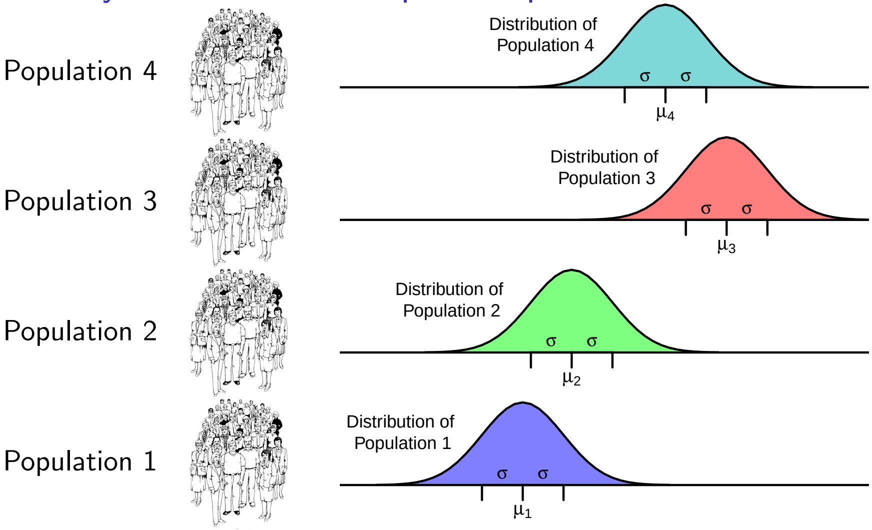


One Way ANOVA — Multiple-Sample Problems



All population distributions are assumed to be normal.
The non-normal case will be discussed in Chapter 6.

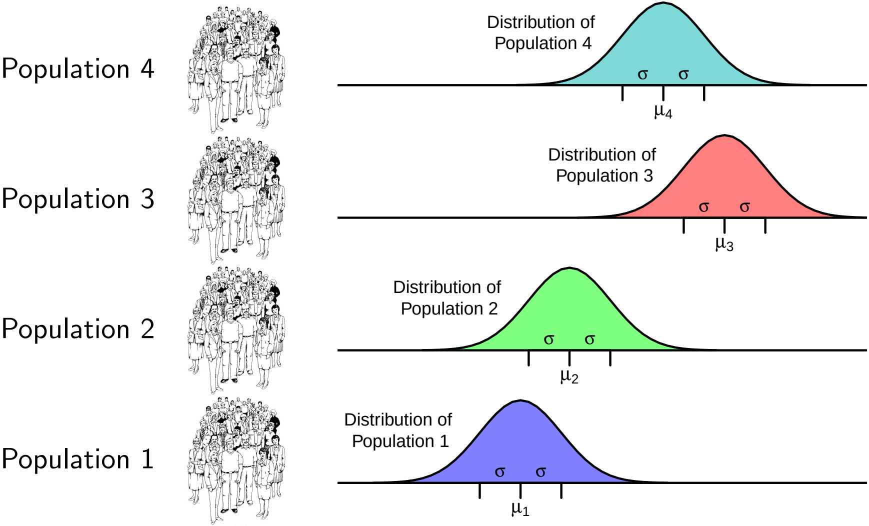
One Way ANOVA — Multiple-Sample Problems



All populations have an identical SD.

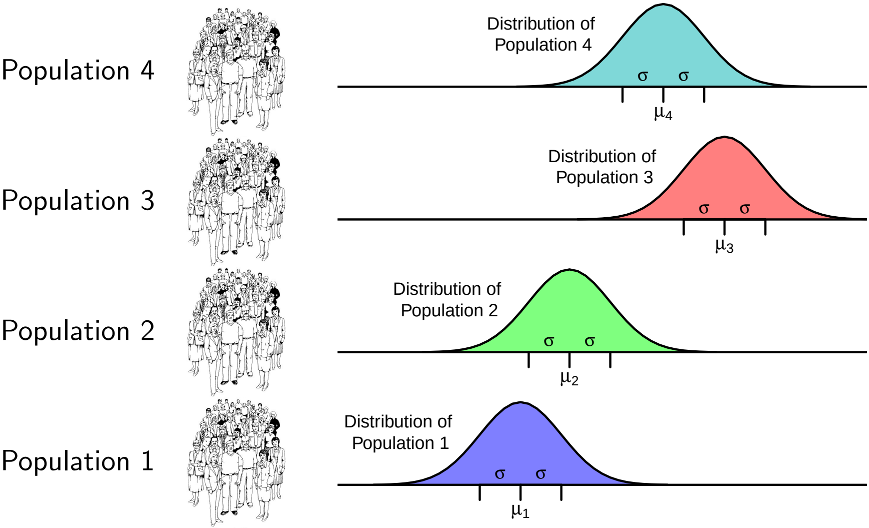
The unequal SDs case will be discussed in Chapter 6.

One Way ANOVA — Multiple-Sample Problems



Goal: comparison of different population means μ_i's.

One Way ANOVA — Multiple-Sample Problems



Data may come from experiments or observational studies.

Models for a 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, $i = 1, 2, \ldots, g$.

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

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

- 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.

Model for a Multi-Sample Observational Study

Data

Random Sample from Population 1 : $y_{11}, y_{12}, \ldots, y_{1n_1}$
Random Sample from Population 2 : $y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2}$
$\vdots$
Random Sample from Population g : $y_{g1}, y_{g2}, \ldots\ldots, y_{gn_g}$

*j*th observation in the *i*th sample		population mean		error (or noise)	
$\downarrow$		$\downarrow$		$\downarrow$	$i = 1, 2, \ldots, g$
y_{ij}	$=$	μ_i	$+$	ε_{ij}	$j = 1, 2, \ldots, n_i$

For both multi-treatment randomized experiments and multiple-sample observational studies, the format of the model and the analysis are the same.

Case Study: Grass/Weed Competition

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.

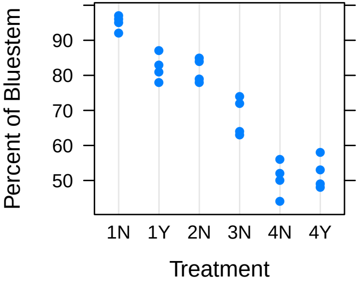
Response: Percentage of living material in each plot that is big bluestem one year after quack grass was seeded.

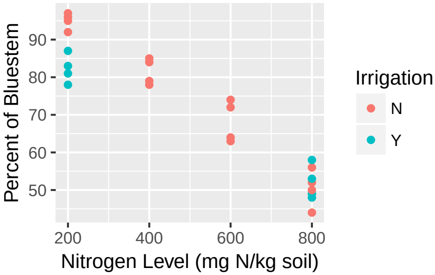
Treatment	1N	1Y	2N	3N	4N	4Y
	97	83	85	64	52	48
	96	87	84	72	56	58
	92	78	78	63	44	49
	95	81	79	74	50	53

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("http://www.stat.uchicago.edu/~yibi/s222/grassweed.t
library(ggplot2)
ggplot(grass, aes(x=trt,y=percent)) + geom_point()+
  ylab("Percent of Bluestem")+ xlab("Treatment")
ggplot(grass, aes(x=Nlevel,y=percent,color=Irrigation)) + geom_point()+
  ylab("Percent of Bluestem")+ xlab("Treatment")
```

Questions of Interest

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

Questions of Interest

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

E.g., 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

$$\frac{\mu_{1Y} + \mu_{4Y}}{2} - \frac{\mu_{1N} + \mu_{4N}}{2}$$

Questions of Interest

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

E.g., 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.

Questions of Interest

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

E.g., 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

Questions of Interest

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

E.g., 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

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

Estimate of Means, Fitted Values, and Residuals

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

- **Estimate** for μ_i is simply the **sample mean** of observations in the corresponding sample/treatment group,

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

Estimate of Means, Fitted Values, and Residuals

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

- **Estimate** for μ_i is simply the **sample mean** of observations in the corresponding sample/treatment group,

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

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

Estimate of Means, Fitted Values, and Residuals

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

- **Estimate** for μ_i is simply the **sample mean** of observations in the corresponding sample/treatment group,

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

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

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

Sum of Squares (1)

$$y_{ij} - \overline{y}_{\bullet\bullet} = (\overbrace{\overline{y}_{i\bullet}}^{\text{add a term}} - \overline{y}_{\bullet\bullet}) + (y_{ij} \overbrace{- \overline{y}_{i\bullet}}^{\text{subtract a term}})$$

Sum of Squares (1)

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

Sum of Squares (1)

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

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

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

Sum of Squares (1)

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

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

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

Summing over the indexes i and j, we get

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

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

Sum of Squares (2)

Observe that

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underbrace{(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})}_{\text{constant in } j}(y_{ij}-\overline{y}_{i\bullet}) \qquad \text{since } \sum_i c x_i = c\sum_i x_i$$

$$= \sum_{i=1}^{g}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})}_{=0\text{, see below}}$$

Sum of Squares (2)

Observe that

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} \underbrace{(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})}_{\text{constant in } j}(y_{ij} - \overline{y}_{i\bullet}) \qquad \text{since } \sum_i c x_i = c \sum_i x_i$$

$$= \sum_{i=1}^{g} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) \underbrace{\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})}_{=0, \text{ see below}}$$

because

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

Sum of Squares (2)

Observe that

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underbrace{(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})}_{\text{constant in } j}(y_{ij}-\overline{y}_{i\bullet}) \qquad \text{since} \sum_i cx_i = c\sum_i x_i$$

$$= \sum_{i=1}^{g}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})}_{=0,\text{ see below}} = 0$$

because

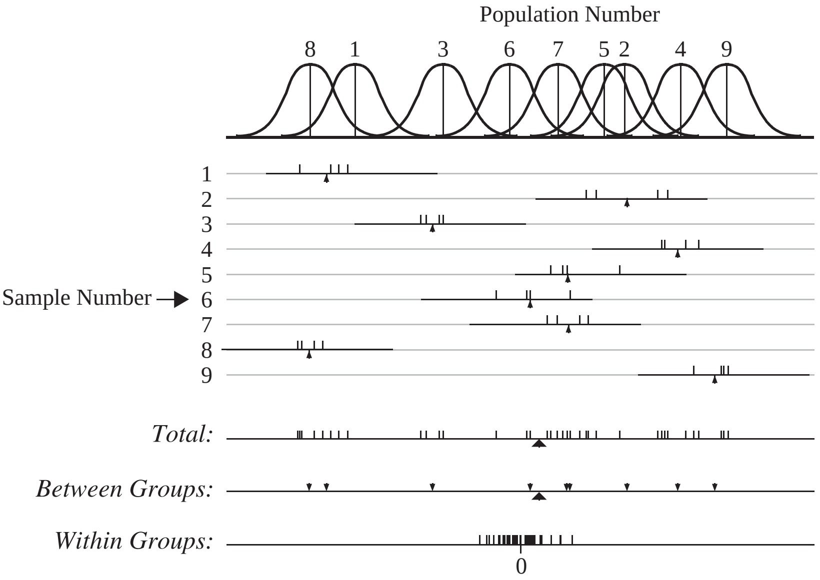
$$\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet}) = y_{i\bullet} - n_i\overline{y}_{i\bullet} = y_{i\bullet} - n_i\left(\frac{y_{i\bullet}}{n_i}\right) = 0$$

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

- 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

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

    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


Population Number
8 1 3 6 7 5 2 4 9
1
2
3
4
5
Sample Number 6
7
8
9
Total:
Between Groups:
Within Groups:
0

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

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

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

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

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

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

and SS_{trt} is independent of SSE.

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

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

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

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

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 $SST = SS_{trt} + SSE$,

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

Mean Squares

The mean squares are the sum of squares divided by the corresponding degrees of freedom.

- MSE = Mean Square Error $= \dfrac{\text{SSE}}{\text{dfE}} = \dfrac{\text{SSE}}{N - g}$
- MS_{trt} = Mean Square for Treatment $= \dfrac{\text{SS}_{trt}}{\text{df}_{trt}} = \dfrac{\text{SS}_{trt}}{g - 1}$

Estimate of the Variance — MSE (1)

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

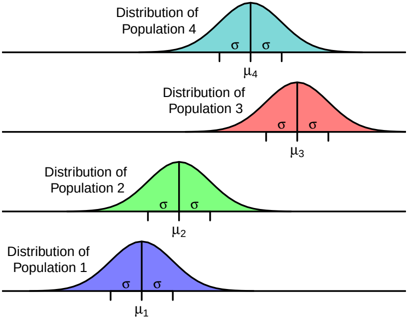
$$s^2 = \frac{\sum_i (y_i - \overline{y})^2}{n-1} \quad \xrightarrow{\text{estimates}} \sigma^2.$$

Estimate of the Variance — MSE (1)

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

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

For the model $y_{ij} = \mu_i + \varepsilon_{ij}$, as all groups have identical variance $\text{Var}(\varepsilon_{ij}) = \sigma^2$, the sample variance s_j^2 of any group can estimate the common variance σ^2.



Group 1: $s_1^2 \xrightarrow{\text{estimates}} \sigma^2$

Group 2: $s_2^2 \xrightarrow{\text{estimates}} \sigma^2$

$\vdots$

Group g: $s_g^2 \xrightarrow{\text{estimates}} \sigma^2$

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

Estimate of the Variance — MSE (2)

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

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

which is simply the **mean square error (MSE)**.

Mean of MSE

Recall in a one sample problem, $y_1, \cdots, y_n$ are *i.i.d.* with variance $\mathrm{Var}(Y_i) = \sigma^2$, then the sample variance s^2 is an unbiased estimate of the variance:

$$\mathbb{E}(s^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2\right) = \sigma^2.$$

Mean of MSE

Recall in a one sample problem, $y_1, \cdots, y_n$ are *i.i.d.* with variance $\mathrm{Var}(Y_i) = \sigma^2$, then the sample variance s^2 is an unbiased estimate of the variance:

$$\mathbb{E}(s^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2\right) = \sigma^2.$$

For a multi-sample (one-way ANOVA) problem $y_{ij} = \mu_i + \varepsilon_{ij}$, we know $y_{i1}, \ldots, y_{in_i}$ are i.i.d. with $\mathrm{Var}(y_{ij}) = \sigma^2$. Thus the sample variance **within** treatment group i

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

is an unbiased estimator of σ^2. Thus

$$\mathbb{E}(\mathrm{SSE}) = \mathbb{E}\Big(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2\Big) = \mathbb{E}\left(\sum_{i=1}^{g}(n_i - 1)s_i^2\right)$$
$$= \sum_{i=1}^{g}(n_i - 1)\sigma^2 = (N - g)\sigma^2$$

So MSE$=$ SSE$/(N - g)$ is an unbiased estimator of σ^2.

One-Way ANOVA Test

A one-way ANOVA test is for testing whether the treatments have different effects or whether the population means are different

$H_0 : \mu_1 = \cdots = \mu_g$ (no diff. btw. treatments/population means)

$H_a : \mu_i$'s not all equal (some diff. btw treatments/population means)

One-Way ANOVA Test

A one-way ANOVA test is for testing whether the treatments have different effects or whether the population means are different

$H_0 : \mu_1 = \cdots = \mu_g$ (no diff. btw. treatments/population means)
$H_a : \mu_i$'s not all equal (some diff. btw treatments/population means)

- Under H_0: $\mu_1 = \cdots = \mu_g$, we expect

$$\bar{y}_{1\bullet} \approx \bar{y}_{2\bullet} \approx \ldots \approx \bar{y}_{g\bullet}$$

which implies $\bar{y}_{i\bullet} \approx \bar{y}_{\bullet\bullet}$ for all i. Hence a large value of $SS_{trt} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$ is evidence against H_0.

One-Way ANOVA Test

A one-way ANOVA test is for testing whether the treatments have different effects or whether the population means are different

$H_0 : \mu_1 = \cdots = \mu_g$ (no diff. btw. treatments/population means)

$H_a : \mu_i$'s not all equal (some diff. btw treatments/population means)

- Under H_0: $\mu_1 = \cdots = \mu_g$, we expect

$$\overline{y}_{1\bullet} \approx \overline{y}_{2\bullet} \approx \ldots \approx \overline{y}_{g\bullet}$$

  which implies $\overline{y}_{i\bullet} \approx \overline{y}_{\bullet\bullet}$ for all i. Hence a large value of $\mathrm{SS}_{trt} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$ is evidence against H_0.
- Larger $\sigma^2 \Rightarrow$ Large noise $\Rightarrow$ Greater variability btw $\overline{y}_{i\bullet}$'s even if H_0 is true. Hence, should consider $\dfrac{\mathrm{SS}_{trt}}{\sigma^2}$

One-Way ANOVA Test

A one-way ANOVA test is for testing whether the treatments have different effects or whether the population means are different

$H_0 : \mu_1 = \cdots = \mu_g$ (no diff. btw. treatments/population means)

$H_a : \mu_i$'s not all equal (some diff. btw treatments/population means)

- Under H_0: $\mu_1 = \cdots = \mu_g$, we expect

$$\overline{y}_{1\bullet} \approx \overline{y}_{2\bullet} \approx \ldots \approx \overline{y}_{g\bullet}$$

  which implies $\overline{y}_{i\bullet} \approx \overline{y}_{\bullet\bullet}$ for all i. Hence a large value of $\mathrm{SS}_{trt} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$ is evidence against H_0.
- Larger $\sigma^2 \Rightarrow$ Large noise $\Rightarrow$ Greater variability btw $\overline{y}_{i\bullet}$'s even if H_0 is true. Hence, should consider $\dfrac{\mathrm{SS}_{trt}}{\sigma^2}$
- The unknown σ^2 is estimated by MSE.

ANOVA F-Statistic

The test statistic is hence the F-statistic.

$$F = \frac{SS_{trt}/(g-1)}{MSE}$$

ANOVA F-Statistic

The test statistic is hence the F-statistic.

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

The larger the variation between groups relative to variation within each group, the stronger the evidence against H_0 and toward H_a

the ANOVA Table

The ANOVA F-statistic

$$F = \frac{\text{SS}_{trt}/(g-1)}{\text{SSE}/(N-g)} = \frac{\text{MS}_{trt}}{\text{MSE}}$$

has an F distribution with $g-1$ and $N-g$ degrees of freedom and is often calculated and displayed in an ANOVA table as follows.

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	$N-1$		

- The last row (Total) is omitted in R output

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	...	$\overline{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

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

$$\mathrm{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 =$$

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

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	...	$\overline{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

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

$$\text{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 =$$

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

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	...	$\overline{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

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

$$\text{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 =$$

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

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	...	$\overline{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

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

$$\text{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\underbrace{(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}_{\text{constant in } j} =$$

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

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$	...	$\bar{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

$$\bar{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{g}\underbrace{\sum\nolimits_{j=1}^{n_i} y_{ij}}_{=y_{i\bullet}=n_i\bar{y}_{i\bullet}} = \frac{1}{N}\sum_{i=1}^{g} n_i\bar{y}_{i\bullet}$$

$$\text{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\underbrace{(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2}_{\text{constant in } j} = \sum_{i=1}^{g} n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2$$

$$\text{SSE} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2 =$$

Calculating SS from the Group Means and Group SDs

Group	1	2	...	g
Group Mean	$\overline{y}_{1\bullet}$	$\overline{y}_{2\bullet}$	...	$\overline{y}_{g\bullet}$
Group SD	s_1	s_2	...	s_g

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

$$\text{SS}_{trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \underbrace{(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}_{\text{constant in } j} = \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$$

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

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

Case Study: Grass/Weed Competition – SS_{trt} and SSE

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

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{g} n_i \overline{y}_{i\bullet}$$

$$SS_{trt} = \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$$

$$SSE = \sum_{i=1}^{g}(n_i - 1)s_i^2$$

Case Study: Grass/Weed Competition – SS_{trt} and SSE

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

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{g} n_i \overline{y}_{i\bullet} = \frac{4}{24}(95 + 82.25 + 81.5 + 68.25 + 50.5 + 52) \approx 71.583$$

$$SS_{trt} = \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$$

$$SSE = \sum_{i=1}^{g}(n_i - 1)s_i^2$$

Case Study: Grass/Weed Competition – SS_{trt} and SSE

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

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{g} n_i \overline{y}_{i\bullet} = \frac{4}{24}(95 + 82.25 + 81.5 + 68.25 + 50.5 + 52) \approx 71.583$$

$$\begin{aligned} SS_{trt} &= \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \\ &= 4(95-71.583)^2 + 4(82.25-71.583)^2 + 4(81.5-71.583)^2 \\ &\quad + 4(68.25-71.583)^2 + 4(50.5-71.583)^2 + 4(52-71.583)^2 \approx 6398.33 \\ SSE &= \sum_{i=1}^{g}(n_i - 1)s_i^2 \end{aligned}$$

Case Study: Grass/Weed Competition – SS_{trt} and SSE

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

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{g} n_i \overline{y}_{i\bullet} = \frac{4}{24}(95 + 82.25 + 81.5 + 68.25 + 50.5 + 52) \approx 71.583$$

$$\begin{aligned}
SS_{trt} &= \sum\nolimits_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \\
&= 4(95-71.583)^2 + 4(82.25-71.583)^2 + 4(81.5-71.583)^2 \\
&\quad + 4(68.25-71.583)^2 + 4(50.5-71.583)^2 + 4(52-71.583)^2 \approx 6398.33 \\
SSE &= \sum\nolimits_{i=1}^{g}(n_i - 1)s_i^2 \\
&= (4-1)(2.16^2 + 3.775^2 + 3.512^2 + 5.56^2 + 5^2 + 4.546^2) \approx 323.49
\end{aligned}$$

Case Study: Grass/Weed Competition — ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment		$SS_{trt} =$ 6398.3		
Error		$SSE =$ 323.49		

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		
Error	$N-g=$ $24-6=18$	SSE $=$ 323.49		

Case Study: Grass/Weed Competition — ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment	$g-1=$	$SS_{trt} =$	$MS_{trt} = SS_{trt}/df_{trt}$	
	$6-1=5$	6398.3	$=6398.3/5 \approx 1279.67$	
Error	$N-g=$	$SSE=$	$MSE = SSE/dfE$	
	$24-6=18$	323.49	$=323.49/18 \approx 17.97$	

Case Study: Grass/Weed Competition — ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment	$g-1=$	$\text{SS}_{trt}=$	$\text{MS}_{trt}=\text{SS}_{trt}/df_{trt}$	$F=\text{MS}_{trt}/\text{MSE}$
	$6-1=5$	6398.3	$=6398.3/5\approx1279.67$	$=\frac{1279.67}{17.97}\approx71.2$
Error	$N-g=$	$\text{SSE}=$	$\text{MSE}=\text{SSE}/dfE$	
	$24-6=18$	323.49	$=323.49/18\approx17.97$	

Case Study: Grass/Weed Competition — ANOVA Table

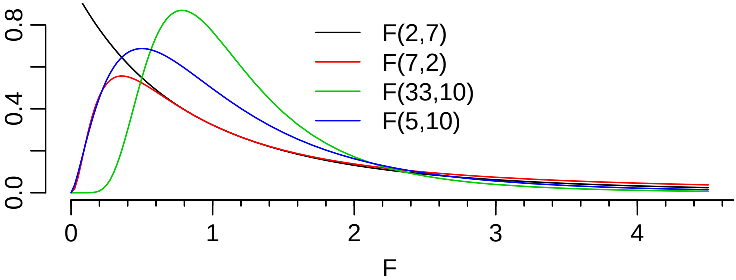
Source	df	Sum of Squares	Mean Squares	F
Treatment	$g-1=$	$SS_{trt}=$	$MS_{trt}=SS_{trt}/df_{trt}$	$F=MS_{trt}/MSE$
	$6-1=5$	6398.3	$=6398.3/5\approx1279.67$	$=\frac{1279.67}{17.97}\approx71.2$
Error	$N-g=$	$SSE=$	$MSE=SSE/dfE$	
	$24-6=18$	323.49	$=323.49/18\approx17.97$	

ANOVA table 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
```

The F Distributions



- An F-distribution has two parameters df1 and df2.
- There is one F-density for each pair of df1 and df2.
- The order of df1 and df2 matters.
  e.g., $F(2,7)$ and $F(7,2)$ are different F-distributions.

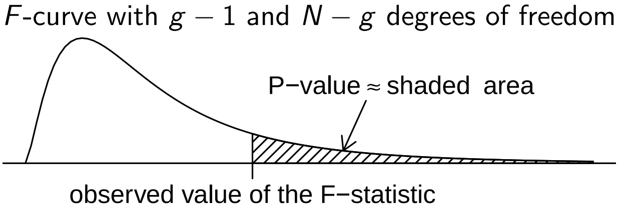
P-value of the One-Way ANOVA Test

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.



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)

Table entries are $F_{.05,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.05,\nu_1,\nu_2}) = .05$.

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03

- The *F*-table above gives the **critical value** at 0.05 significance level for deciding if H_0 should be rejected

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

Table entries are $F_{.05,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.05,\nu_1,\nu_2}) = .05$. significance level

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03

- The F-table above gives the **critical value** at 0.05 significance level for deciding if H_0 should be rejected
- For df1 = 5, df2 = 18, if the F-statistic exceeds $F_{0.05,df1=5,df2=18} = 2.71$, p-value < 0.01 and H_0 is rejected at 0.01 level

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

Table entries are $F_{.01,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.01,\nu_1,\nu_2}) = .01$.

	ν_1															
ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.45	9.38	9.29
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69

- The F-table above gives the **critical value** at 0.01 significance level for deciding if H_0 should be rejected

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

Table entries are $F_{.01,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.01,\nu_1,\nu_2}) =$ .01 significance level

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.45	9.38	9.29
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69

- The F-table above gives the **critical value** at 0.01 significance level for deciding if H_0 should be rejected
- For df1 $= 5$, df2 $= 18$, if the F-statistic exceeds $F_{0.01,df1=5,df2=18} = 4.25$, p-value < 0.01 and H_0 is rejected at 0.01 level

What Does "ANOVA" Stands For?

"ANOVA" is the shorthand for "ANalysis Of VAriance."
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.)

Experimental Units v.s. Measurement Units

Experimental units are the smallest groupings of the experimental material that could have gotten different treatments.

Measurement units are the actual objects on which the response is measured.

- In many cases, the measurement units are just the experimental units
- Sometimes a measurement unit is only *part* of an experimental unit.

Experimental Units v.s. Measurement Units

- 12 pens of young turkeys are randomly assigned 3 different diets (20 turkeys per pen)
  - A measurement unit is one turkey, and an experimental unit is a whole pen of turkeys.
  - Sample size is 4 per diet, not 4×20 per diet

- A class full of students is assigned a certain pedagogical intervention.
  - Suppose classes of students are assigned to two different pedagogy scheme. A measurement unit is one student, and an experimental unit is a whole class of students