MATH 4720 / MSSC 5720

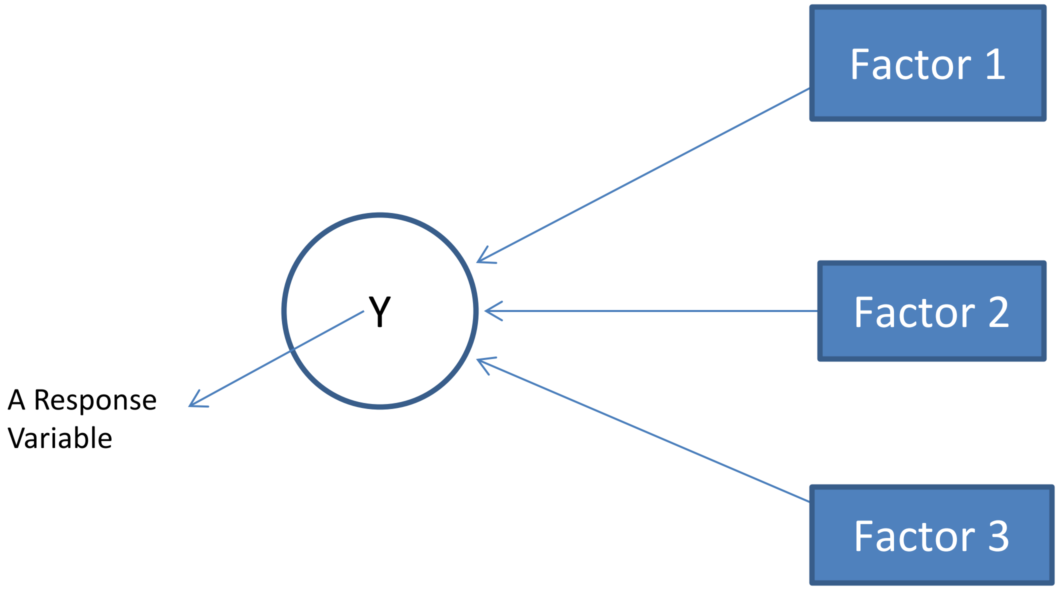
Instructor: Mehdi Maadooliat

Chapter 8 (Part A)

Department of Mathematical and Statistical Sciences

ANALYSIS OF VARIANCE (ANOVA)

- **ANOVA is one of the most popular statistical tools of analyzing data.**

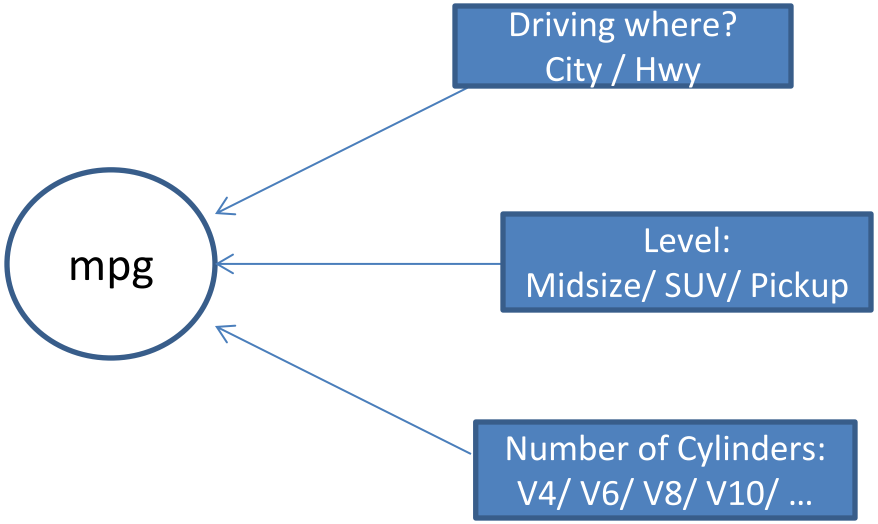


- **Does Y (the response) depends on any of the factors?**

ANOVA EXAMPLES

- **Example 1: You are doing a research on mpg (miles per gallon) for a brand of automobiles.**

- **Question: What effects mpg?**

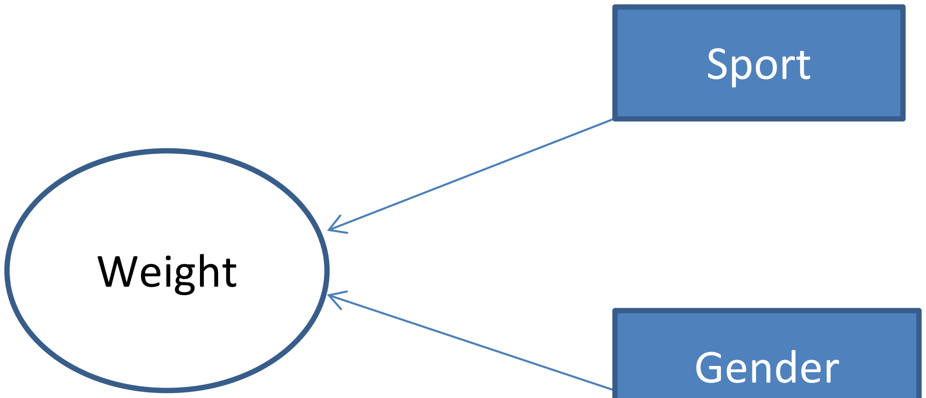


- **Does Y (the response) depends on any of the factors?**

ANOVA EXAMPLESLE

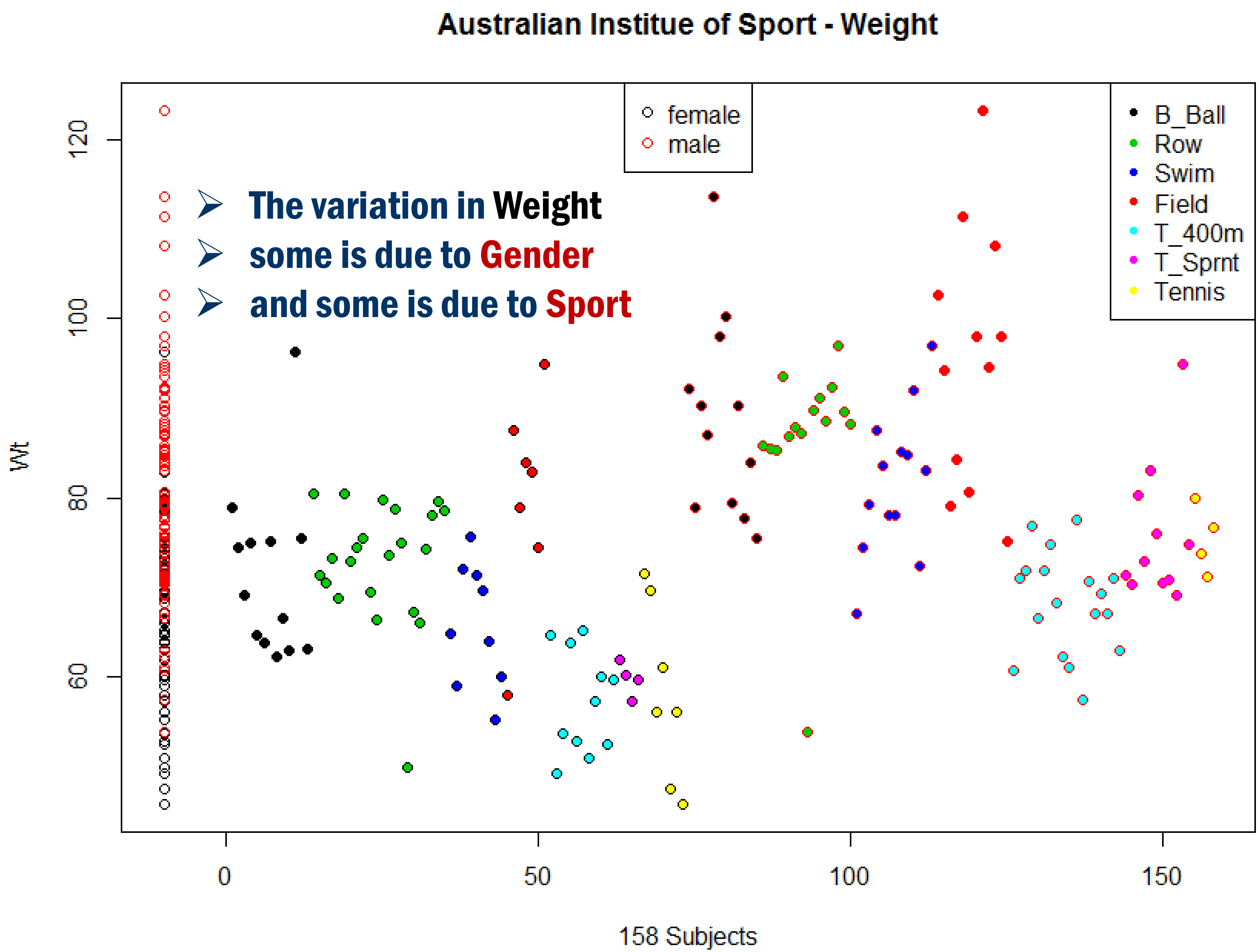
- **Example 2: (Australian Institute of Sport)**

- **Research Question: Does body weight (Wt) depend on**
  - Sport
  - Gender



- **In R (Australian Institute of Sport)**

EXAMPLE CONT'D



CONCEPT

- **Variation(Wt) =**

  Variation(Gender) + Variation(Sport)

  + Variation(Error)

- **These variation can be described by Sums of Squares** $\sum(\ldots)^2$

- **SS(Wt) = SS(Gender) + SS(Sport) + SS(Error)**
- $df_W = df_G + df_S + df_E$

- df **is the degrees of freedom that represent the effective number of terms in the sums of squares**

- **In R:** aov(Wt ~ sport + gender, data=ais2)

TEST STATISTICS

- **F-Statistics**

- **Gender:** **Test Statistics** $F_1 = \dfrac{\frac{SS(Gender)}{df_G}}{\frac{SS(Error)}{df_E}} = \dfrac{MS_G}{MSE}$

  - **If** $F_1 > F_\alpha(df_G, df_E)$, **then**

    gender is a significant factor

- **Sport:** **Test Statistic** $F_2 = \dfrac{\frac{SS(Sport)}{df_S}}{\frac{SS(Error)}{df_E}} = \dfrac{MS_S}{MSE}$

  - **If** $F_2 > F_\alpha(df_S, df_E)$, **then**

    sport is a significant factor

BACK TO CONCEPT

- The sums of squares are not always easily available. For different factor-designs, there are different sums of squares.

- For One-Factor design, sums of squares are easy to compute.

- **One Factor ANOVA:**

Treatment Levels

	1	2	3	. . .	t
	y_{11}	y_{21}	y_{31}		y_{t1}
	y_{12}	y_{22}	y_{32}		y_{t2}
	.	.	.		.
	.	.	.		.
	y_{1n_1}	y_{2n_2}	y_{3n_3}		y_{tn_t}
	=====	=====	=====	=====	=====
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\bar{y}_{3.}$		$\bar{y}_{t.}$
St.dev.	s_1	s_2	s_3		s_t

BACK TO CONCEPT

- **Grand Mean:** $\bar{y}_{..} = \frac{n_1\bar{y}_{1.}+n_2\bar{y}_{2.}+\cdots+n_t\bar{y}_{t.}}{\mathrm{n}_1+\mathrm{n}_2+\cdots+\mathrm{n}_t}$

- **Total Variability:** $\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2$
- **Variability Between Samples:** $\sum_{i=1}^{t} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$
- **Variability Within Samples:** $\sum_{i=1}^{t}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$

- $$\sum\sum\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum\sum\left(y_{ij}-\bar{y}_{i.}\right)^2$$
$$SS(Total) = SSB + SSE$$

df's: $\underset{df_{Total}}{\sum n_i - 1}$ $\quad$ $\underset{df_B}{t-1}$ $\quad$ $\underset{df_E}{\sum n_i - t}$

BACK TO CONCEPT

- $\sum\sum\left(y_{ij}-\bar{y}_{..}\right)^2=\sum n_i(\bar{y}_{i.}-\bar{y}_{..})^2+\sum\sum\left(y_{ij}-\bar{y}_{i.}\right)^2$

$$SS(Total) \quad = \quad SSB \quad + \quad SSE$$

df's: $\underbrace{\sum n_i - 1}_{df_{Total}}$ $\quad \underbrace{t-1}_{df_B}$ $\quad \underbrace{\sum n_i - t}_{df_E}$

- H_0**: There is no difference between the Treatments**
- H_a**: At least one of the treatment is different from the rest**

- **Test Statistics:** $F=\dfrac{SS_B/df_B}{SS_E/df_E}$
- **Decision Rule:**

  The Factor is significant if $F>F_\alpha(df_1=df_B,\ df_2=df_E)$

- **Applet**

FORMULATION IN TERMS OF HYPOTHESIS PROBLEM

- **One Factor ANOVA:**

Treatment Levels

1	2	3	. . .	t
y_{11}	y_{21}	y_{31}	. . .	y_{t1}
y_{12}	y_{22}	y_{32}	. . .	y_{t2}
.	.	.		.
.	.	.		.
y_{1n_1}	y_{2n_2}	y_{3n_3}	. . .	y_{tn_t}
$N(\mu_1, \sigma_1^2)$	$N(\mu_2, \sigma_2^2)$	$N(\mu_3, \sigma_3^2)$	. . .	$N(\mu_t, \sigma_t^2)$

- $H_0: \mu_1 = \mu_2 = \cdots = \mu_t$
- $H_a: \ \mu_i \neq \mu_j$ **for some pairs** (i, j)

FORMULATION IN TERMS OF HYPOTHESIS PROBLEM CONT'D

- H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$
- H_a: $\mu_i \neq \mu_j$ **for some pairs** (i, j)

- **Assumptions:**
  - $\sigma_1 = \sigma_2 = \cdots = \sigma_t$
  - **Data is generated from normal distribution for each treatment.**

- **TS** $F = \dfrac{SS_B/df_B}{SS_E/df_E}$
  - $SS_B = \sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ $\quad\quad df_B = t - 1$
  - $SS_E = \sum (n_i - 1)s_i^2$ $\quad\quad df_E = \sum n_i - t$

- **Decision Rule: Reject H_0 in favor of H_a if**
  - $F > F_\alpha\,(df_B, df_E)$

- **Applet**

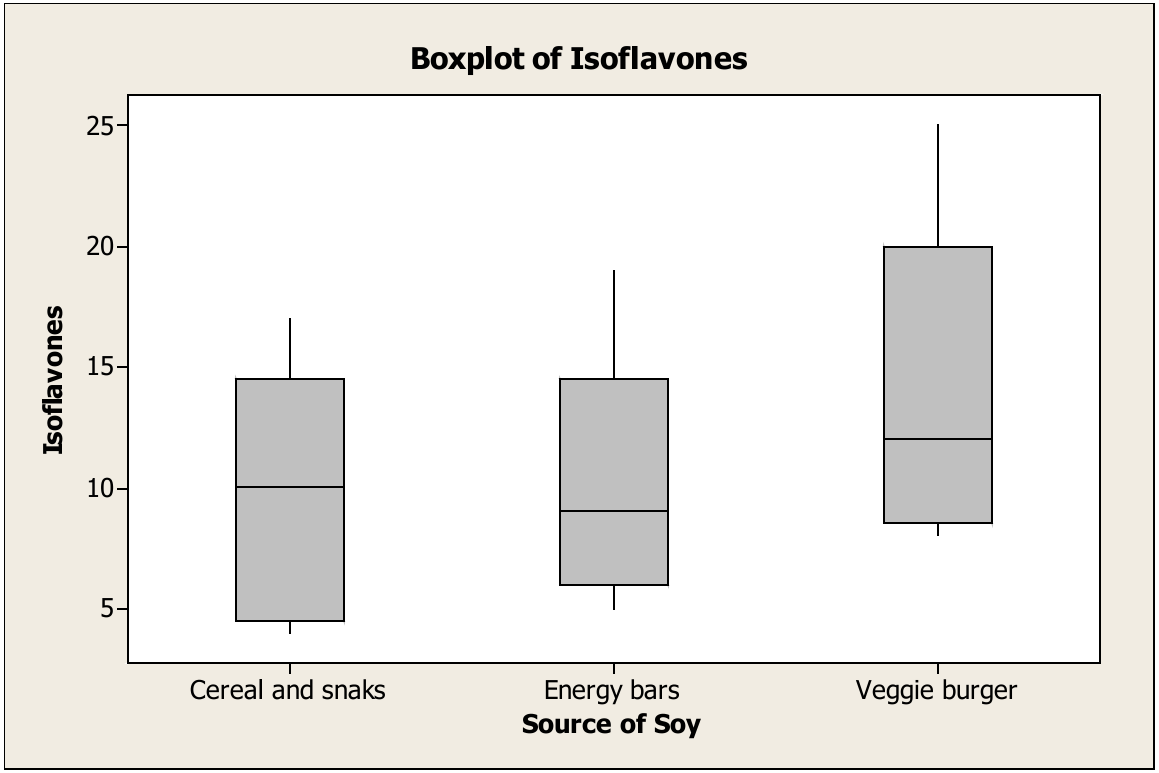
ANOVA TABLE

Source of Variation	df	Sum of Squares	Mean Square	F	p-value
Group (Between)	$t-1$	$\sum n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 = SS_B$	$\dfrac{SS_B}{df_B} = MS_B$	$\dfrac{MS_B}{MS_E} = F_{\text{calc}}$	$\Pr(F > F_{\text{calc}})$
Error (Within)	$N-t$	$\sum (n_i-1)s_i^{\,2} = SS_E$	$\dfrac{SS_E}{df_E} = MS_E$		
Total	$N-1$	$\sum (y_{ij}-\bar{y}_{\bullet\bullet})^2 = SS_T$			

- **Here:**
  - $N = \sum_i^t n_i$
  - $SS_T = SS_B + SS_E$
  - MS_E **is the pooled sample variance, an estimate for** σ^2
  - $R^2 = \frac{SS_B}{SS_T}$ **is the proportion of the total variation explained by the groups.**

BOOK EXAMPLE 8.1

- A hypothesis is that a nutrient “Isoflavones” varies among three types of food items: (1) Cereals and snacks, (2) energy bars, and (3) veggie burgers. A sample of five each is taken and the amount of isoflavones is measured.



EXAMPLE 8.1 CONT'D

- **Cereal and snacks:** $n_1 = 5$, $\bar{y}_1 = 9.20$, $s_1^2 = 33.7$
- **Energy bars:** $n_2 = 5$, $\bar{y}_2 = 10.00$, $s_2^2 = 29.0$
- **Veggie burger:** $n_3 = 5$, $\bar{y}_3 = 13.80$, $s_3^2 = 46.7$

- **Is there a sufficient evidence to conclude that the amount of isoflavones varies among these food items?** $\alpha = 0.05$.

- $H_0: \mu_1 = \mu_2 = \mu_3$
- $H_a: \mu_i \neq \mu_j$ **for some pairs** (i, j)

- **Assumptions:**
  - $\sigma_1 = \sigma_2 = \sigma_3$
  - **Data is generated from normal distribution for each type of food.**

EXAMPLE 8.1 ASSUMPTIONS

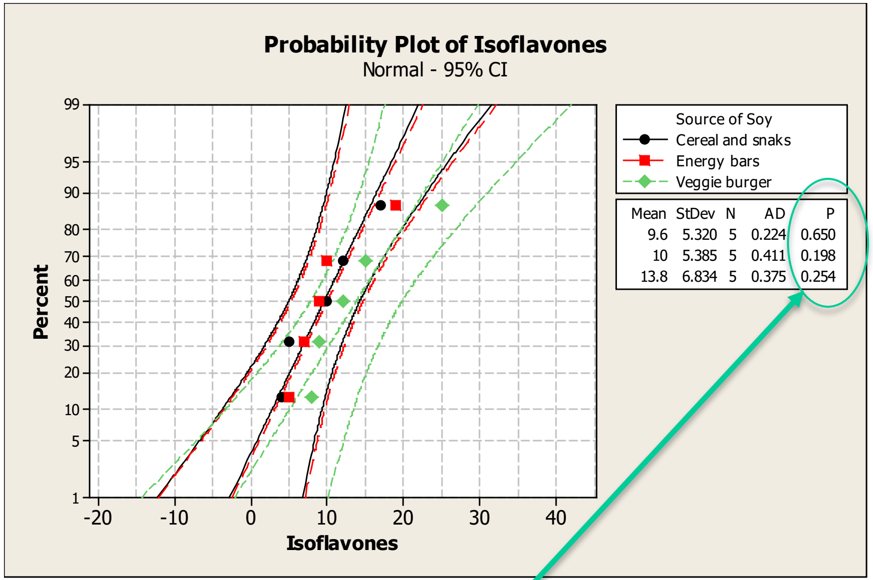
- $H_0: \sigma_1 = \sigma_2 = \sigma_3$
  - **In R:** car::leveneTest(exmp8.1$isof, exmp8.1$source)
  - lawstat::levene.test(exmp8.1$isof, exmp8.1$source)

Test for Equal Variances: Isoflavones versus Source of Soy

```
95% Bonferroni confidence intervals for standard deviations

   Source of Soy  N    Lower    StDev    Upper
Cereal and snaks  5  2.87498  5.31977  20.4752
     Energy bars  5  2.91032  5.38516  20.7269
   Veggie burger  5  3.69318  6.83374  26.3023


Bartlett's Test (Normal Distribution)
Test statistic = 0.30, p-value = 0.861


Levene's Test (Any Continuous Distribution)
Test statistic = 0.11, p-value = 0.896
```

- **Levene's test $p - value$ is 0.896. Fail to reject equality of the variances**

- H_0: **Data is generated from normal distribution for each type of food.**



- **Large $p - values$. Fail to reject Normality assumption.**

EXAMPLE 8.1 CONT'D

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_a: $\mu_i \neq \mu_j$ **for some pairs** (i, j)

- **TS** $F = \dfrac{SS_B/df_B}{SS_E/df_E}$
  - $SS_B = \sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ $\qquad df_B = t - 1$
  - $SS_E = \sum(n_i - 1)s_i^2$ $\qquad df_E = \sum n_i - t$

- $\bar{y}_{..} = \dfrac{n_1\bar{y}_{1.} + n_2\bar{y}_{2.} + n_3\bar{y}_{3.}}{\mathrm{n}_1 + \mathrm{n}_2 + \mathrm{n}_3} = \dfrac{5*9.2+5*10.0+5*13.8}{5+5+5} = 11.0$
- $df_B = (t - 1) = 3 - 1 = 2$
- $df_E = \sum n_i - t = (5 + 5 + 5) - 3 = 12$

EXAMPLE 8.1 CONT'D

- $SS_B = \sum n_i(\bar{y}_{i.} - \bar{y}_{..})^2$
  $= 5 * (9.2 - 11.0)^2 + 5 * (10.0 - 11.0)^2 + 5 * (13.8 - 11.0)^2$
  $= 60.40$
- $SS_E = \sum (n_i - 1)s_i^2$
  $= (5 - 1) * 33.0 + (5 - 1) * 29.0 + (5 - 1) * 46.7$
  $= 437.60$
- **TS** $F = \frac{SS_B/df_B}{SS_E/df_E} = \frac{60.40/2}{437.60/12} = 0.83$
- $F_\alpha(df_1 = 2, df_2 = 12) = 3.89$
- **Conclusion: Is $F > 3.89$? No. Fail to reject H_0. We cannot conclude that the amount of isoflavones vary among the food items.**
- **F Calculator**

EXAMPLE 8.1 ANOVA TABLE

Source of Variation	df	Sum of Squares	Mean Square	F	p-value
Group (Between)	$t-1$	$\sum n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = SS_B$	$\frac{SS_B}{df_B} = MS_B$	$\frac{MS_B}{MS_E} = F_{\text{calc}}$	$\Pr(F > F_{\text{calc}})$
Error (Within)	$N-t$	$\sum (n_i - 1)s_i^2 = SS_E$	$\frac{SS_E}{df_E} = MS_E$		
Total	$N-1$	$\sum (y_{ij} - \bar{y}_{\bullet\bullet})^2 = SS_T$			

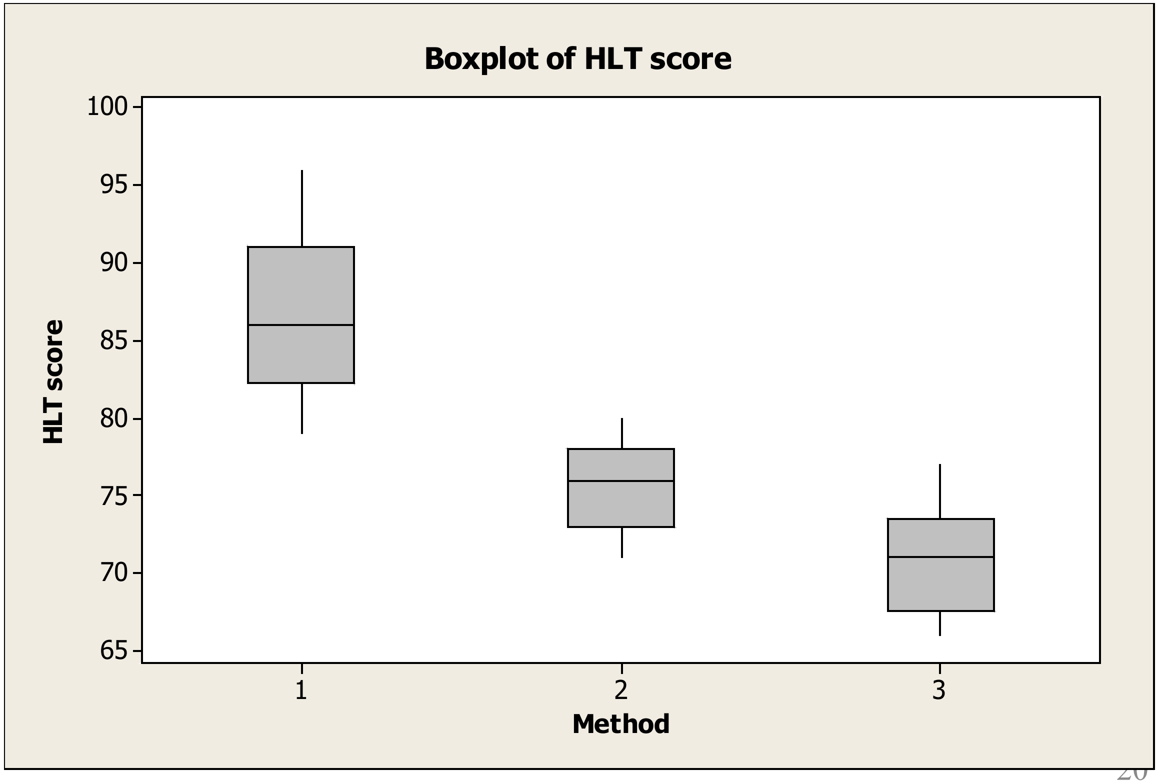
- **In R**
  - summary(aov(isof ~ source, data=exmp8.1))

One-way ANOVA: Isoflavones versus Source of Soy

```
Source          DF     SS    MS     F      P
Source of Soy    2   60.4  30.2  0.83  0.460
Error           12  437.6  36.5
Total           14  498.0

S = 6.039   R-Sq = 12.13%   R-Sq(adj) = 0.00%
```

BOOK EXAMPLE 8.2

- A clinical psychologist compares **three** methods for reducing the hospitality levels in university students. **HLT score** is used to measure the degree of hospitality. Randomly they assigned 8 students to method 1, 7 students to method 2 and 9 students to method 3. Each student was given HLT test at the end of semester



EXAMPLE 8.2 CONT'D

- **Is there a sufficient evidence to conclude difference among mean scores? Use** $\alpha = 0.05$.

- $H_0: \mu_1 = \mu_2 = \mu_3$
- $H_a: \mu_i \neq \mu_j$ **for some pairs** (i, j)

- **Assumptions:**
  - $\sigma_1 = \sigma_2 = \sigma_3$
  - **Data is generated from normal distribution for each type of food.**

EXAMPLE 8.2 ASSUMPTIONS

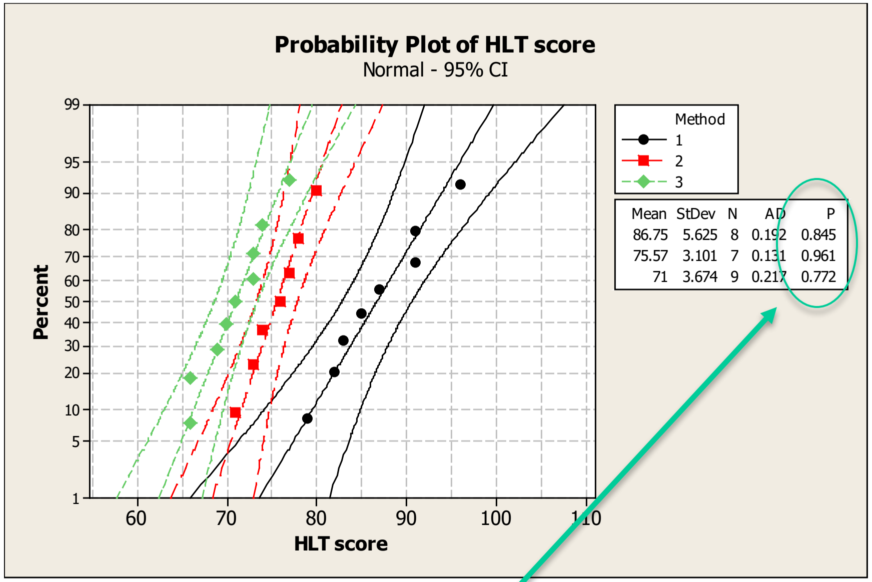
- H_0: $\sigma_1 = \sigma_2 = \sigma_3$
  - **In R:** car::leveneTest(exmp8.2$HLT, exmp8.2$method)
  - lawstat::levene.test(exmp8.2$HLT, exmp8.2$method)

Test for Equal Variances: HLT score versus Method

```
95% Bonferroni confidence intervals for standard deviations

Method  N    Lower     StDev    Upper
     1  8  3.41853  5.62520  13.7768
     2  7  1.82797  3.10146   8.4154
     3  9  2.29049  3.67423   8.3206


Bartlett's Test (Normal Distribution)
Test statistic = 2.46, p-value = 0.292


Levene's Test (Any Continuous Distribution)
Test statistic = 1.68, p-value = 0.210
```

- **Levene's test** $p - value$ **is** 0.210**. Fail to reject equality of the variances**

- H_0: **Data is generated from normal distribution for each type of food.**



- **Large** $p - values$**. Fail to reject Normality assumption.**

EXAMPLE 8.2 CONT'D

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_a: $\mu_i \neq \mu_j$ **for some pairs** (i, j)

One-way ANOVA: HLT score versus Method

Source	DF	SS	MS	F	P
Method	2	1090.6	545.3	29.57	0.000
Error	21	387.2	18.4		
Total	23	1477.8			

S = 4.294   R-Sq = 73.80%   R-Sq(adj) = 71.30%

- **In R:** summary(aov(HLT ~ method, data=exmp8.2))

- **TS** $F = \frac{SS_B/df_B}{SS_E/df_E} = \frac{1090.6/2}{387.2/21} = 29.57$
- $F_\alpha(df_1 = 2, df_2 = 21) = 3.47$

- **Conclusion: Is** $F > 3.47$? **Yes. Reject** H_0**.**
- **F Calculator**

NORMALITY ASSUMPTION FAILS
NON-PARAMETRIC METHOD

MARQUETTE UNIVERSITY
Be The Difference.

- **In the ANOVA method, we assume that the sample from each treatment level is drawn from normal population. What if the distribution is non-normal.**

- **The Kruskal-Wallis Test**

- H_0: **All t distributions are identical**
- H_a: **Not all distributions are the same.**

- **TS:**
  1. **Rank all samples from the lowest to the highest.**
  2. **The test statistics H is similar to the F-statistic based on the ranks.**

- **Decision Rule: Reject H_0 if $H > \chi^2_\alpha(df = t - 1)$**
- **In R:**
  - kruskal.test(x, g)

WHAT IF EQUALITY OF THE VARIANCES FAIL?

- **The assumption that the sample are generated from normal distribution is not very important as long as the total sample size is large.**

- **Note that conceptually the test statistic** $F = \frac{SS_B/df_B}{SS_E/df_E}$ **still makes sense.**

- **The major problem is with the assumption** $\sigma_1 = \sigma_2 = \cdots = \sigma_t$**. If this cannot be assumed, F- test must not be used.**

- **If** $H_0\colon\ \sigma_1 = \sigma_2 = \cdots = \sigma_t$ **is rejected, then one approach is to transform the data if the variances** σ^2 **is a function of the mean** μ**.**