Analysis of Variance (Factorial Experiments)

Also referred to as ANOVA[1]

November 20, 2017

[1] HMS, 2017, v1.2

Chapter References

- Diez: 5.5
- Navidi, Chapter 9

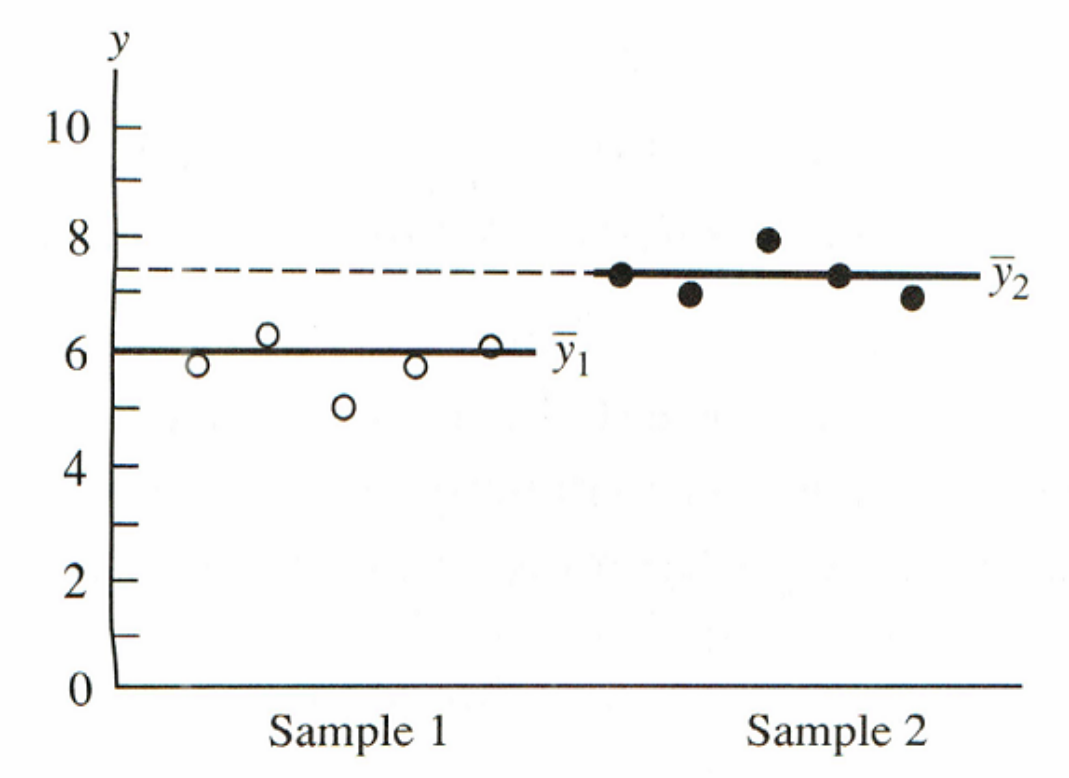
Comparing two means

To compare whether the means from two samples are the same we use a two-sampled t-Test.

What do we do if we have multiple samples?

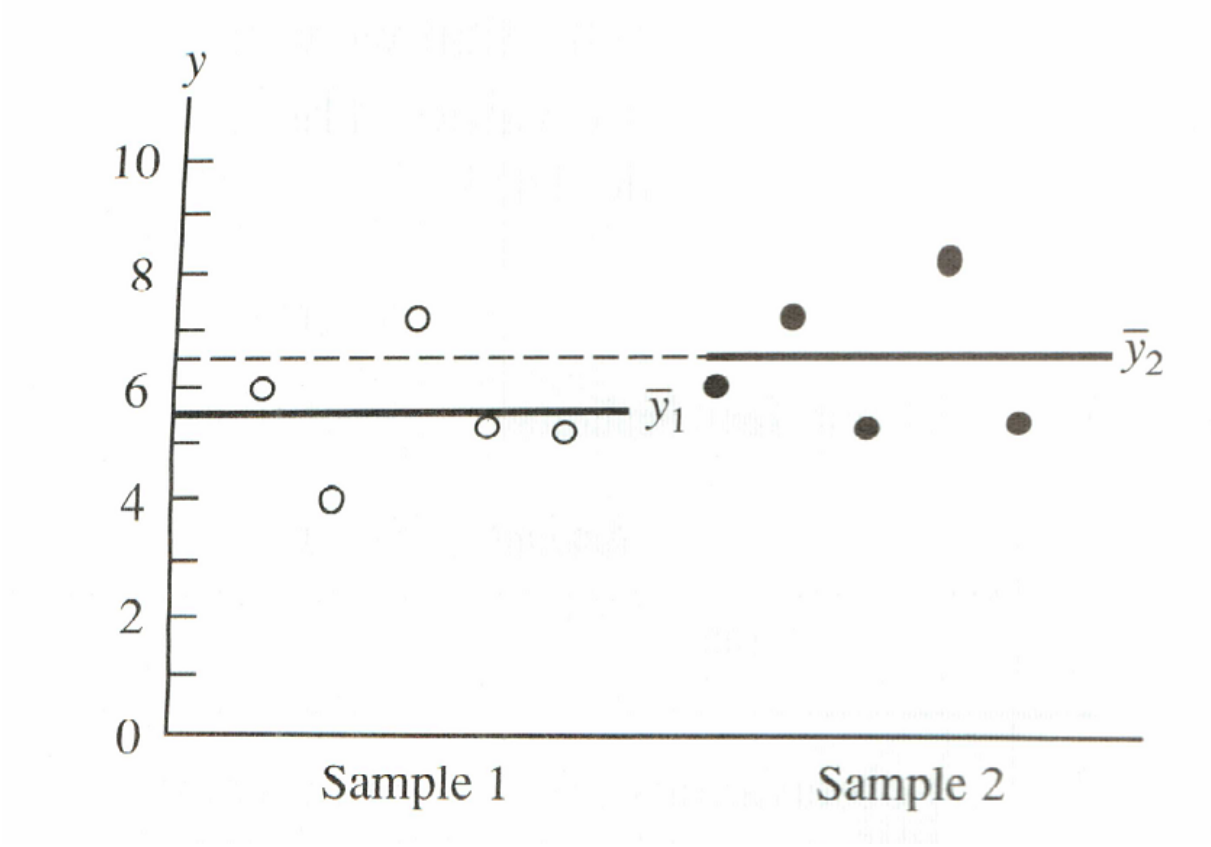
Comparing two means

Bold line represents the mean



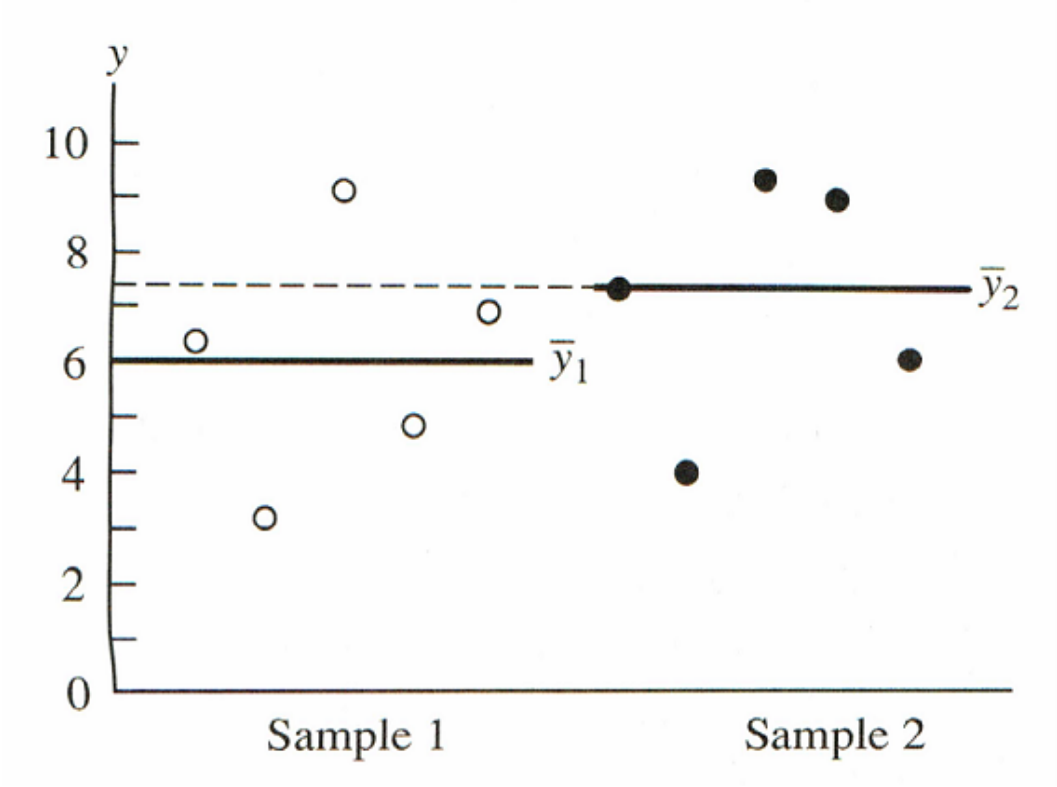
Comparing two means

Bold line represents the mean



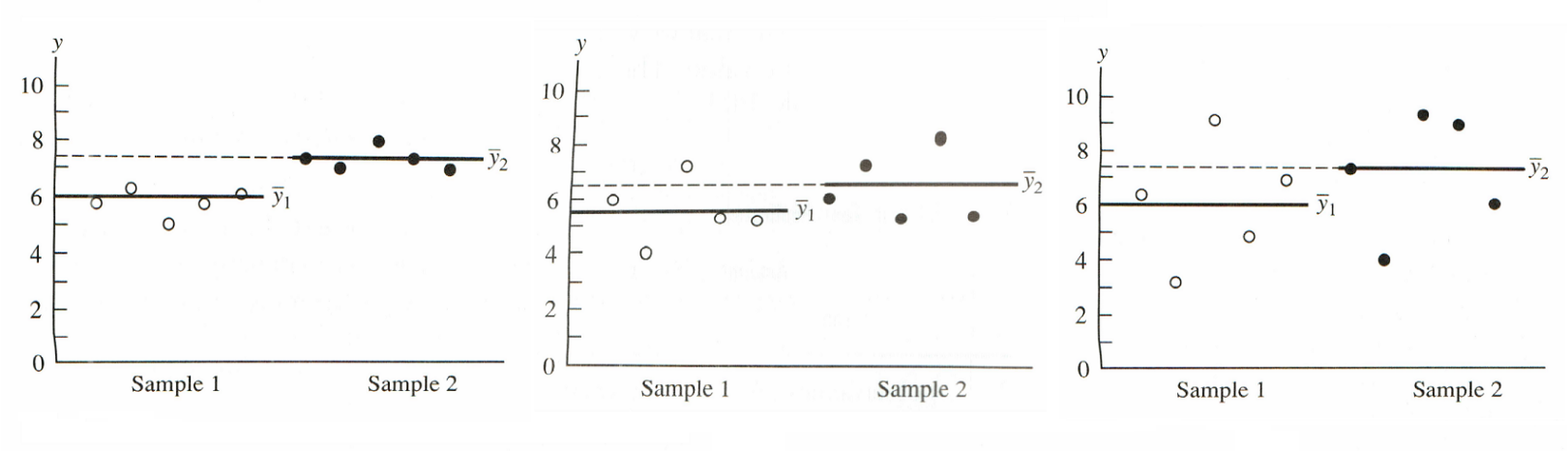
Comparing two means

Bold line represents the mean



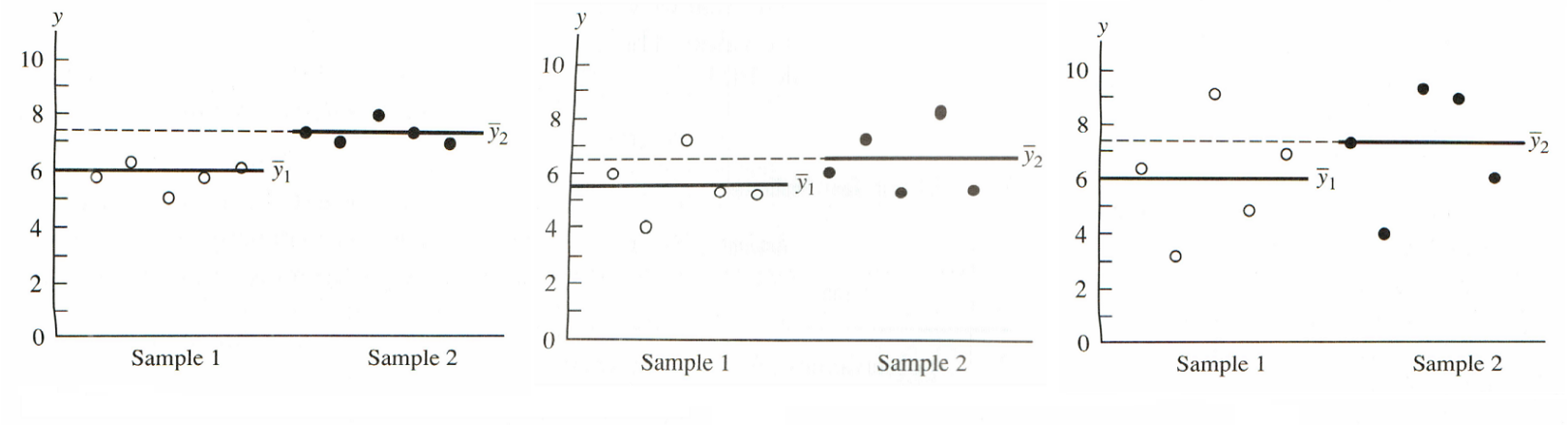
Comparing two means

What are we doing when you assess whether the two samples have the same mean or not?



Comparing two means

What are you doing when you assess whether the two samples have the same mean or not?



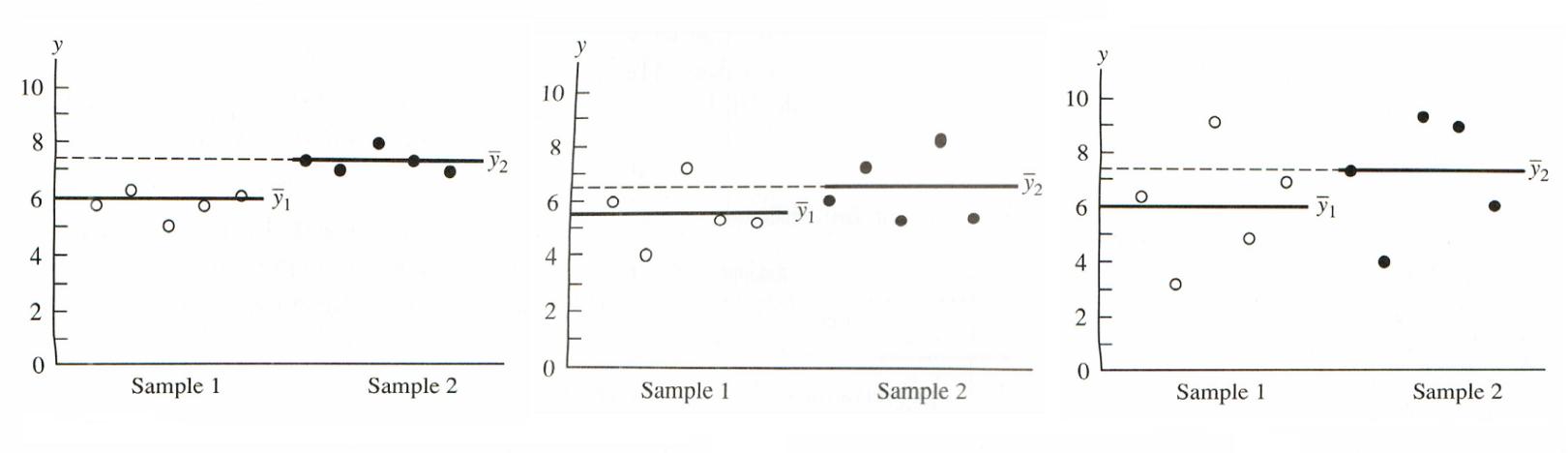
We visually compare the distance (variation) **between** the means to the variation **within** the means.

e.g in the third panel, the variation within the samples is greater than the distance between the two means.

Therefore, the larger the within variation is to the between variation the more likely we'll consider the means the same (or no evidence to suggest otherwise).

Comparing two means

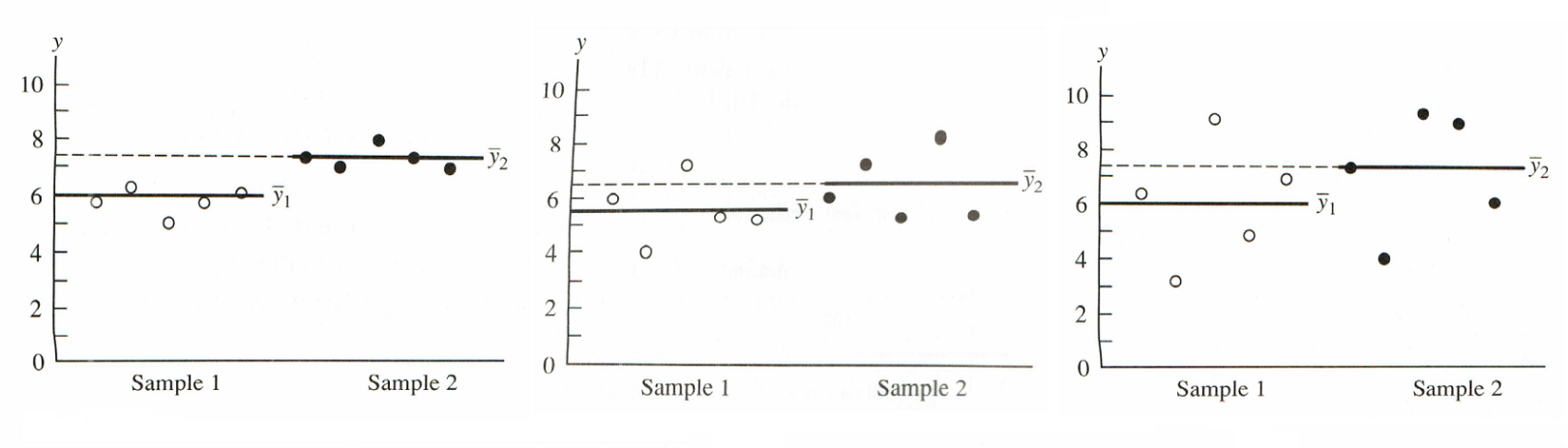
To contrast: In the first panel, the variation within the means is less that the distance between the means, therefore this suggests that the means are different.



This is what the **Analysis of Variance** does.

Comparing two means

The question is then: How do we measure within and between variation?



Comparing two means

The question is then: How do we measure within and between variation?

We'll measure variation using the sum of squares:

$$\text{Sum of Squares} = \sum (x_i - \overline{x})^2$$

Comparing two means

The question is then: How do we measure within and between variation?

$$\text{Total Sums of Squares} = \text{Variation Between Means} + \text{Variation Within Means}$$

Later on we will abbreviate the phrase 'Sum of Squares' with SS.

Comparing two means

Let t be the number of means (also more commonly called the treatments or groups)

Let n_i be the number of data points in group i.

Let N be the total number of data points in the entire set $N = \sum n_i$

Group 1: $x_{11}, x_{12}, \ldots, x_{1n_1}$

Group 2: $x_{21}, x_{22}, \ldots, x_{2n_2}$

$\ldots$

Group r: $x_{t1}, x_{t2}, \ldots x_{tn_t}$

$$\text{Total } SS_T = \sum_{i=1}^{t} \sum_{k=1}^{n_i} (x_{ik} - \overline{x})^2$$

where $\overline{x}$ is the grand mean (i.e mean of all data points)

Comparing two means

Let's look at this more closely

The deviation of any particular data point x_{ij} relative to the grand mean $\overline{x}$ is:

$$x_{ij} - \overline{x}$$

If we square this and sum over all data points we get what we saw in the previous slide, the total sums of squares:

$$\text{Total } SS_T = \sum_{i=1}^{t} \sum_{k=1}^{n_i} (x_{ik} - \overline{x})^2$$

Comparing two means

Let us decompose $x_{ij} - \overline{x}$ as follows:

$$(x_{ij} - \overline{x}) = (x_{ij} - \overline{x}_i) + (\overline{x}_i - \overline{x})$$

where $\overline{x}_i$ is the mean for treatment i.

$(x_{ij} - \overline{x}_i)$ is the deviation of the data point from its treatment mean – within variation.

$(\overline{x}_i - \overline{x})$ is the deviation of the grand mean from the mean of a treatment – between variation

Comparing two means

$$(x_{ij} - \overline{x}) = (x_{ij} - \overline{x}_i) + (\overline{x}_i - \overline{x})$$

Square both sides and sum over every data point:

$$\begin{aligned}\sum_i^t \sum_j^{n_i} (x_{ij} - \overline{x})^2 &= \sum_i^t \sum_j^{n_i} [(x_{ij} - \overline{x}_i) + (\overline{x}_i - \overline{x})]^2 \\ &= \sum_i^t \sum_j^{n_i} (x_{ij} - \overline{x}_i)^2 + \sum_i^t \sum_j^{n_i} (\overline{x}_i - \overline{x})^2 \\ &\quad + 2 \sum_i^t \sum_j^{n_i} (x_{ij} - \overline{x}_i)(\overline{x}_i - \overline{x})\end{aligned}$$

Theorem:

$$\sum_i \sum_j x_i y_{ij} = \sum_i \left(x_i \sum y_{ij} \right)$$

Comparing two means

Theorem:

$$\sum_i \sum_j x_i y_{ij} = \sum_i \left(x_i \sum y_{ij} \right)$$

Using this theorem we rearrange the last term such that:

$$\begin{aligned} & 2 \sum_i^t \sum_j^{n_i} (x_{ij} - \overline{x}_i)(\overline{x}_i - \overline{x}) \\ & = 2 \sum_i (\overline{x}_i - \overline{x}) \sum_j (x_{ij} - \overline{x}_i) \end{aligned}$$

Comparing two means

Given:

$$= 2 \sum_i (\overline{x}_i - \overline{x}) \sum_j (x_{ij} - \overline{x}_i)$$

Look at this term:

$$\sum_j (x_{ij} - \overline{x}_i)$$

j is the jth data element in the i group. For a given i, it sums up the deviations between a data element and the mean for that group.

$$\sum (x_i - \overline{x}) =$$

$$= x_1 - \overline{x} + x_2 - \overline{x} + \cdots + x_n - \overline{x}$$

$$= x_1 - \sum \frac{x_i}{n} + x_2 - \sum \frac{x_i}{n} + \cdots + x_n - \sum \frac{x_i}{n}$$

$$= (x_1 + x_2 + \cdots + x_n) - n \left(\sum \frac{x_i}{n} \right)$$

$$= \sum x_i - \sum x_i = 0$$

Comparing two means

What this means is that

$$2\sum_i(\overline{x}_i-\overline{x})\sum_j(x_{ij}-\overline{x}_i)=0$$

However, recall that:

$$\sum_i^t\sum_j^{n_i}(x_{ij}-\overline{x})^2=$$

$$=\sum_i^t\sum_j^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_i^t\sum_j^{n_i}(\overline{x}_i-\overline{x})^2$$

$$+2\sum_i^t\sum_j^{n_i}(x_{ij}-\overline{x}_i)(\overline{x}_i-\overline{x})$$

Therefore:

$$\sum_i^t\sum_j^{n_i}(x_{ij}-\overline{x})^2=\sum_i^t\sum_j^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_i^t\sum_j^{n_i}(\overline{x}_i-\overline{x})^2$$

Comparing two means

$$\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_{i}^{t}\sum_{j}^{n_i}(\overline{x}_i-\overline{x})^2$$

Note there is no j in the brackets of the third term, i.e we're just adding up the bracket term n_i times:

$$\begin{aligned}\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x})^2&=\sum_{i}^{t}\sum_{j}^{n_i}(\overline{x}_i-\overline{x})^2\\&=\sum_{i}^{t}n_i(\overline{x}_i-\overline{x})^2\end{aligned}$$

Comparing two means

In summary:

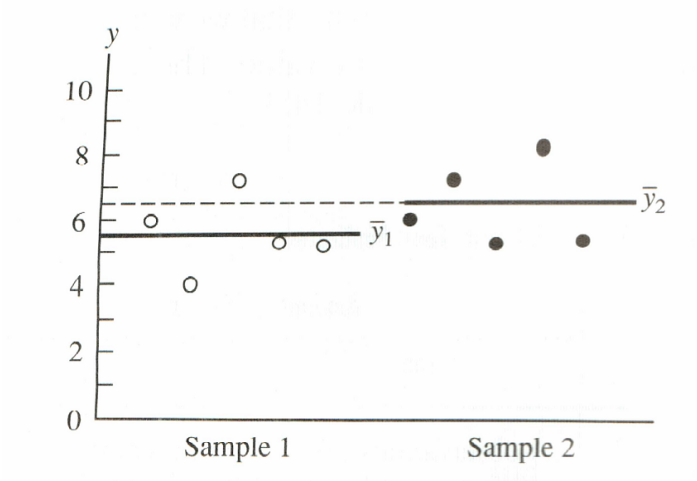
$$\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_{i}^{t}n_i(\overline{x}_i-\overline{x})^2$$

The term on the left will be recognized to be the total sum of squares:

$$SS_T=\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x}_i)^2+\sum_{i}^{t}n_i(\overline{x}_i-\overline{x})^2$$

Comparing two means

Variation within is presumably due to random errors in collecting the sample.



We could compute this random variation by computing the sum of squares between sample points and their corresponding sample means. This variation is often called the sum of squared errors, or SS_E:

$$SS_E = \sum_{i=1}^{t} \sum_{k=1}^{n_i} (x_{ik} - \overline{x}_i)^2$$

where $\overline{x}_i$ is the mean for group i.

Comparing two means

$$SS_E = \sum_{i=1}^{t} \sum_{k=1}^{n_i} (x_{ik} - \overline{x}_i)^2$$

where $\overline{x}_i$ is the mean for group i.

Compare with formal derivation, we see that the second term is SS_E.

$$\sum_{i}^{t} \sum_{j}^{n_i} (x_{ij} - \overline{x})^2 = \sum_{i}^{t} \sum_{j}^{n_i} (x_{ij} - \overline{x}_i)^2 + \sum_{i}^{t} n_i (\overline{x}_i - \overline{x})^2$$

Comparing two means

What about variation between groups/treatments?

For this we look at the variation between the grand mean and the individual group means. i.e how do the group means vary within the grand mean?

This variation is called the treatment sum of squares, SS_{T_R} and from the formal derivation we see that:

$$SS_{T_R} = \sum_{i=1}^{t} n_i(\overline{x}_i - \overline{x})^2$$

where $\overline{x}_i$ is the mean for group i.

Summary

This is the important result:

$$\sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x})^2 = \sum_{i}^{t}\sum_{j}^{n_i}(x_{ij}-\overline{x}_i)^2 + \sum_{i}^{t} n_i(\overline{x}_i-\overline{x})^2$$

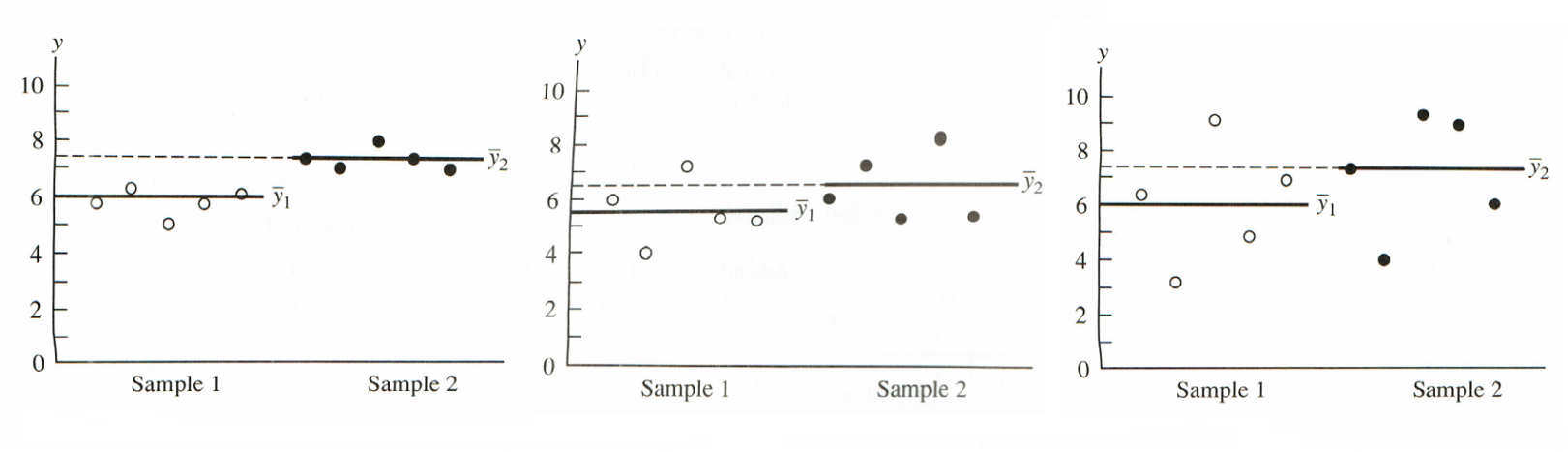
$$SS_T = SS_E + SS_{T_r}$$

$$\text{Total} = \text{Within} + \text{Between}$$

But what do we do with it?

What we'd like to do is compare SS_{T_r} with SS_E. The larger is SS_{T_r} the more likely the means are different.

Summary



Large: SS_{T_r}/SS_E Medium: SS_{T_r}/SS_E Small: SS_{T_r}/SS_E

F Test

We could look at the ratio:

$$\frac{SS_{T_r}}{SS_E}$$

Better still we can divide each term by the appropriate degrees of freedom so that each term now become a variance. We can test equality of the variance using an F test.

F Test

Better still we can divide each term by the appropriate degrees of freedom so that each term now become a variance. We can test equality of the variances using a F test.

$$F = \frac{SS_{T_r}/(t-1)}{SS_E/(N-t)} = \frac{MST}{MSE}$$

where MST is the mean square between treatments and MSE is the mean square error (within treatments).

Procedure

Given a table of data such as:

Amess et al (1978) conducted a study where they looked at the effects of three different ventilation treatments on patients with cardiac bypass surgery. After 24 hours, the folic acid level was measured in red blood cells (proxy for Vit B12 metabolism).

Group	Patients								
1	243	251	275	291	347	354	380	392	
2	206	210	226	249	255	273	285	295	309
3	241	258	270	293	328				

Null Hypothesis H_o: The mean levels of folic acid are the same in each group $\mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis H_1: The level of folic acid after 24hrs differ $\mu_1 \neq \mu_2 \neq \mu_3$

Assumptions Before you Being

- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other
- Normality of errors. It is assumed that error terms are normally distributed.
- Equal error variance across treatments. i.e all error terms have the same variance: homogeneity of variances.

Procedure

- Assumptions hold?
- Set up H_o and H_1.
- Set the significance level, eg 5%
- Compute $SS_{T_r} = \sum_{i=1}^{t} n_i(\overline{x}_i - \overline{x})^2$
- Compute $SS_E = \sum_{i}^{t} \sum_{j}^{n_i} (x_{ij} - \overline{x}_i)^2$
- Compute the MST_r and MSE
- Compute F statistic: $F = MST_r/MSE$
- Find p-value in a F table with $t-1$ and $N-t$ degrees of freedom.

Walk Through Example

We'll assume for now that the required assumptions hold. Here is the data:

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

H_o: Three means across the group are the same: $\mu_1 = \mu_2 = \mu_2$

Set the significance level to 5%.

Walk Through Example

Its standard practice to form a table with the following structure:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments				
Error				
Total				

The Error is sometimes also called the 'Residuals' or 'Within'.

Walk Through Example

Sample printout from a statistics package:

One-way ANOVA: Observations versus Cotton Weight %

Source	DF	SS	MS	F	P
Cotton Weight %	4	475.76	118.94	14.76	0.000
Error	20	161.20	8.06		
Total	24	636.96			

Walk Through Example

Enter the degrees of freedom first $(t-1)$ and $(N-t)$

$t = 3$ and $N = 12$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments		2		
Error		9		
Total		11		

Walk Through Example

Compute some basic statistics:

Individual group Means:

$$\overline{x}_1 = (7+3+6+6)/4 = 5.5 \quad \overline{x}_2 = 6.0 \quad \overline{x}_3 = 6.0$$

$$\text{Grand mean} = \frac{1+7+3+\ldots+7+6+7}{12} = 5.8333 = \frac{5+6+6}{3}$$

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

Walk Through Example

Compute some basic statistics:

Individual group variances

$$s_1^2 = \frac{(5.5-7)^2 + (5.5-3)^2 + (5.5-6)^2 + (5.5-6)^2}{4-1} = 3.0$$

$$s_2^2 = 2.0$$

$$s_3^2 = 2.0$$

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

Walk Through Example

Next calculate the SS_E:

$$SS_E = \sum_{i}^{t} \sum_{j}^{n_i} (x_{ij} - \overline{x}_i)^2$$

$$SS_E = (7 - 5.5)^2 + (3 - 5.5)^2 + \ldots + (6 - 6)^2 + (7 - 6)^2 = 21$$

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

Walk Through Example

Enter SS_E into the table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments		2		
Error	21	9		
Total		11		

Walk Through Example

Next calculate the SS_{T_r}:

$$SS_{T_r} = \sum_{i=1}^{t} n_i(\overline{x}_i - \overline{x})^2$$

$n_i = 4$

$$SS_{T_r} = 4 \times [(5.5 - 5.83)^2 + (6 - 5.83)^2 + (6 - 5.83)^2] = 0.6667$$

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

Walk Through Example

Enter SS_{T_r} into the table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments	0.6667	2		
Error	21	9		
Total	21.6667	11		

Walk Through Example

Next calculate the mean squares:

$$MST = 0.6667/(3-1) = 0.3333$$

$$MSE = 21/(12-3) = 2.3333$$

Group				
1	7	3	6	6
2	6	5	5	8
3	4	7	6	7

Walk Through Example

Enter SS_{T_r} into the table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments	0.6667	2	0.3333	
Error	21	9	2.3333	
Total	21.6667	11		

Walk Through Example

Next calculate the F statistic:

$$F = 0.3333/2.3333 = 0.14286$$

And enter it into the table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments	0.6667	2	0.3333	0.14286
Error	21	9	2.3333	
Total	21.6667	11		

Walk Through Example

At 5% significance look up the corresponding F value in the F table.
Numerator df $= 3 - 1 = 2$
Denominator df $= 12 - 3 = 9$

F value with these degrees of freedom and using the 0.025 one tailed table, the value is 5.7147

This is much greater than 0.14286, therefore we accept H_o.

The actual p-value is 0.87 (Used online calculator to get this) $\gg 0.025$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Treatments	0.6667	2	0.3333	0.14286
Error	21	9	2.3333	
Total	21.6667	11		

Walk Through Example

Let's make a change:

Group				
1	17	13	16	16
2	6	5	5	8
3	4	7	6	7

Set the significance level to 5%.

Class Exercise: Determine whether we accept or reject H_o.

Assumptions

Checking the Assumptions

Normality check: Carry out a probability plot of the residuals $x_{ij} - \overline{x}_i$

If the data is normally distributed the residuals should also be normally distributed with a mean of zero and the probability plot will yield a straight line.

Equal variances: Plot the residuals for each group versus the sample means, $\overline{x}_i$. If the spreads look roughly the same then assume the variances are also the same.

Two Factor ANOVA

Use a two factor ANOVA if you have more than one type of factor to consider. Table shows weight loss due to combination of diet and exercise.

	Diet Plan		
Exercise Plan	1	2	3
A	10.2	11.1	9.5
B	8.5	10.4	9.4
C	7.1	12.6	8.4

Why are they important?

Two Factor ANOVA

Another Example:

Plecebo	Medication 1	Medication 2	Medication 3
A			
B			
C			

Gender	Medication 1	Medication 2	Medication 3
Male			
Female			

Two Factor ANOVA

Another Example:

Heart Value	Age (41-50)	Age (51-60)	Age (61-70)	Age (70-)
Type 1				
Type 2				
Type 3				

Two Factor ANOVA

Whereas with a one-factor ANOVA:

$$SS_T = SS_E + SS_{T_r}$$

With a two factor ANOVA we have to accommodate more:

$$SS_T = SS_E + SS_A + SS_B + SS_{AB}$$

SS_{AB} is called the interaction term.

What are interactions?

Two Factor ANOVA

Its the possibility of interactions that makes two-way ANOVA interesting. If there are no interactions, a two-way analysis effectively reduces to two one-way analyses.

Drinking alcohol increases the chance of throat cancer, as does smoking. However, people who both drink and smoke have an even higher chance of getting throat cancer. The combination is an example of these risk factors interacting.

An interaction implies that the effect of one variable differs depending on the level of another variable.

The effect of smoking on the probability of getting throat cancer is greater for people who drink than for people who do not drink: the effect of smoking differs depending on whether drinkers or nondrinkers are being considered.

Two Factor ANOVA

What can be tested? **Three hypotheses:**

Null Hypothesis #1: There are no differences in the means due to the first factor (exercise) irrespective of the second factor.

Null Hypothesis #2: There are no differences in the means due to the second factor (diet) irrespective of the first factor.

Null Hypothesis #3: There are no interaction effects between the first and second factors.

Two Factor ANOVA

Null Hypothesis #1: There are no differences in the means due to the first factor (exercise) irrespective of the second factor.

Could be tested with a one factor ANOVA.

Null Hypothesis #2: There are no differences in the means due to the second factor (diet) irrespective of the first factor.

Could be tested with a one factor ANOVA.)

Null Hypothesis #3: There are no interaction effects between the first and second factors.

Cannot be tested with a one factor ANOVA.

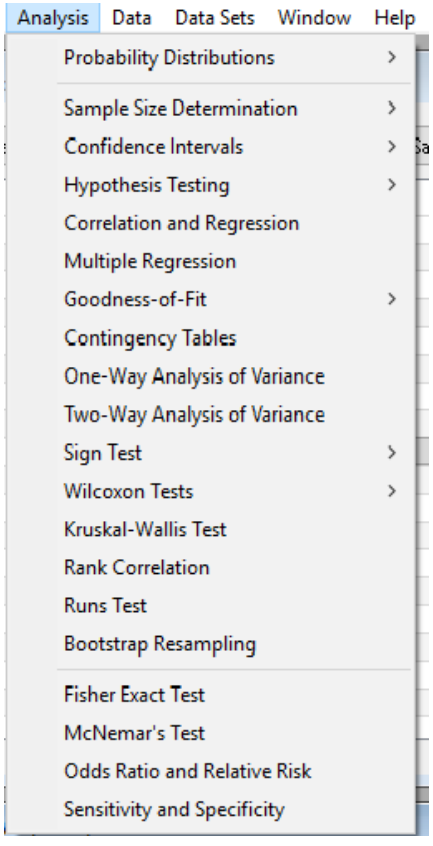
HardCore Stats Software

`http://r4stats.com/articles/popularity/`

SPSS	(UW licence) otherwise its $99 per month
R	(Free)
SAS	(UW licence) $1000 to $50,000
Stata	(UW licence) $1000 to $2000
GraphPad Prism	$390
MATLAB	$1000s
etc	

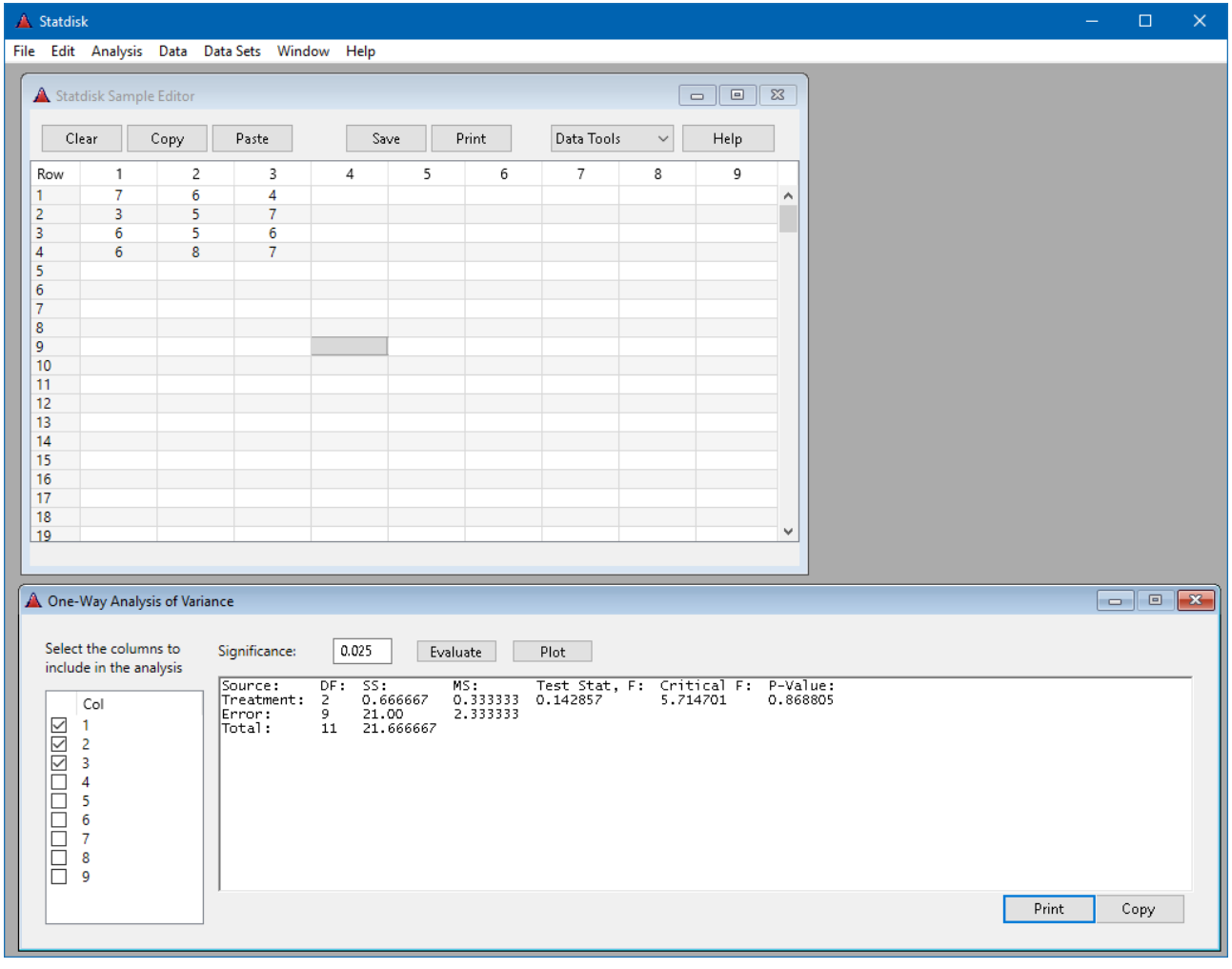
Software: StatDisk

StatDisk:
https://www.statdisk.org/



Software

StatDisk (Windows and Mac):
https://www.statdisk.org/



Software

With Python you can use a variety of libraries, including scipy, pandas, statsmodels

```
import scipy.stats
x1 = [7,3,6,6]; x2 = [6,5,5,8]; x3 = [4,7,6,7]
print scipy.stats.f_oneway (x1, x2, x3)
F_onewayResult(statistic=0.14285714285714288,
pvalue=0.868051120637752)
```

Software: Excel

Use Data Tab → Data Analysis Data Analysis → One Factor

	A	B	C	D	E	F	G	H	I	J	K	L
	7	6	4		Anova: Single Factor							
	3	5	7									
	6	5	6		SUMMARY							
	6	8	7		*Groups*	*Count*	*Sum*	*Average*	*Variance*			
					Column 1	4	22	5.5	3			
					Column 2	4	24	6	2			
					Column 3	4	24	6	2			
10					ANOVA							
11					*Source of Varia*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
12					Between	0.666667	2	0.333333	0.142857	0.868805	5.714705	
13					Within Gr	21	9	2.333333				
14												
15					Total	21.66667	11					
16												
17												