statstutor

community project

encouraging academics to share statistics support resources

Statistical Methods 10. Introduction to Analysis of Variance (ANOVA)

Based on materials provided by Coventry University and Loughborough University under a National HE STEM Programme Practice Transfer Adopters grant

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Workshop outline

- Motivation for ANOVA
- Checking assumptions
- ANOVA using SPSS
- Multiple comparisons – post hoc tests

Participants should have previous experience of:

- Descriptive Statistics – see Workshop 3
- SPSS – see Workshop 7
- Two sample tests – see Workshop 8

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 1

- ❑ Amount of oil used by four machines (litres/week)
- ❑ Recorded over 6 sampled periods
- ❑ Does this sample data provide evidence that oil consumption differs between the machines?

$\Rightarrow$ Create summary statistics and error bar charts

$\Rightarrow$ Describe the data

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Oil data

Machine number gives 4 data groups (known as a **factor**)

Note: This example has the same number of data values for each group, but this is not necessary (as in the unpaired t-test)

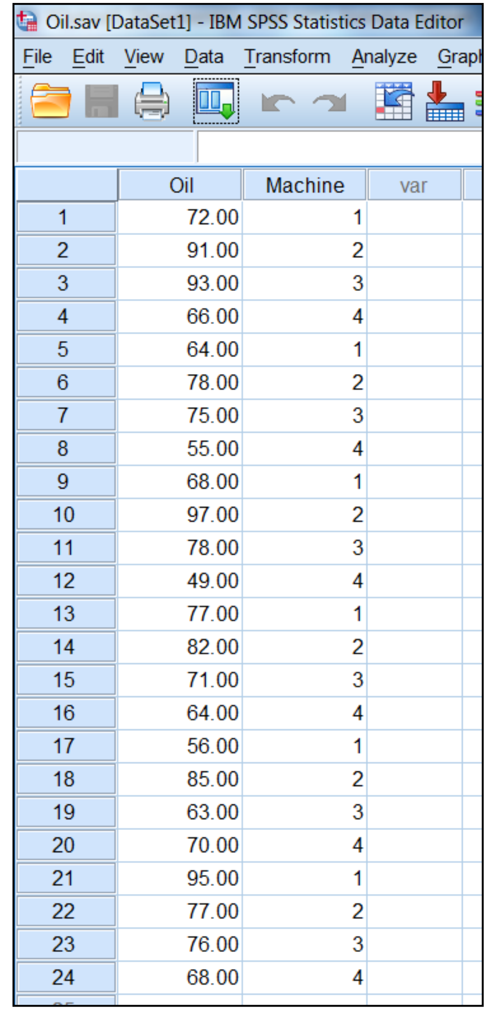
Machine	**1**	**2**	**3**	**4**
Oil consumption	72	91	93	66
	64	78	75	55
	68	97	78	49
	77	82	71	64
	56	85	63	70
	95	77	76	68

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Oil data in SPSS

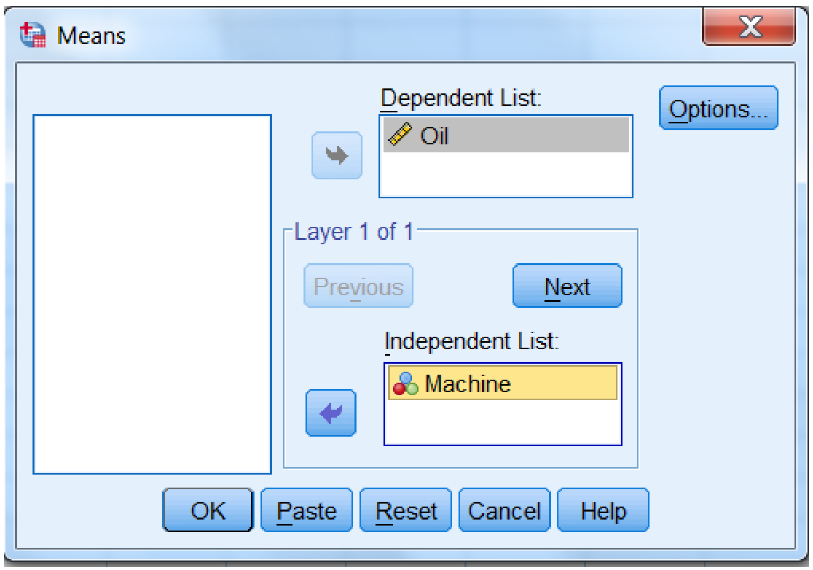
- Open the file Oil.sav
- Oil data is given in a single column with the *Machine* variable indicating the machine it refers to



	Oil	Machine	var
1	72.00	1	
2	91.00	2	
3	93.00	3	
4	66.00	4	
5	64.00	1	
6	78.00	2	
7	75.00	3	
8	55.00	4	
9	68.00	1	
10	97.00	2	
11	78.00	3	
12	49.00	4	
13	77.00	1	
14	82.00	2	
15	71.00	3	
16	64.00	4	
17	56.00	1	
18	85.00	2	
19	63.00	3	
20	70.00	4	
21	95.00	1	
22	77.00	2	
23	76.00	3	
24	68.00	4	

Simple statistics

- Analyze - Compare means – means
- Add *Oil* and *Machine* as shown



Report

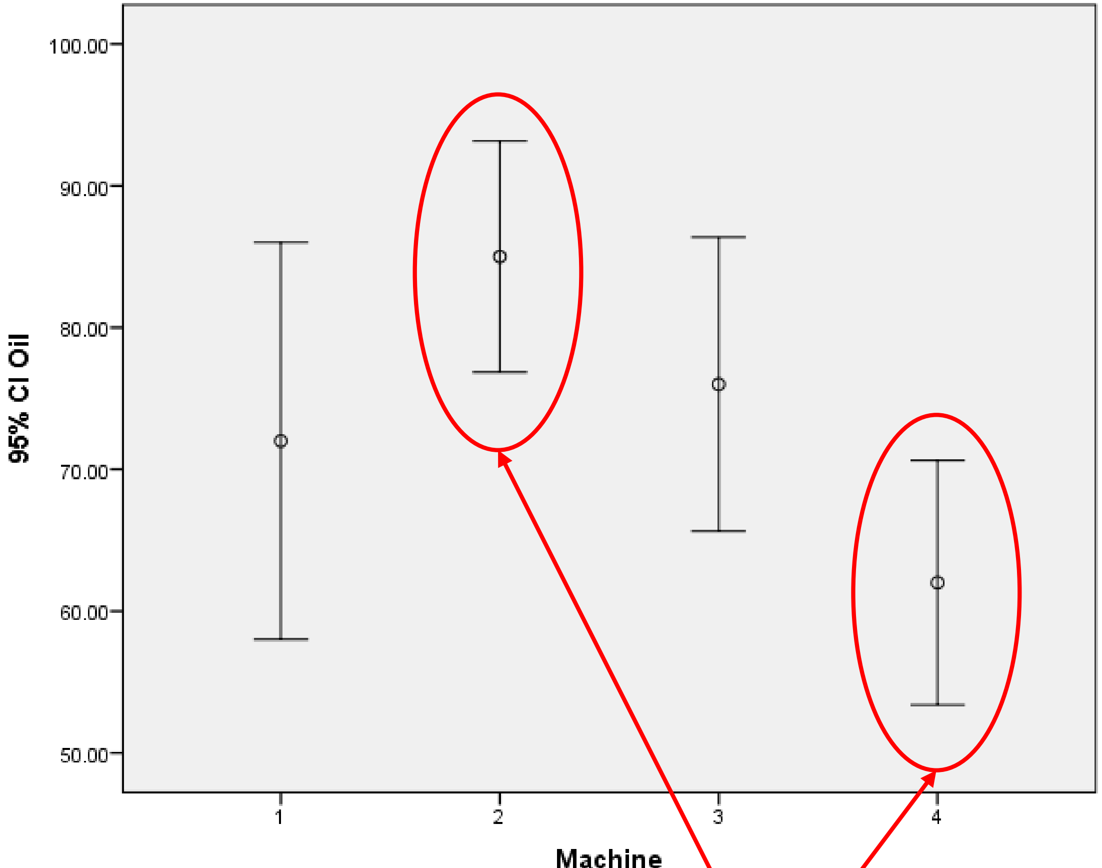
Oil

Machine	Mean	N	Std. Deviation
1	72.0000	6	13.34166
2	85.0000	6	7.77174
3	76.0000	6	9.87927
4	62.0000	6	8.22192
Total	73.7500	24	12.60521

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Error bar chart (Oil v. Machine)



Error bar charts are better for larger samples.

They show the means and their confidence intervals

Non-overlapping confidence intervals indicate possible significant differences

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Initial observations

- There appear to be differences between the sample means, i.e. variation between groups
- But there is also variation within groups
- Can we conclude that there are differences between groups (population means)?
- We need an objective approach – this is known as **ANOVA**

Introduction to ANOVA

- ANOVA is a multiple group extension of the two sample independent t test used to compare two groups (population means)
- ANOVA is used to compare several groups (population means)
- Called ANOVA from **AN**alysis **O**f **VA**riance
- (The name is therefore a bit confusing because it appears to be a **means** test, not a variance test)

Introduction to ANOVA

- Better than doing lots of two sample tests, e.g. 6 tests for 4 machines
- For every test, there is a probability that we reject H_0 when it is true
- This probability is 0.05 for testing at a significance level of 0.05
- Doing several tests increases the probability of making a wrong inference of significance (Type I error)
- E.g. for our example, the probability of a wrong inference, assuming they are all equally randomly distributed and that these events are independent is $1 - 0.95^6 = 1 - 0.735 = 0.265$, i.e. more than 1 in 4

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

The ANOVA model

$$y_{ij} = \mu + m_i + e_{ij}$$

- y_{ij} denotes oil consumption for the j^{th} measurement of the i^{th} machine
- The parameter m_i denotes how the consumption for machine i differs from the overall mean μ
- e_{ij} denotes the error for the j^{th} measurement of the i^{th} machine
- The ANOVA model assumes that all these errors are normally distributed with zero mean and equal variances

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Testing

- In our example, we test the hypothesis:

  $$H_0\text{: } m_1 = m_2 = m_3 = m_4 = 0$$

  Or, more simply, that the machine means are the same

- Intuitively, this is done by looking at the difference between means relative to the difference between observations, i.e. is the mean to mean variation greater than you would expect by chance?

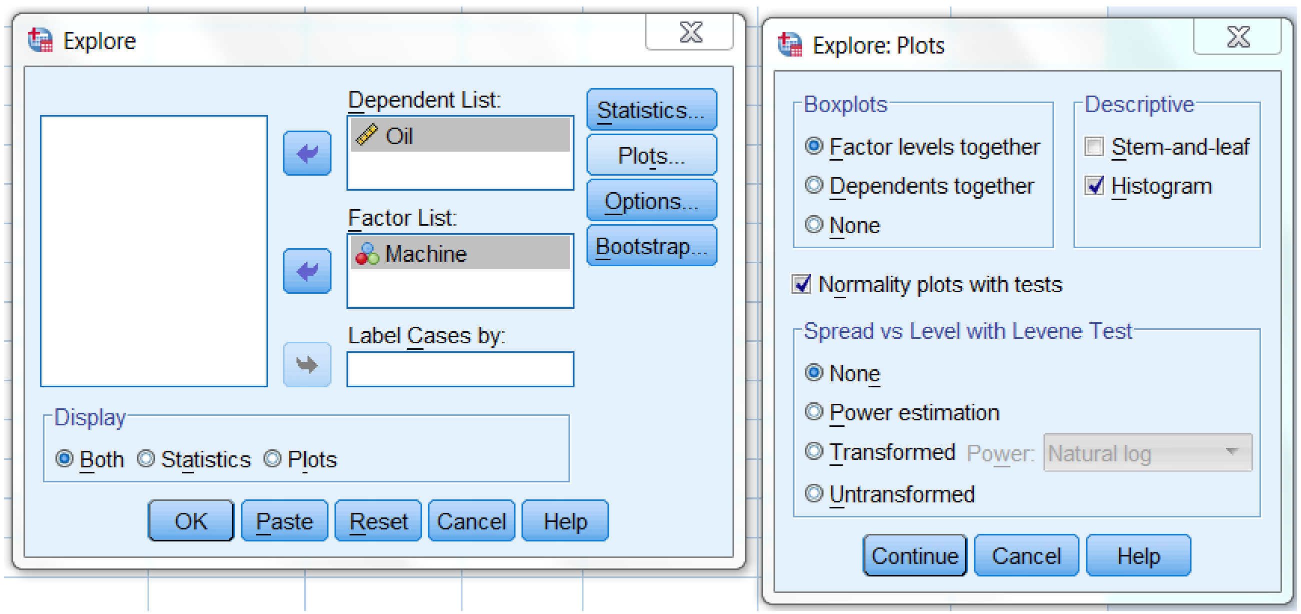
Assumptions

(Similar to the two-sample unpaired t-test)

1. The dependent values y_{ij} are normally distributed for each i. However, if there are many groups there is a danger of a Type I error.
2. The errors e_{ij} for the whole data set are normally distributed. But we must estimate the sample means ($\mu + m_i$) first. (This theoretically follows from Assumption 1, but it is worth testing separately with small samples.)
3. The variances of each group are equal

Assumption 1: Testing each group for normality

- Analyze – Descriptive Statistics – Explore
- Choose the variables as shown
- Select Plots… and choose Histogram and Normality plots with tests as shown



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Tests of Normality

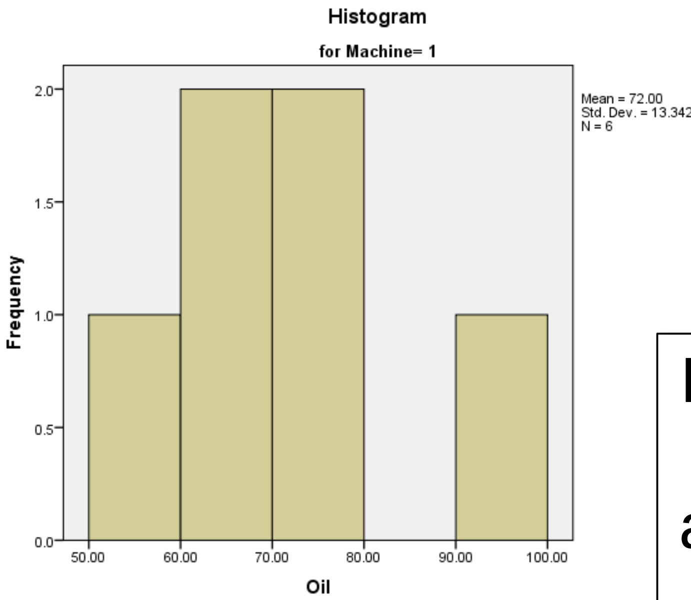
	Machine	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Oil	1	.187	6	.200*	.950	6	.741
	2	.167	6	.200*	.932	6	.593
	3	.253	6	.200*	.933	6	.607
	4	.263	6	.200*	.888	6	.310

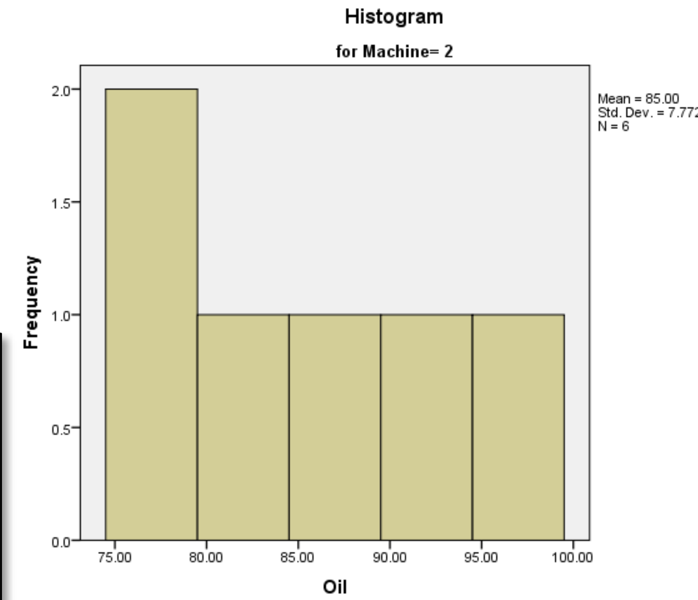
a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

- Shapiro-Wilk test significance levels are all greater than 0.1 (look at this test first for small to medium sizes, up to one or two thousand)
- No evidence that individual machine data is not normally distributed

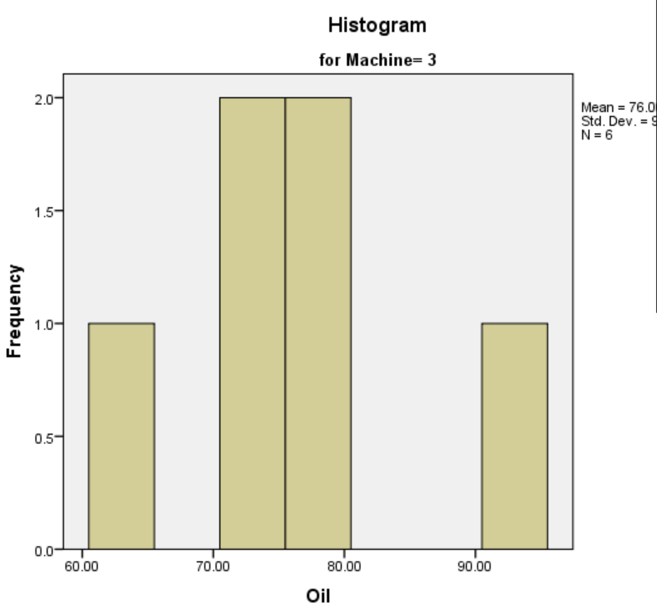
Peter Samuels
Birmingham City University

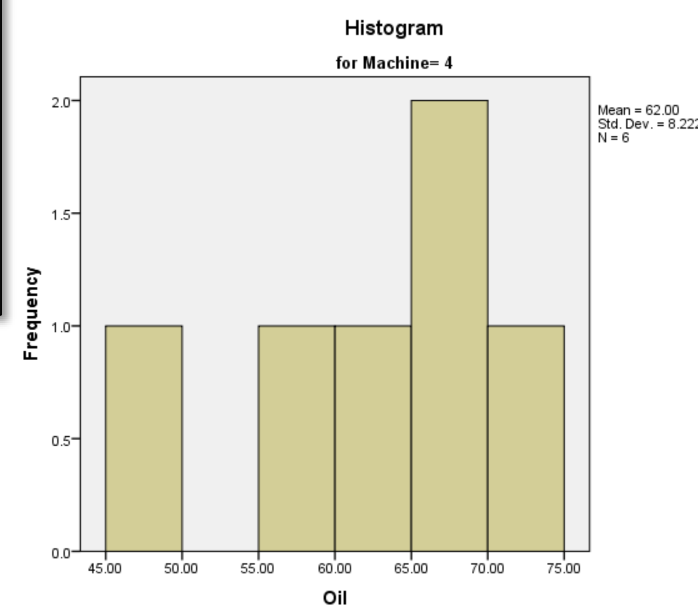
Reviewer: Ellen Marshall
University of Sheffield





Histograms are acceptable, taking into account the small sample sizes



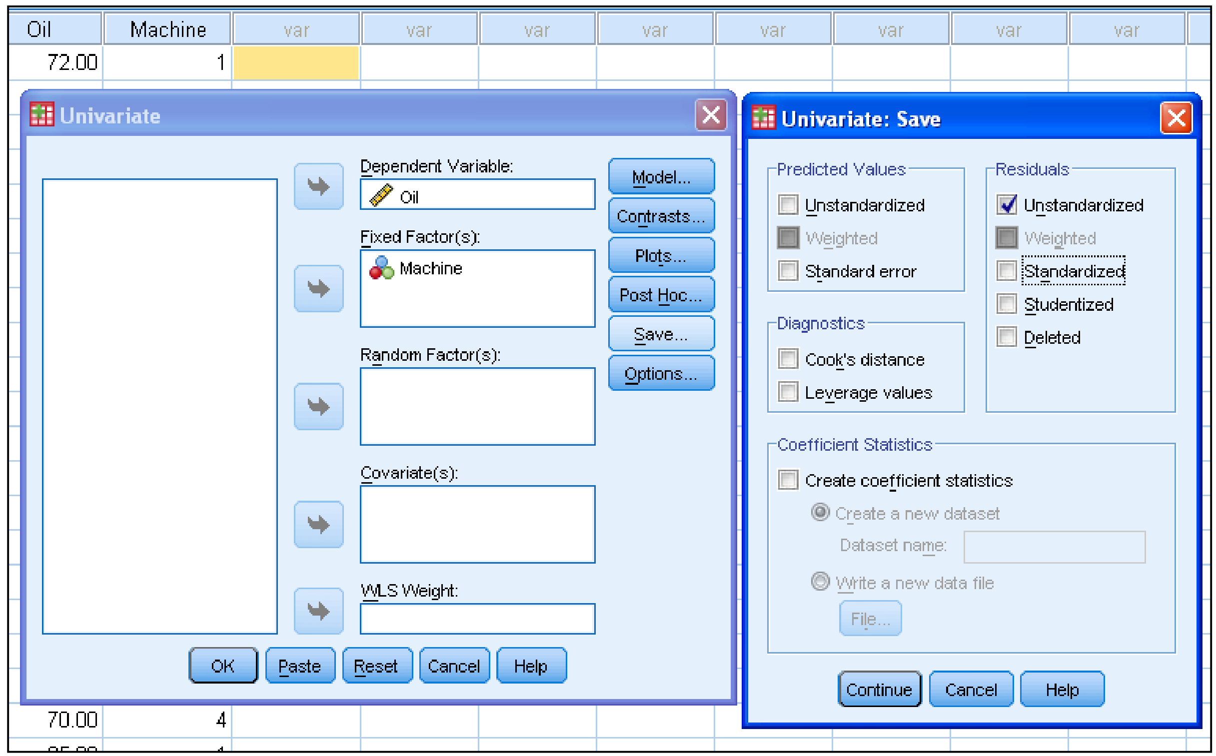


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Assumption 2: Testing errors for normality

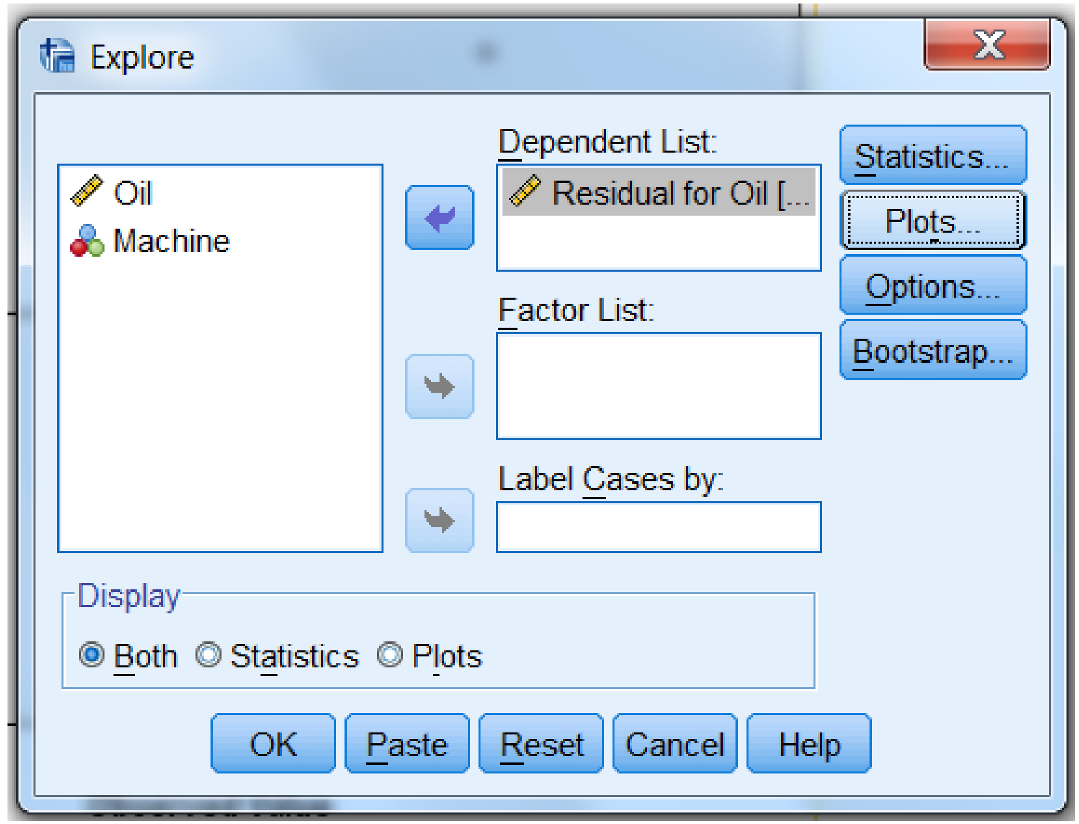
- First create the residuals
- Select Analyze – General linear model – Univariate
- Add the variables as shown
- Select Save…
- Choose Unstandard-ised residuals
- Based on estimates of m_i



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Select Analyze – Descriptive Statistics – Explore
- Add the residual variable as shown
- Keep the Plots… settings as before



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for Oil	.094	24	.200*	.972	24	.721

a. Lilliefors Significance Correction

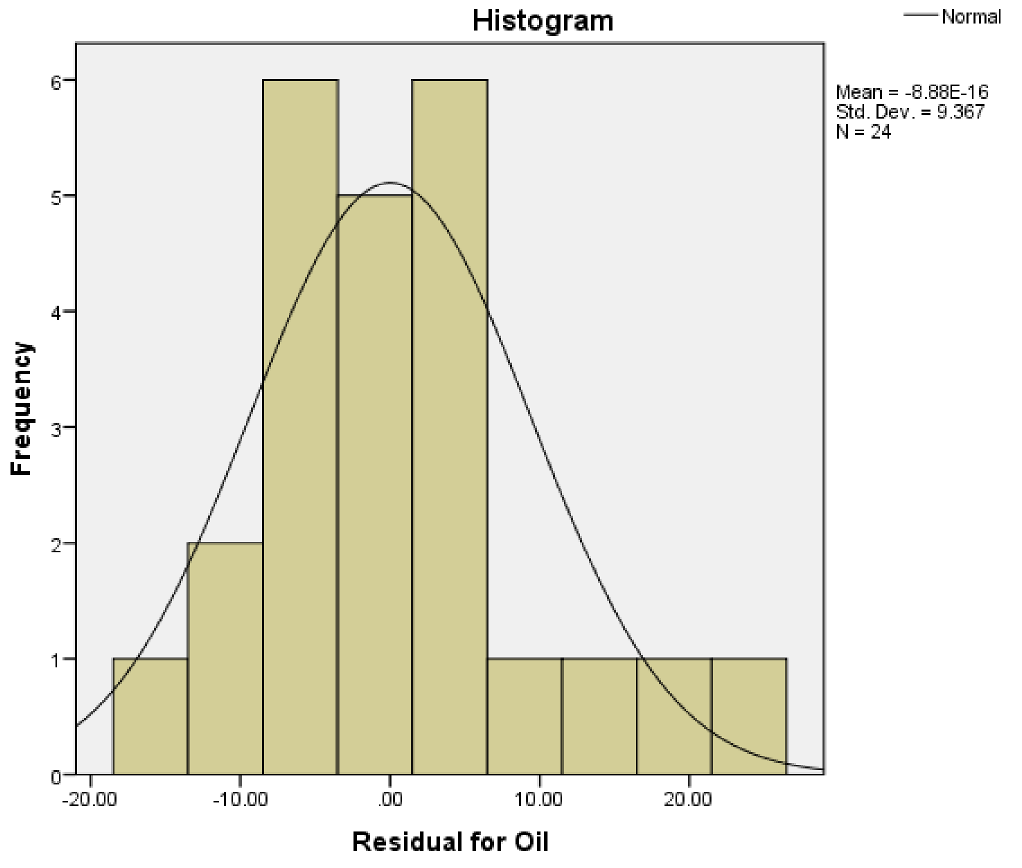
*. This is a lower bound of the true significance.

- Significance level of Shapiro-Wilk test is greater than 0.1
- No evidence that the residuals are not normally distributed
- However, a slightly higher threshold is required than usual because we have already estimated the group means $\mu + m_i$ (and thus reduced the degrees of freedom)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

The histogram is again acceptable. The sample size is now 24. A normal curve approximation has been added using the Chart Editor window.

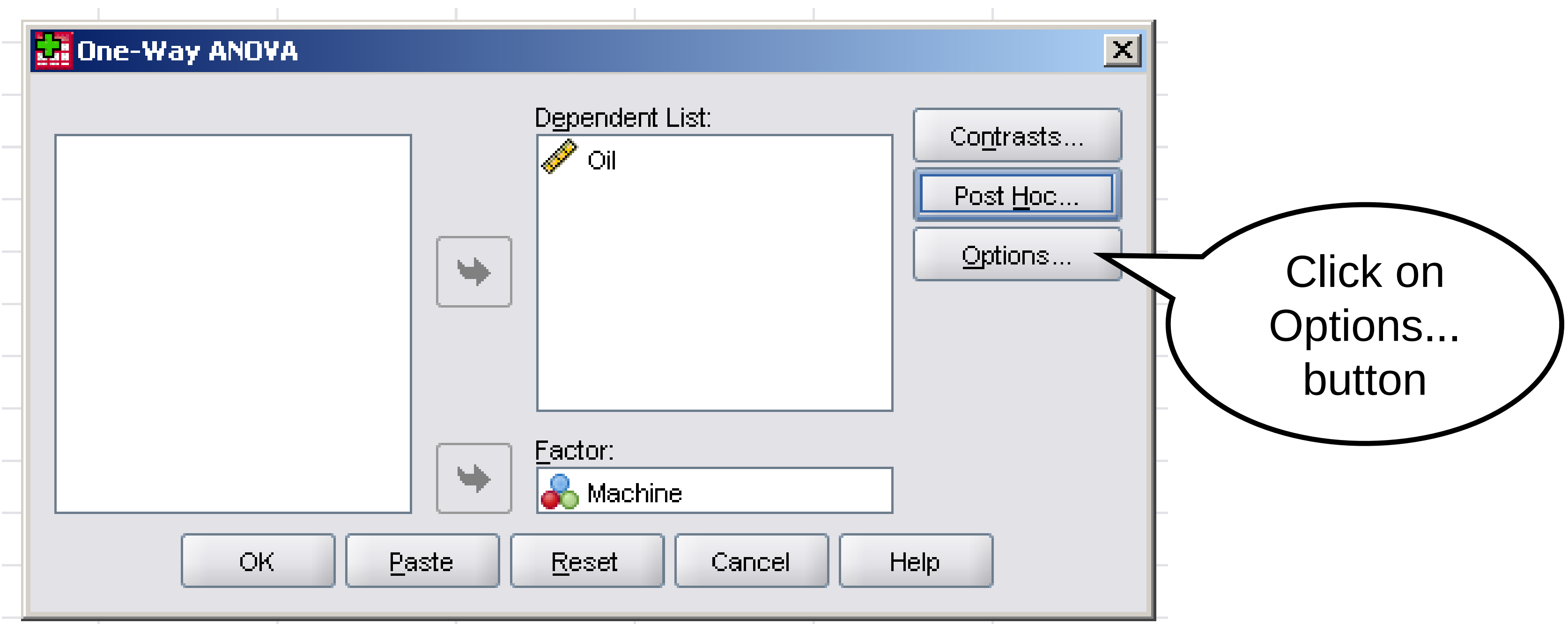


Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

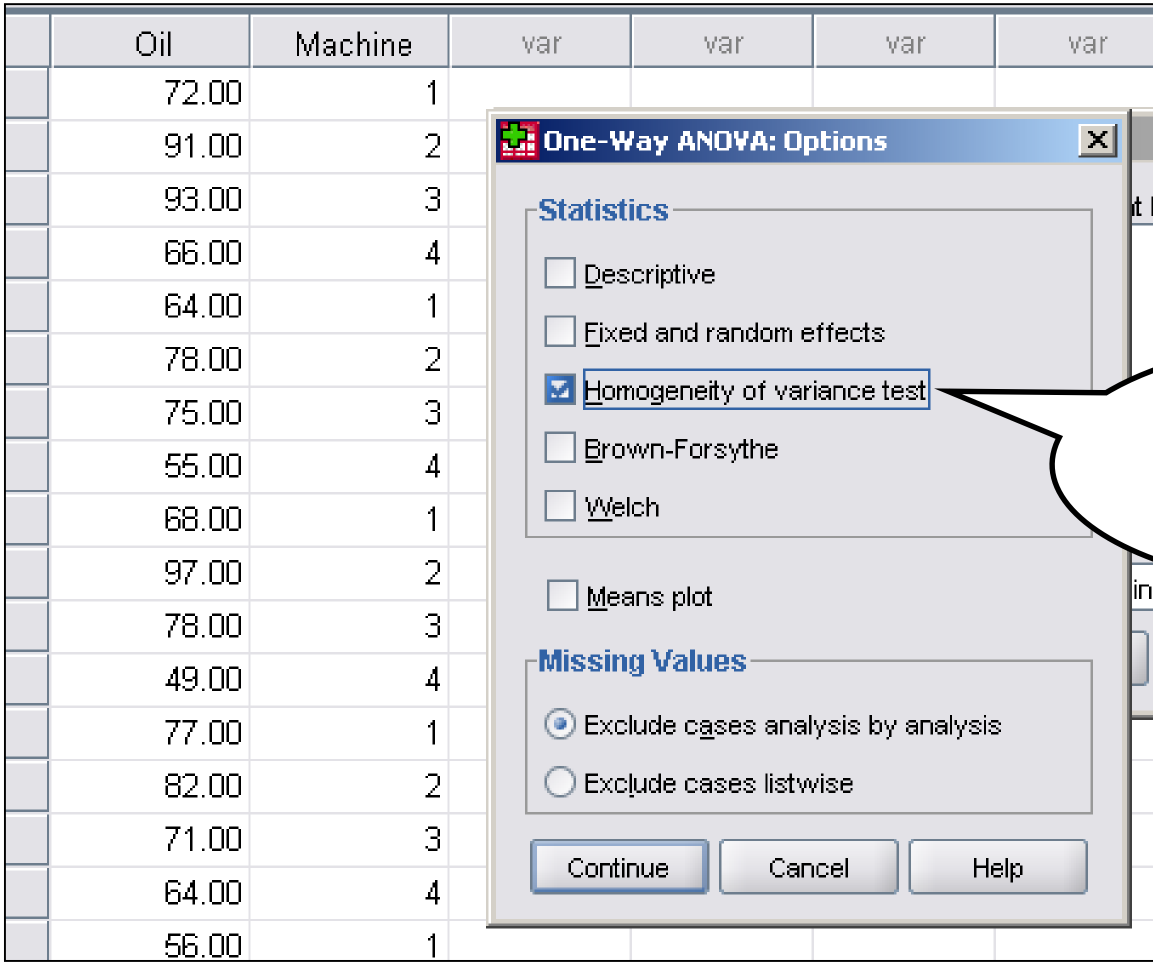
Assumption 3: Equal variances for Oil data

Analyze → Compare Means → One-Way ANOVA



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield



Click on Homogeneity of variance test

- This carries out a Levene's test for homogeneity of variance
- Null hypothesis: the variances are equal

Test of Homogeneity of Variances

Oil

Levene Statistic	df1	df2	Sig.
.361	3	20	.782

- Significance value > 0.1 so we have no evidence to doubt assumption of equal variances

Example 2

- A research project involving three different designs of a new product
- Tested by 60 people
- Each person was assigned to assess one product, providing in an overall performance score out of 100
- 20 people per product

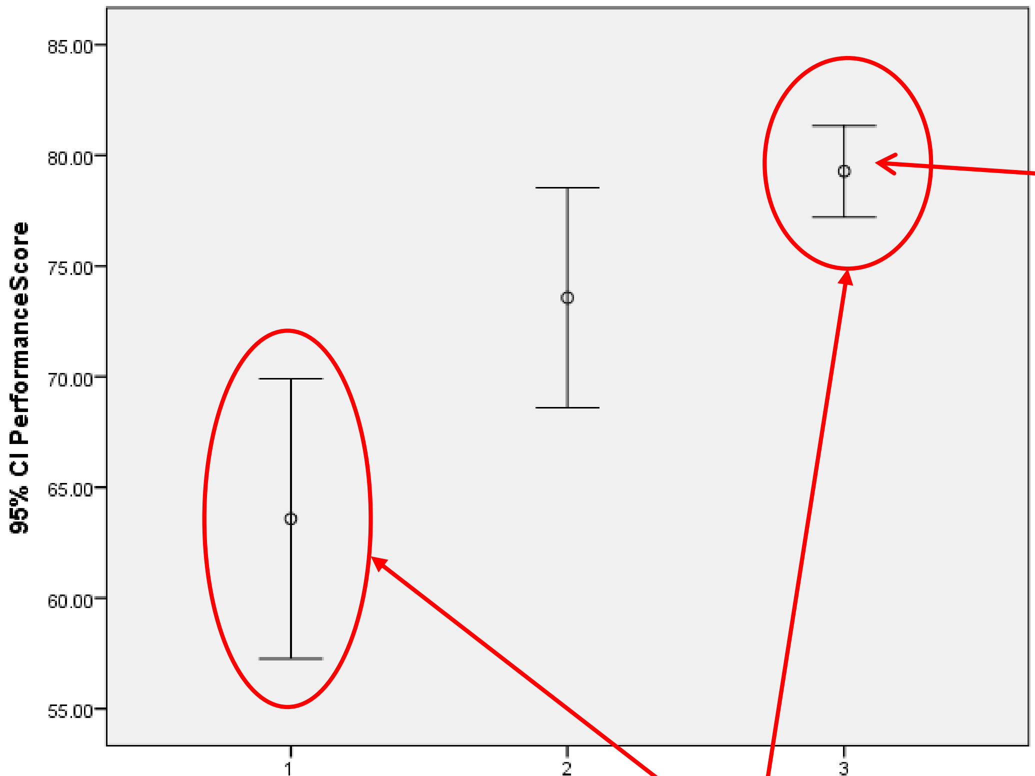
$\Rightarrow$ Create summary statistics and an error bar chart

$\Rightarrow$ Describe the data

$\Rightarrow$ Test the ANOVA assumptions

$\Rightarrow$ Interpret the output

Error bar chart (*PerformanceScore* v. *Design*)



Performance scores for *Design 3* seems to be quite different from the other two groups, especially *Design 1*.

The variance of *Design 3* also seems to be smaller.

As before, these confidence intervals clearly don't overlap, indicating likely significant differences

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Check normality of each group

- Analyze – Descriptive Statistics – Explore
- Select *PerformanceScore* in the Dependent list and *Design* as the factor
- Select Normality plots with tests and Histograms under Plots…

Tests of Normality

	Design	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
PerformanceScore	1	.139	20	.200*	.957	20	.494
	2	.134	20	.200*	.948	20	.344
	3	.153	20	.200*	.962	20	.582

*. This is a lower bound of the true significance.

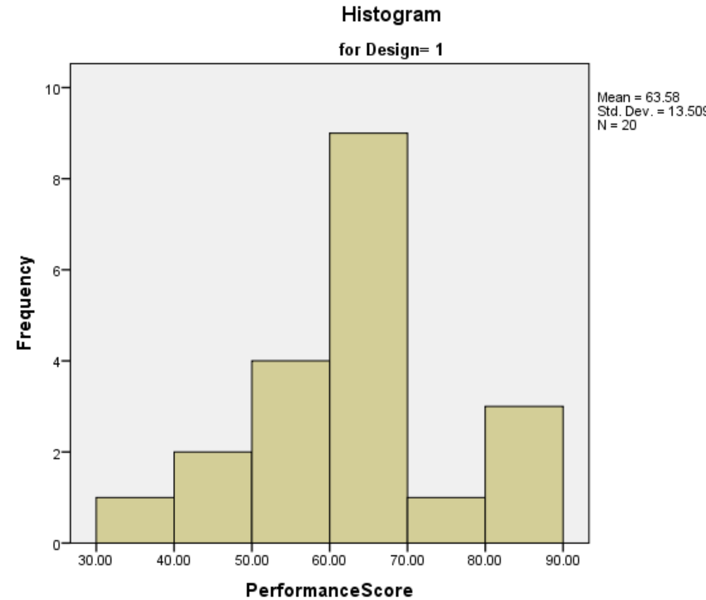
a. Lilliefors Significance Correction

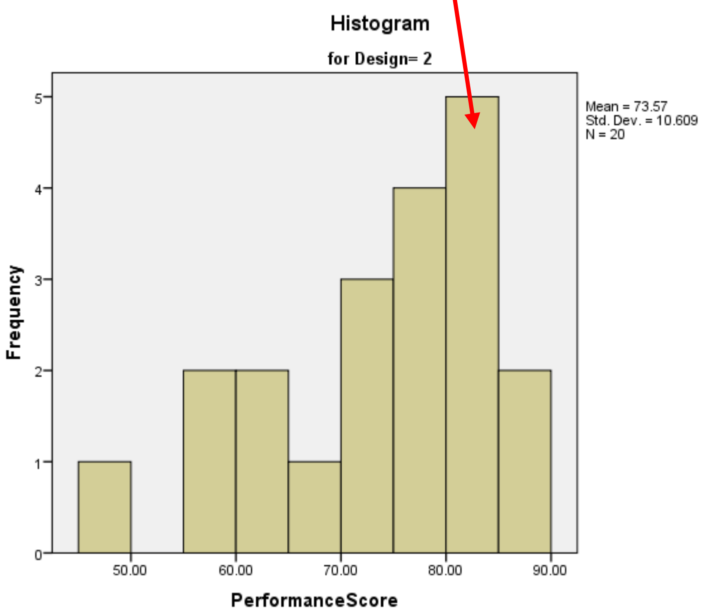
- No evidence that individual groups are not normally distributed

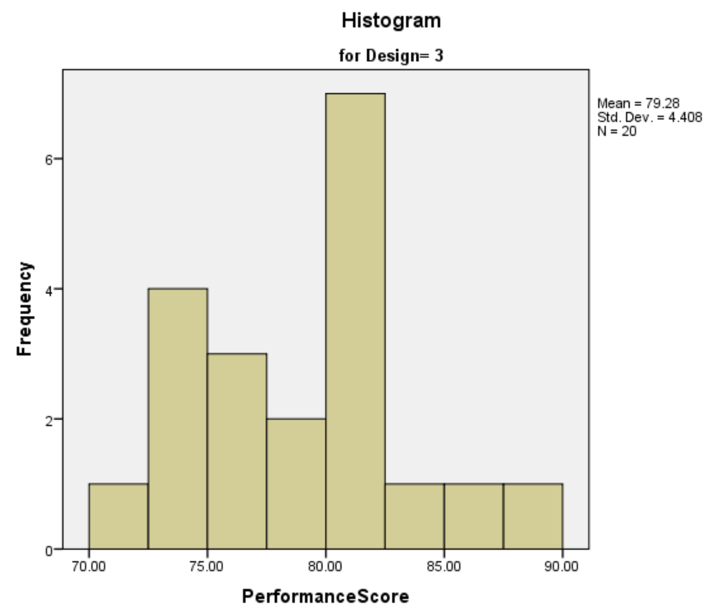
Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Histograms are fairly acceptable, although *Design 2* appears to have a slight negative skew (although it is less than twice its standard error)







Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Normality of errors check

- Analyze – General Linear Model – Univariate
- Save… Unstandardised Residuals as before
- Analyze – Descriptive Statistics – Explore
- Select *Residual for PerformanceScore* as the variable
- Select Plots… Normality plots with tests

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual for PerformanceScore	.123	60	.025	.957	60	.032

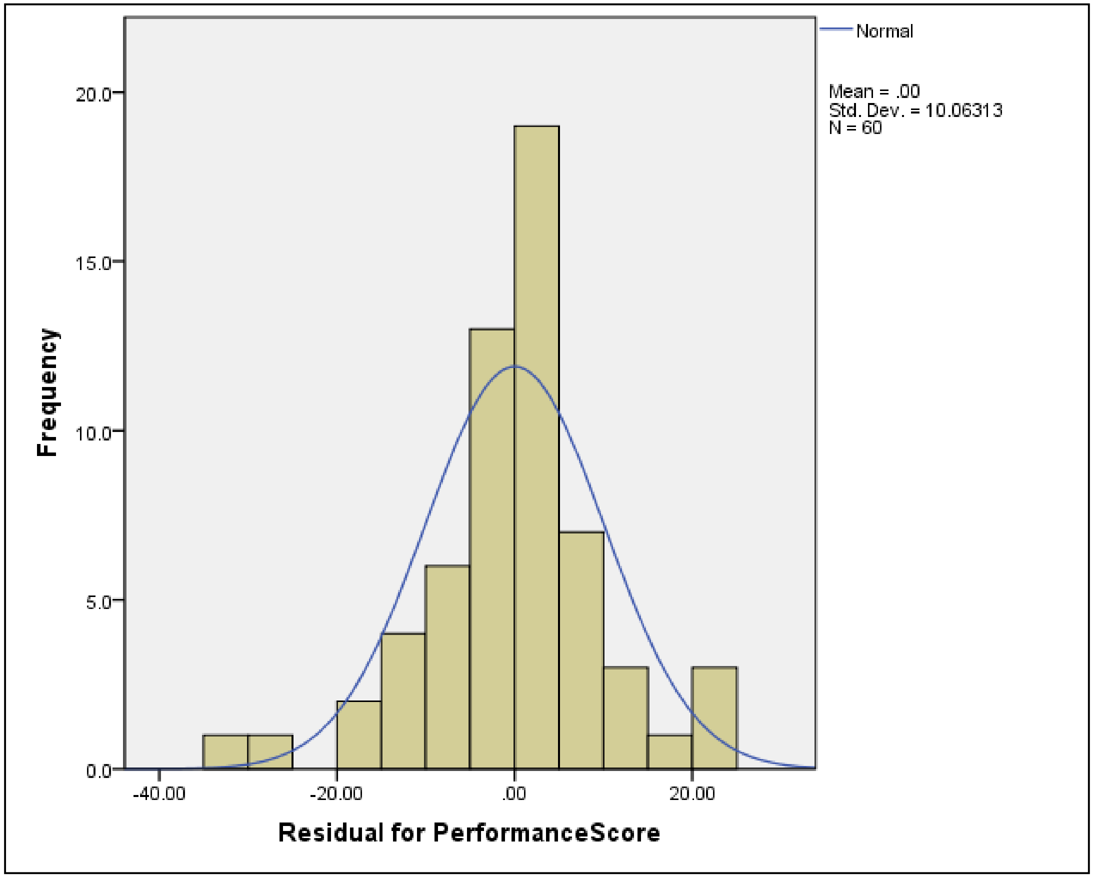
a. Lilliefors Significance Correction

- Evidence that residuals are not normally distributed from Shapiro-Wilk test ($p < 0.05$)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

- Kurtosis looks a bit high – it is 1.553
- Its standard error is 0.608
- So it is more than twice its standard error



Equality of variances check

- Analyze – Compare Means – One-Way ANOVA
- Select Options… and Homogeneity of variance test

Test of Homogeneity of Variances

PerformanceScore

Levene Statistic	df1	df2	Sig.
4.637	2	57	.014

- Significance value < 0.05 so we do have evidence to reject the assumption of equal variances

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Robustness of ANOVA

- ANOVA is quite robust to changes in skewness but not to changes in kurtosis. Thus, it should not be used when:

$$\frac{|Kurtosis|}{Standard\ Error\ of\ Kurtosis} > 2$$

for any group.

- Otherwise, provided the group sizes are equal and there are at least 20 degrees of freedom, ANOVA is quite robust to violations of its assumptions
- However, the variances must still be equal

Source:

Glass, G. V., Peckham, P. D. and Sanders, J. R. (1972) Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance, *Review of Educational Research*, 42(3), pp. 237-288.

Robustness calculation for Example 2

Group	Kurtosis	Standard Error of Kurtosis	$\frac{\lvert Kurtosis \rvert}{Standard\ Error\ of\ Kurtosis}$
1	0.493	0.992	0.497 < 2
2	0.435	0.992	0.439 < 2
3	0.115	0.992	0.116 < 2

- ❑ Group sizes are equal
- ❑ Total degrees of freedom = 20 + 20 + 20 – 1 = 59 > 20
- ❑ All OK so far
- ❑ However, ANOVA cannot be used because the variances are not equal

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Summary of findings: ANOVA assumptions

Example	1	2
Normality of groups	No evidence of non-normality	No evidence of non-normality
Normality of residuals	No evidence of non-normality	Evidence of non-normality
Equality of variances	No evidence of non-equality	Evidence of non-equality
Robustness	N/A	Satisfied apart from non-equality of variances

Peter Samuels
Birmingham City University

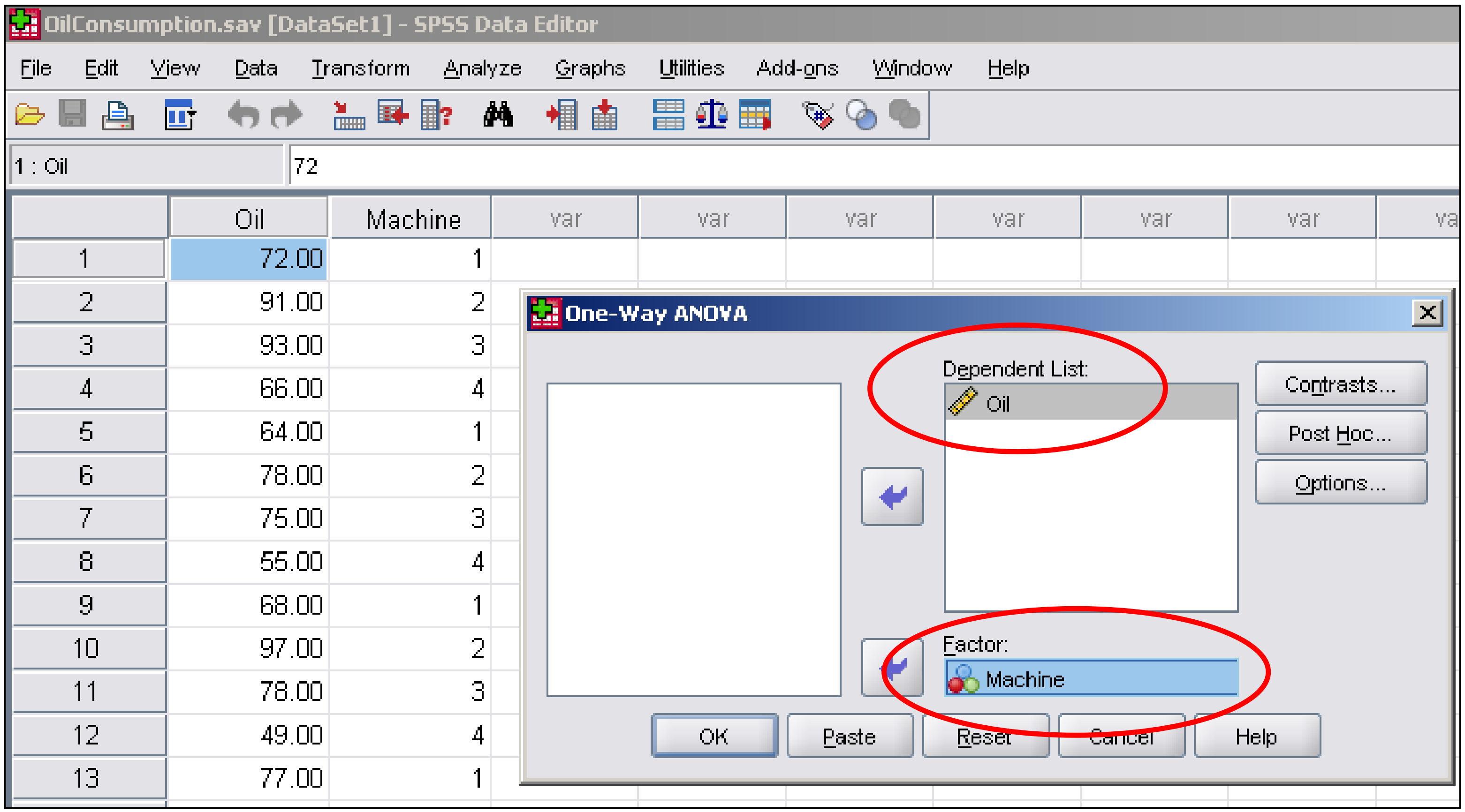
Reviewer: Ellen Marshall
University of Sheffield

What if these assumptions are in doubt?

- If normality assumptions are in doubt:
  - Use a **non-parametric** test: Kruskal-Wallis (general) or Jonckheere-Terpstra (where the groups are in a sequence and you wish to look for a linear trend)
  - Select Analyze – Nonparametic Tests – Independent Samples… then select these tests on the Settings tabs after selecting Customise Tests
- If variances assumption in doubt:
  - Use the **Brown-Forsythe** or **Welch** test (the Welch test is more powerful except where there is an extreme mean with a large variance when the Brown-Forsyth is better)
  - Select ANOVA and click on Options… button and select the **Brown-Forsythe** and **Welch** options
  - Use the significance values there instead

Example 1

- All 3 assumptions are OK so use normal ANOVA
- Analyze – Compare Means – One-Way ANOVA



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

SPSS output

ANOVA

Oil

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1636.500	3	545.500	5.406	.007
Within Groups	2018.000	20	100.900		
Total	3654.500	23			

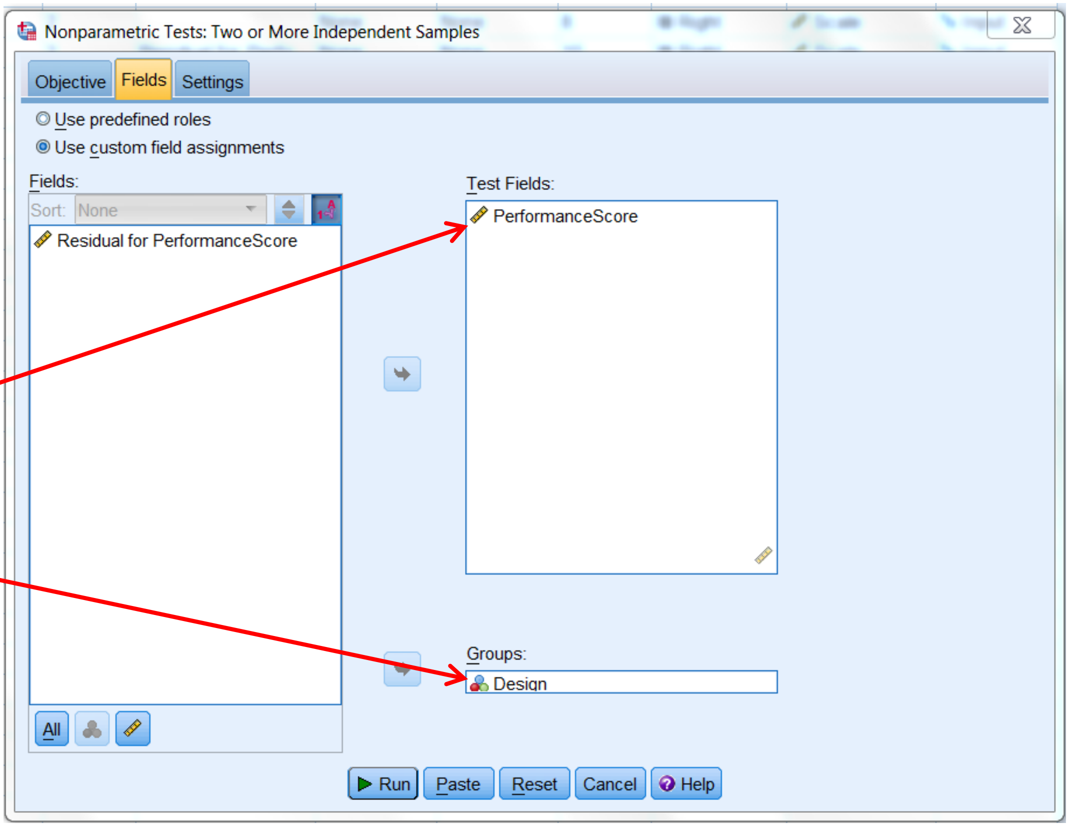
- Significant at 0.01
- So there is strong evidence of differences in mean oil consumption between the four machines

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 2

- Normality cannot be assumed and groups are not ordered so use the Kruskal-Wallis test
- Select Analyze – Nonparametric tests – Independent Samples…
- Add *PerformanceScore* and *Design* on the Groups tab



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of PerformanceScore is the same across categories of Design.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- Give a p-value < 0.001
- Very strong evidence that there are differences between the groups

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

However, ANOVA was robust for Example 2 apart from the differences in variances so we can also use the Brown-Forsythe or Welch test:

Robust Tests of Equality of Means

PerformanceScore

	Statistic[a]	df1	df2	Sig.
Welch	13.278	2	30.962	.000
Brown-Forsythe	12.048	2	40.540	.000

a. Asymptotically F distributed.

- Both tests are significant at the 0.001 level
- Thus there is very strong evidence that the means are not equal

Multiple comparisons

- What if we conclude there are differences between the groups?
- We don't know where differences are!
- We can do **post-hoc** tests to compare each pair of groups
- Similar to 2-sample tests but adjusted for the multiple testing issue

Which post hoc test?

- For equal group sizes and similar variances, use **Tukey (HSD)** or, for guaranteed control over Type I errors (more conservative), use **Bonferroni**
- For slightly different group sizes, use **Gabriel**
- For very different group sizes, use **Hochberg's GT2**
- For unequal variances, use **Games-Howell**

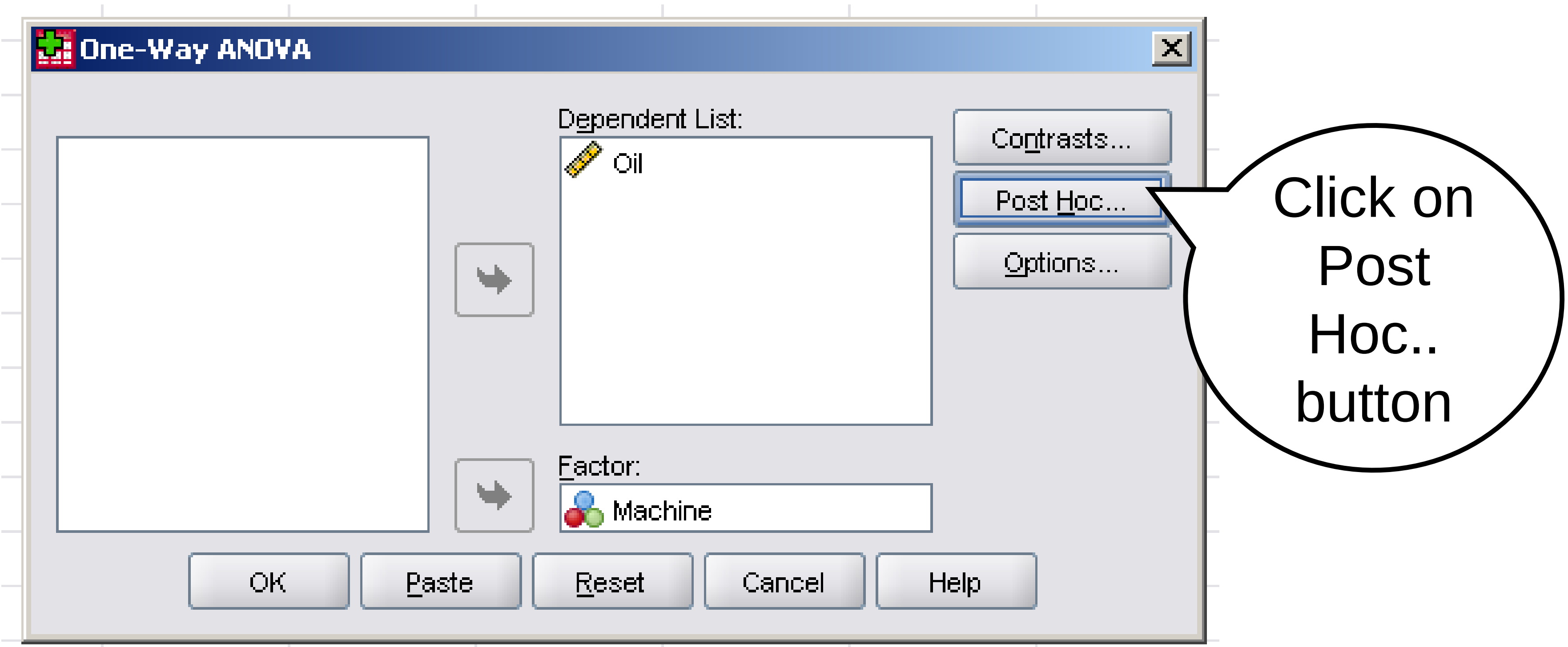
Source: (Field, 2013: 459)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 1

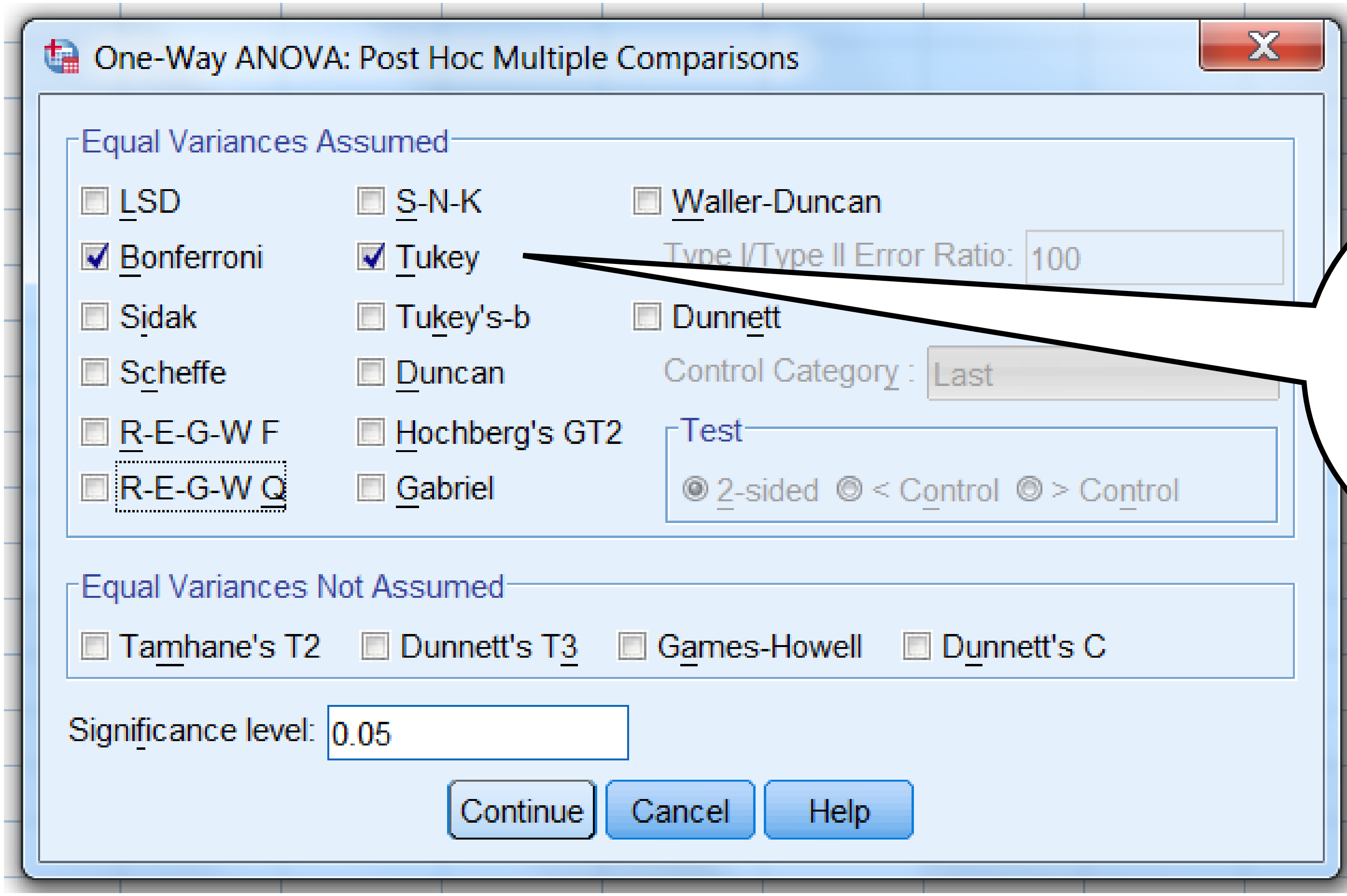
Analyze – Compare Means – One-Way ANOVA



Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multiple comparisons in SPSS



Choose Tukey and Bonferoni tests

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multiple Comparisons

Dependent Variable:Oil

	(I) Machine	(J) Machine	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	1	2	-13.00000	5.79943	.146	-29.2322	3.2322
		3	-4.00000	5.79943	.900	-20.2322	12.2322
		4	10.00000	5.79943	.338	-6.2322	26.2322
	2	1	13.00000	5.79943	.146	-3.2322	29.2322
		3	9.00000	5.79943	.427	-7.2322	25.2322
		4	23.00000*	5.79943	.004	6.7678	39.2322
	3	1	4.00000	5.79943	.900	-12.2322	20.2322
		2	-9.00000	5.79943	.427	-25.2322	7.2322
		4	14.00000	5.79943	.107	-2.2322	30.2322
	4	1	-10.00000	5.79943	.338	-26.2322	6.2322
		2	-23.00000*	5.79943	.004	-39.2322	-6.7678
		3	-14.00000	5.79943	.107	-30.2322	2.2322
Bonferroni	1	2	-13.00000	5.79943	.219	-29.9756	3.9756
		3	-4.00000	5.79943	1.000	-20.9756	12.9756
		4	10.00000	5.79943	.600	-6.9756	26.9756
	2	1	13.00000	5.79943	.219	-3.9756	29.9756
		3	9.00000	5.79943	.818	-7.9756	25.9756
		4	23.00000*	5.79943	.005	6.0244	39.9756
	3	1	4.00000	5.79943	1.000	-12.9756	20.9756
		2	-9.00000	5.79943	.818	-25.9756	7.9756
		4	14.00000	5.79943	.153	-2.9756	30.9756
	4	1	-10.00000	5.79943	.600	-26.9756	6.9756
		2	-23.00000*	5.79943	.005	-39.9756	-6.0244
		3	-14.00000	5.79943	.153	-30.9756	2.9756

*. The mean difference is significant at the 0.05 level.

- Only significant difference for Tukey HSD is between Machines 2 and 4
- Strong evidence ($p < 0.01$) that Machine 2 uses more oil than Machine 4
- Significance levels are higher and confidence interval bounds are smaller than for Bonferroni, as expected

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Multiple comparisons conclusions

- Only significant difference is between Machines 2 and 4
- Strong evidence ($p < 0.01$) with both tests that Machine 2 uses more oil than Machine 4
- 95% confidence interval for difference between machines is approximately 7 to 39 litres/week
- No evidence of differences in oil usage between other machines (because all the other confidence intervals for Tukey HSD contain 0)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Example 2

- As normality cannot be assumed, need to use nonparametric tests

Hypothesis Test Summary

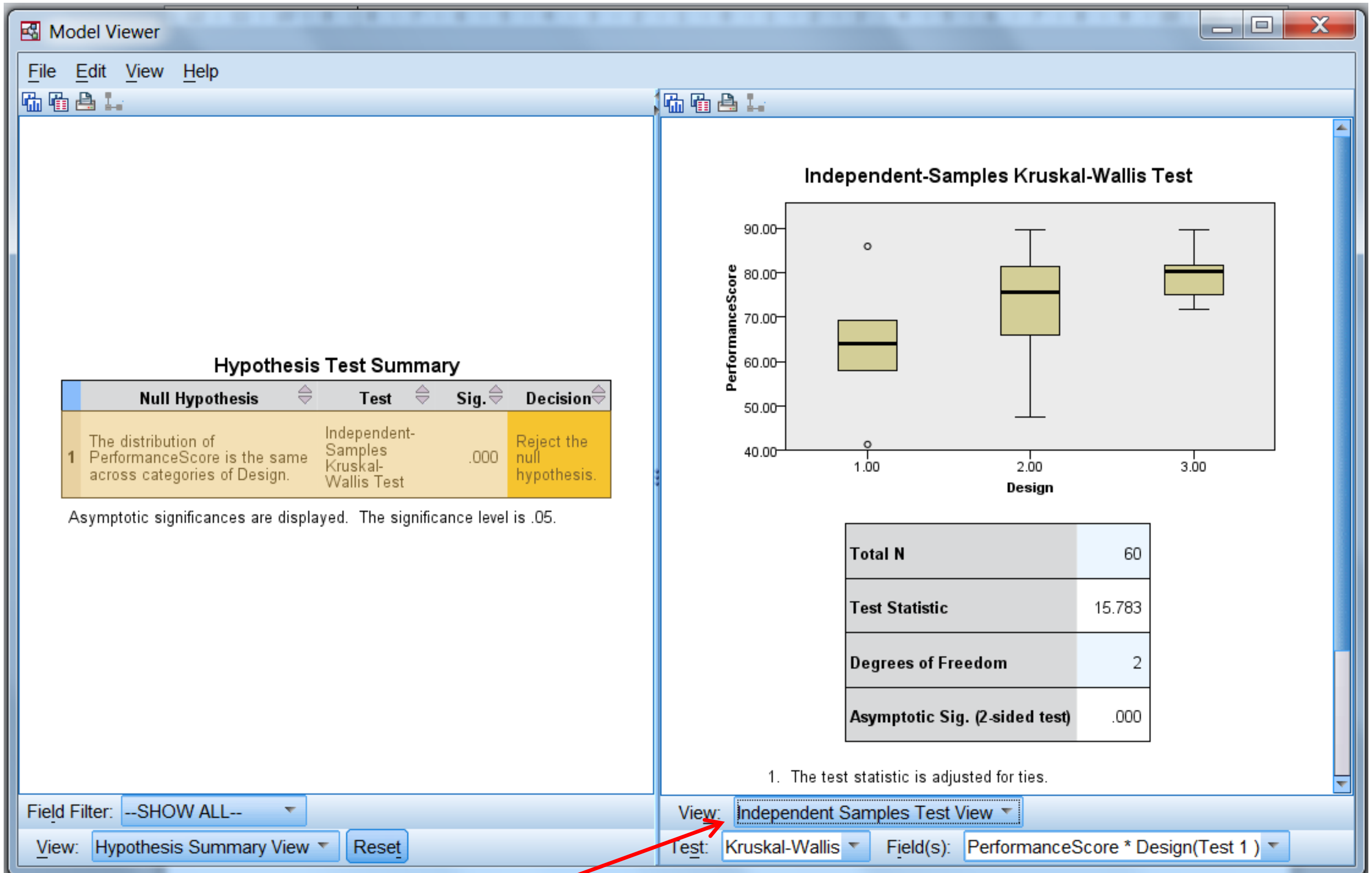
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of PerformanceScore is the same across categories of Design.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Double-click on this note to open the Model Viewer dialogue box

Peter Samuels
Birmingham City University

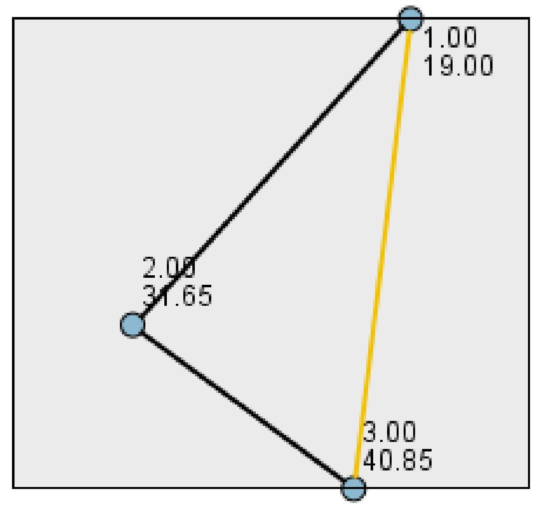
Reviewer: Ellen Marshall
University of Sheffield



Change the view option to Pairwise Comparisons

- The adjusted significance values are corrected using an equivalent to the Bonferroni correction for parametric ANOVA
- Very strong evidence of a difference between groups 1 and 3
- Weak evidence of a difference between groups 1 and 2

Pairwise Comparisons of Design



Each node shows the sample average rank of Design.

Sample 1-Sam...	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
0-1	-12.650	5.523	-2.291	.022	.066
0-2	-21.850	5.523	-3.956	.000	.000
1-2	-9.200	5.523	-1.666	.096	.287

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

However, as ANOVA was robust apart from the equality of variances assumption we can also use the Games-Howell post hoc test:

More powerful conclusions than the nonparametric tests

Multiple Comparisons

PerformanceScore
Games-Howell

(I) Design	(J) Design	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
1	2	-9.98789*	3.84079	.035	-19.3762	-.5996
	3	-15.69947*	3.17733	.000	-23.6566	-7.7424
2	1	9.98789*	3.84079	.035	.5996	19.3762
	3	-5.71158	2.56883	.086	-12.1043	.6812
3	1	15.69947*	3.17733	.000	7.7424	23.6566
	2	5.71158	2.56883	.086	-.6812	12.1043

*. The mean difference is significant at the 0.05 level.

- ❑ Very strong evidence of differences between groups 1 and 3
- ❑ Evidence of differences between groups 1 and 2
- ❑ Weak evidence of differences between groups 2 and 3

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Recap

We have considered:

- Describing multiple groups:
  - Scatter plots
  - Means and standard deviations
  - Boxplots
- Checking assumptions:
  - Normality of each group (Shapiro-Wilk and Kolmogorov Smirnov)
  - Normality of errors (creating unstandardised residuals, then as above)
  - Equality of variances (Levene's test)
  - Robustness to violations of assumptions (kurtosis, group sizes and degrees of freedom)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield

Recap (2)

- Carrying out the ANOVA test
- Unequal variances alternatives (Brown-Forsythe and Welch)
- Nonparametric alternatives: Kruskal-Wallis (general) and Jonckheere-Terpstra (linear)
- Post hoc tests (Tukey, Bonferroni, Gabriel and Hochberg's GT2)
- Unequal variances alternative (Games-Howell)
- Nonparametric alternatives (Kruskal-Wallis pairwise comparisons)

Peter Samuels
Birmingham City University

Reviewer: Ellen Marshall
University of Sheffield