Chapter 10

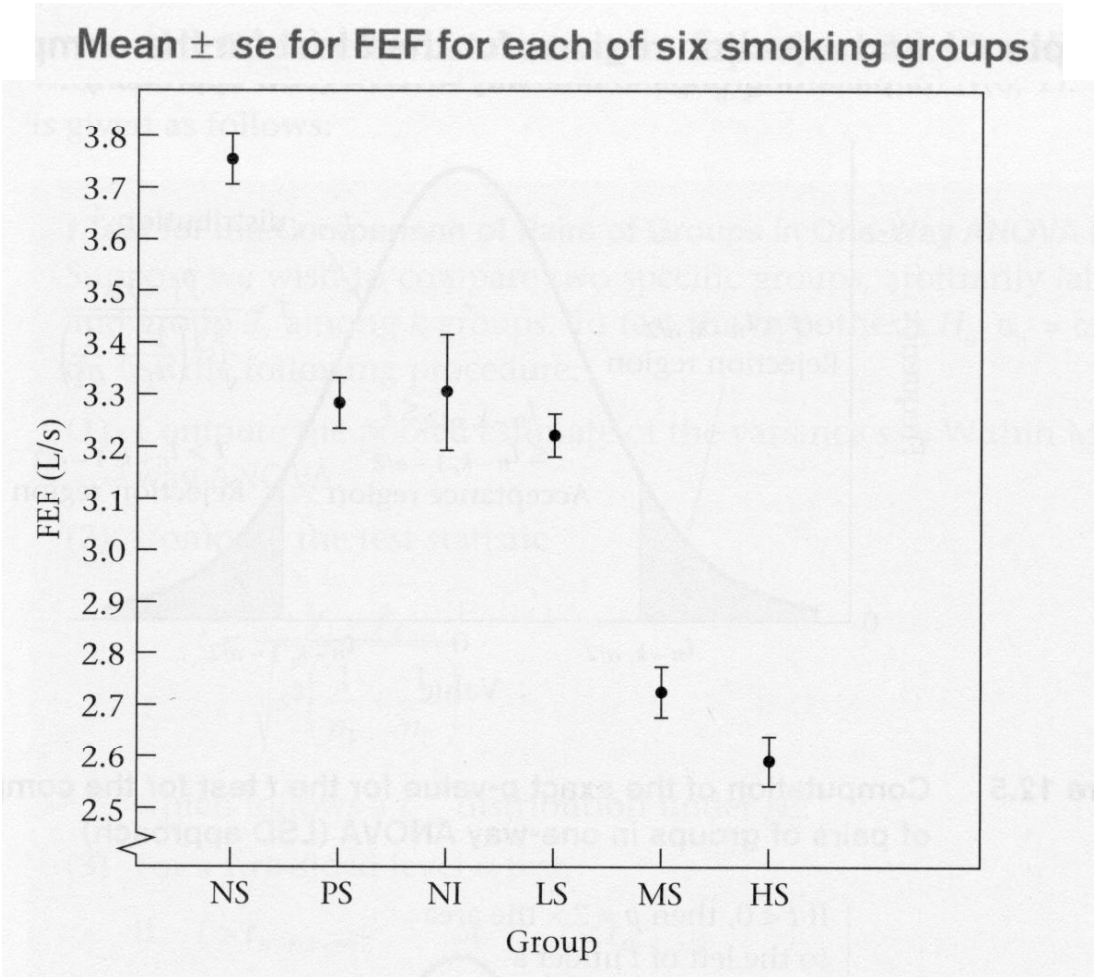
One-Way Analysis of Variance – testing means of three or more groups

Overview:

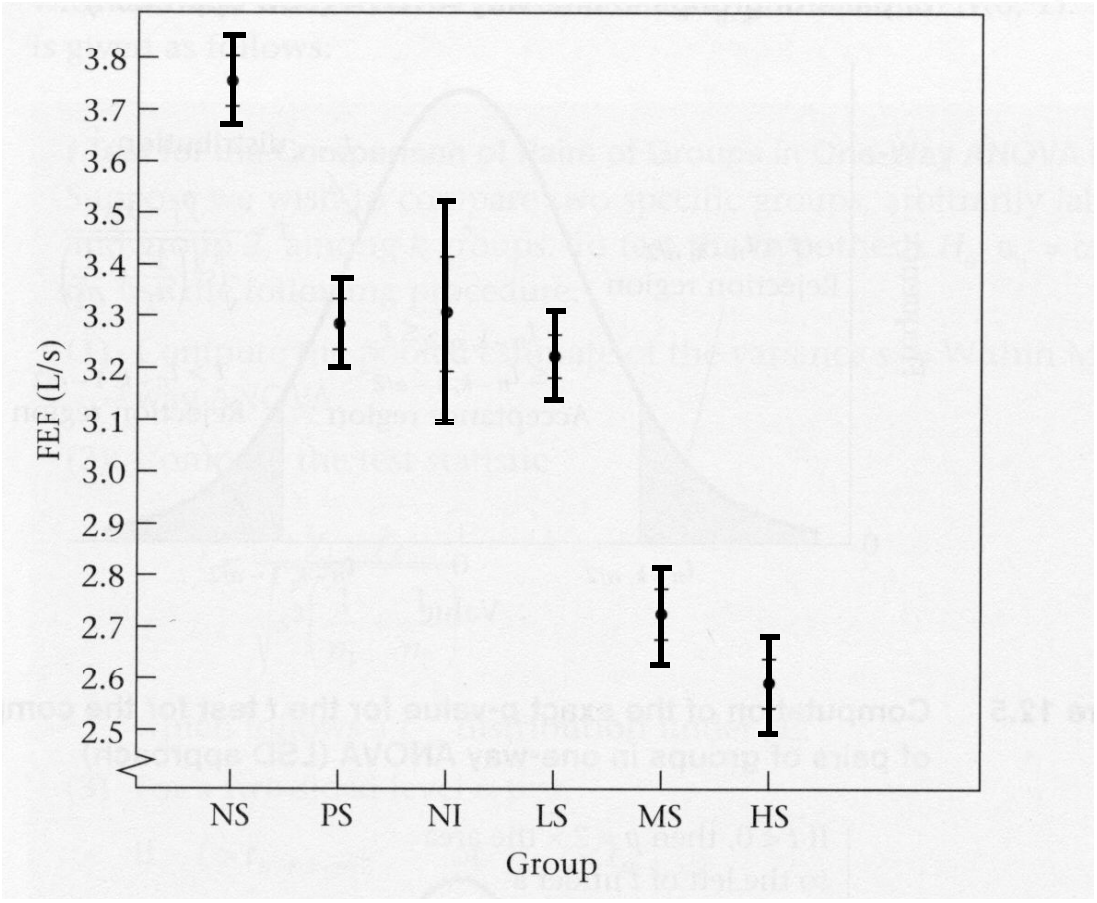
10.1 Overview of One-Way ANOVA

Example 1 (Rosner page 558): White and Froeb studied to assess whether or not passive smoking has a measurable effect on pulmonary health. Pulmonary function were measured using Forced expiratory flow rate (FEF) in the following six groups:
(Source: NEJM, 302 (13), 720-723, 1980)

(1) **Nonsmokers (NS):** People who never smoked and were not exposed to smoking either at home or on the job. (N=200)
(2) **Passive smokers (PS):** People who never smoked and were not exposed to smoking at home, but exposed on the job for 20 + years (N=200)
(3) **Non-inhaling smokers (NI):** people who smoked pipes, cigars or cigarettes, but who did not inhale. (N=50)
(4) **Light smokers (LS):** People who smoked and inhaled 1-10 cigarettes per day for 20 or more years. (N=200)
(5) **Moderate smokers (MS):** People who smoked and inhaled 11-39 cigarettes per day for 20 or more years. (N=200)
(6) **Heavy smokers (HS):** People who smoked and inhaled 40 or more cigarettes per day for 20 or more years. (N=200)



Mean ± 95% CI for FEF for each of six smoking groups



Let's consider with using the knowledge you learned so far, how do you want to analyze this data to compare FEF among groups.

In chapter 8, we learned how to compare means of 2 groups using Students t-tests. Assuming FEF are normally distributed within each group, do you think you can use Student's t-test here?

If you use Student's t-test, you may need to conduct more than 1 test. How many tests do you think you need to perform?

Problem with Multiple Comparisons (Inflation of Type I error)

With 6 groups, you can perform up to 15 Student' t-tests (6 x 5 x ½ = 15) comparing any 2 groups (NS vs PS, NS vs NI, and so on...) This causes a problem of multiple comparisons. The more tests you conduct, chances to observe results with $p<0.05$ becomes higher. For example, if you detect a test with statistical significance with $p<0.05$ among 20 tests you performed, the probability detecting significance at least with one test is 64% when there is no underlying association.

We call this "inflation of type I error" often viewed as "problem of multiple comparisons". To account for this, we may need to use more strict criteria to detect significance difference. For example, divide alpha level (type I error) by the number of tests being performed. With 15 tests being conducted, you need to observe $p < 0.05/15 = 0.0033$ in order to claim two group means differ (**Bonferroni adjustment**).

What SPSS does is, instead, multiply observed p-value by 15 (total number of comparisons)

Probability of having at least one test with p ＜0. 05=1-0.95^k
K: a total number of pair-wise comparisons being performed

K	Probability	Bonferroni correction*
1	0.05	0.05
2	0.10	0.10
5	0.23	0.25
10	0.40	0.50
20	0.64	1.00
30	0.79	1.00
100	0.99	1.00

* P x a total number of pair-wise comparisons

Comparison of Pairs of Groups in One-Way ANOVA: Bonferroni Multiple Comparisons Procedure

Do either of the following, not both!!!

Bonferonni adjusted alpha level

When you have a total of 15 pairs of two means to compare, use α, Type I error, 0.05 / 15=0.0033.
Reject the null if p < 0.0033

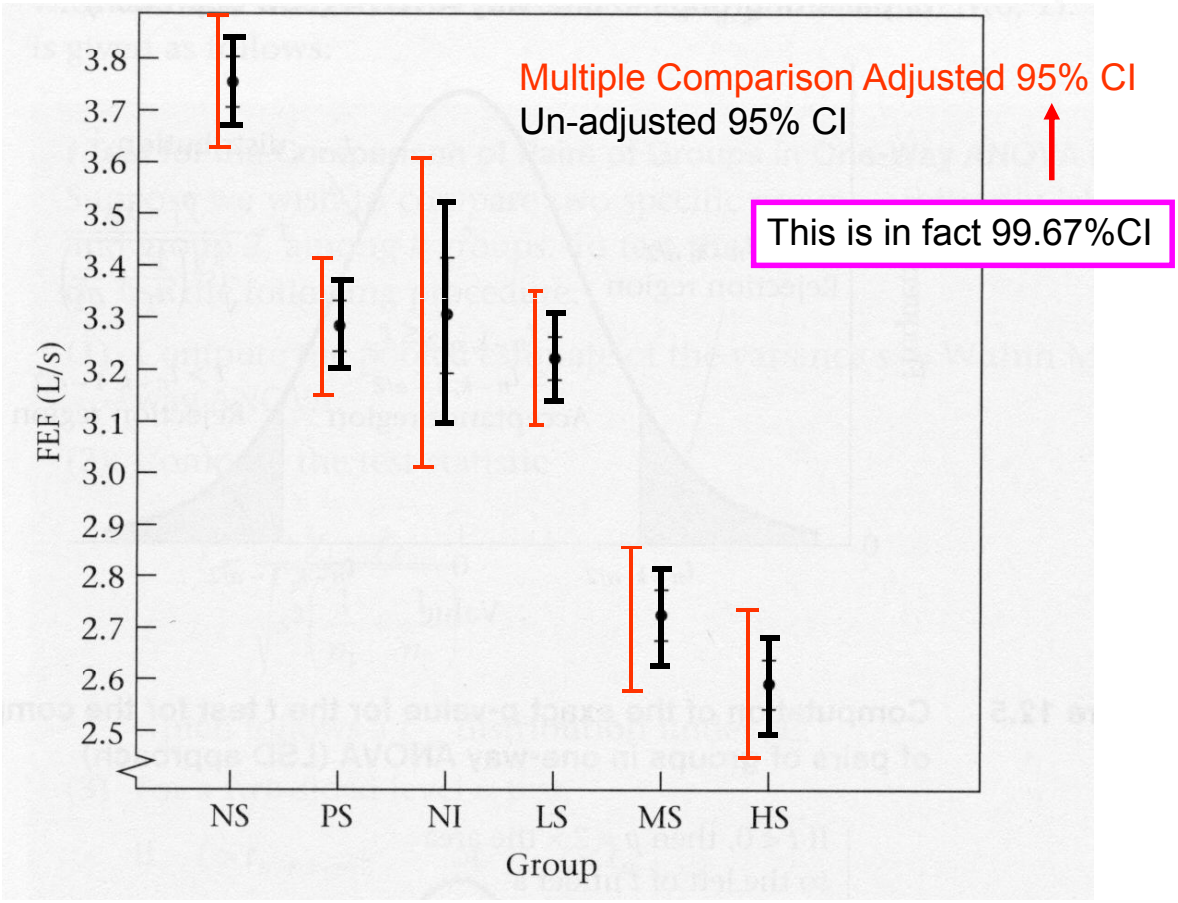
This is my recommendation

Bonferonni adjusted p-value

This is what SPSS does

When you have a total of 15 pairs of two means to compare, multiply p-value by 15.
Boferroni-adjusted p-value = original p-value x 15
Reject the null if Bonferroni-adjusted p-value < 0.05

Mean ± 95% CI for FEF for each of six smoking groups



Thus, performing many t-tests is difficult because we lose analytical power by using more strict criteria to reject the null hypothesis (i.e., adjustment for multiple comparisons, Bonferroni .

Instead, in order to assess the association between smoking and FEF, we will use a global test for the difference in means (**ANOVA test**).

Comparing means of more than 2 groups:

Conducted ANOVA test

No → Multiple comparison Adjustment MUST

Yes

- Significant → May not have to do Multiple comparison adjustment
- Not Significant → Can't even perform Post-hoc pair-wise comparisons

10.2. Global test for overall comparison of group means:

Ho: $Mean_{NS} = Mean_{PS} = Mean_{NI} = Mean_{LS} = Mean_{MS} = Mean_{HS}$

Which means that Mean FEF are the same for all groups.
You need to remember that this is the test for equivalence of means, rejecting this does not imply any directional association between the order of smoking categories and FEF.

We need to perform ad-hoc pair-wise tests to assess the directional association.

In order to test the global hypothesis, we use a technique called "analysis of variance".

Analysis of Variance compares:

Between group variability v.s. Within group variability

Between group variability = Sum (mean of each group – over all mean)^2 over all groups

Within group variability = Sum (each observation – group mean)^2 over all groups

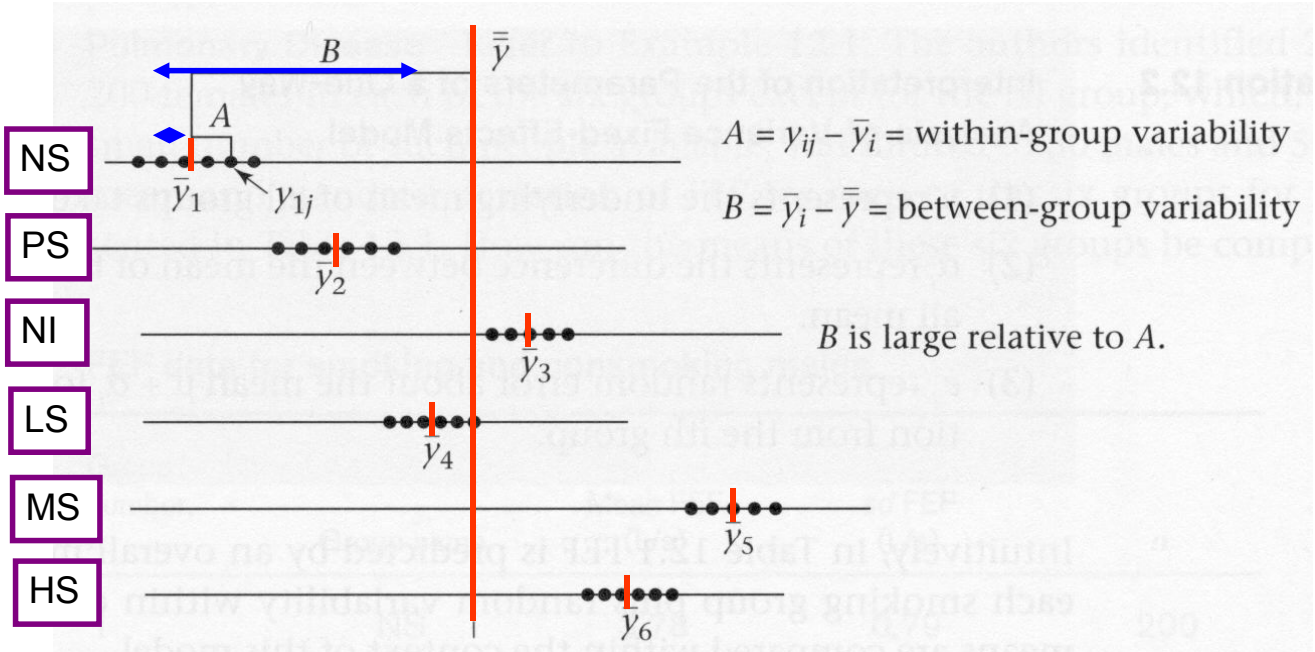
Between group variability (B) > Within group variability (A)

p-value for the test is smaller

↓

Reject the null, indicating group means are not the same

Schematic presentation when ANOVA may reject the null, I.e., Non-equivalence of the means



Between group variability (B) < Within group variability (A)

p-value for the test is larger

↓

Not reject the null, indicating group means are the same

Schematic presentation when ANOVA may not reject the null, I.e., Equivalence of the means

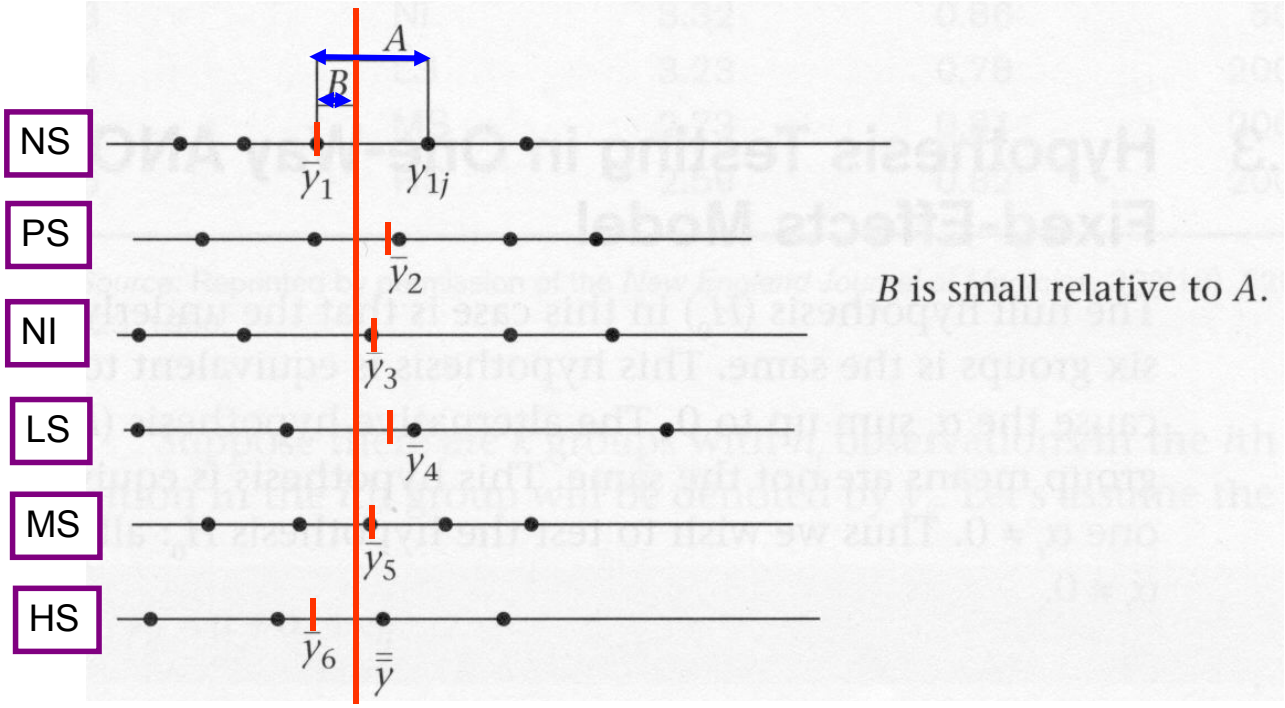


Table 12.3 ANOVA table for FEF data in Table 12.1

	SS	*df*	MS	*F* statistic	*p*-value
Between	184.38	5	36.875	58.0	$p < .001$
Within	663.87	1044	0.636		
Total	848.25				

Reject the global null, indicating that at least one mean differs.

You may conclude that there is a statistically significant association detected between smoking and FEF. But we don't know any thing about the direction of the association yet. We need to perform pair-wise analysis to determine which level of smoking differ.

10.3.1 Post Hoc pair-wise comparisons in One-Way ANOVA: Student's T-test vs LSD

Student's t-test comparing group 1 and 2 means

SD: estimated
Using 2 group data

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Least squares Difference (LSD) comparing group 1 and 2 means

SD: estimated
Using data from all groups

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

10.3.3. Post Hoc pair-wise comparisons in One-Way ANOVA: Bonferroni procedure adjusting for multiple comparisons:

As indicated previously, conducting many statistical tests for significance may cause inflation of type I error. Several procedures have been developed to deal with this problem. The basic idea of these procedures is to ensure that the overall probability of declaring any significant differences between all possible pairs of groups is maintained at some fixed significance level (say α). One of the simplest and most widely used procedure is the method of Bonferroni adjustment .

similar

Comparison #	Groups		t-test P-value	LSD P-value	Bonferroni Adjusted p-value
1	1 (NS)	2 (NI)	<.0001	<.0001	<.0001
2	2 (NI)	3 (PS)	0.564	0.557	1.00
3	3 (PS)	4 (LS)	0.237	0.215	1.00
4	4 (LS)	5 (MS)	<.0001	<.0001	<.0001
5	5 (MS)	6 (HS)	0.043	0.045	0.675
6	1 (NS)	3 (PS)	<.0001	<.0001	<.0001
7	2 (NI)	4 (LS)	0.184	0.172	1.00
8	3 (PS)	5 (MS)	<.0001	<.0001	<.0001
9	4 (LS)	6 (HS)	<.0001	<.0001	<.0001
10	1 (NS)	4 (LS)	<.0001	<.0001	<.0001
11	2 (NI)	5 (MS)	<.0001	<.0001	<.0001
12	3 (PS)	6 (HS)	<.0001	<.0001	<.0001
13	1 (NS)	5 (MS)	<.0001	<.0001	<.0001
14	2 (NI)	6 (HS)	<.0001	<.0001	<.0001
15	1 (NS)	6 (HS)	<.0001	<.0001	<.0001

Bonferroni multiplies LSD p-value by 15, then you can compare with alpha 0.05 for significance.

Or you can perform Bonferroni adjustment in your head by using alpha=0.05/15 = 0.0033 for t-test or for LSD.

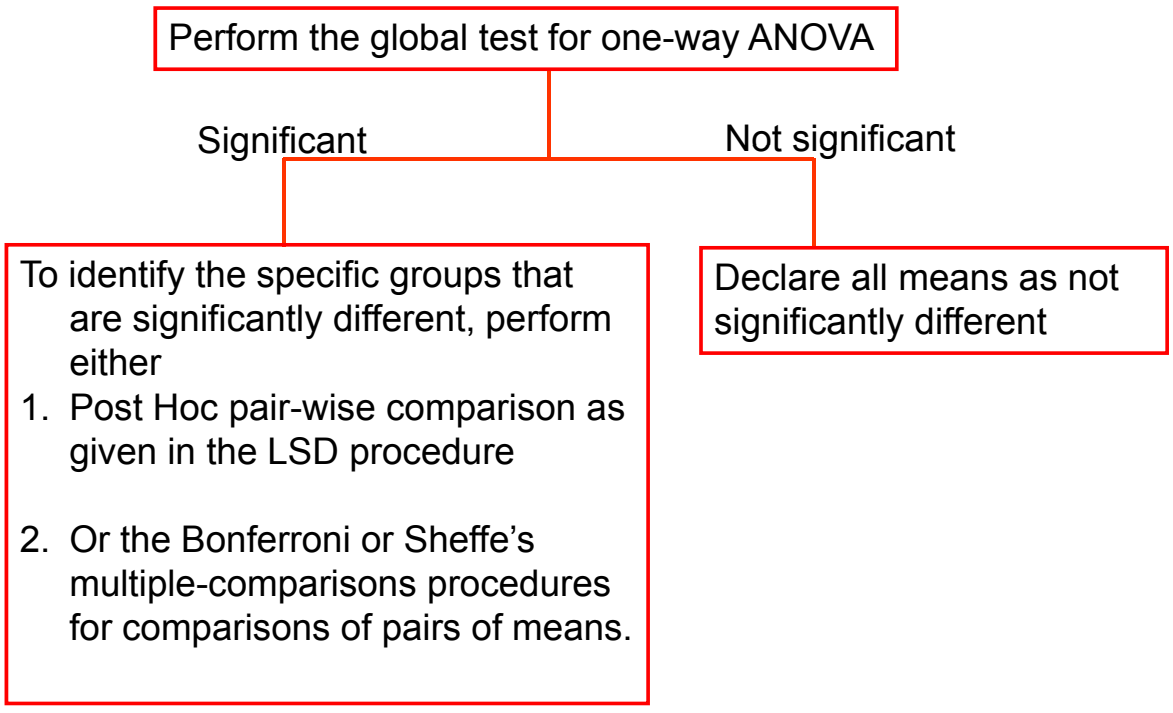
Recommendation to or not to control for multiple comparisons:

Adjusting for multiple comparisons are **highly controversial**, some people perform adjustment, others do not.

Multiple-comparisons procedures should be used if there are many groups and not all comparisons between individual groups have been planned in advance.

However if there are **relatively few groups** which have been **planned in advance** (stated in a protocol), and the global test for equivalence of means is significant then prefers to use ordinary t-test (i.e., LSD procedure) rather than adjusting p-value (or alpha level) for multiple-comparisons (i.e., not using Bonferroni). (Rosner, Fundamentals of Biostatistics, Duxburry Press)

General procedure for comparing the means of k independent, normally distributed samples



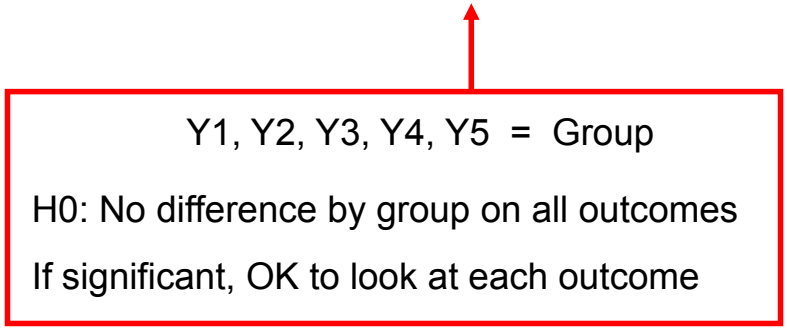
Bonferroni adjustment for multiple outcomes within the same study????

In Stephen Senn's book "Statistical Issues in Drug Development" (Senn S, statistical issues in Drug Development, John Wiley & Sons Ltd, Chichester, England),on page 143
" It can be claimed that, if all tests conducted are reported not only significant but non-significant results, then there should be no problem" (Note that even on this viewpoint, selectively reporting those tests which are significant, whilst ignoring the others, does cause a bias, However if all tests which are to be performed are reported with the order stated in the trial protocol.)

On page 144 -145,
"In general, the probability of making at least on type I error depends upon correlations between the outcomes. The Bonferroni correction is rather pessimistic and will be conservative where as may usually be expected to be the case, clinical outcomes are positively correlated."

How to avoid inflation of Type I error with multiple outcome variables.

1. Select a fewer number of outcomes to use in analysis.
2. Summarize multiple outcomes into one measures, such as using average score.
3. Using a global test to test multiple outcomes simultaneously such as multivariate analysis of variance (**MANOVA**) test.

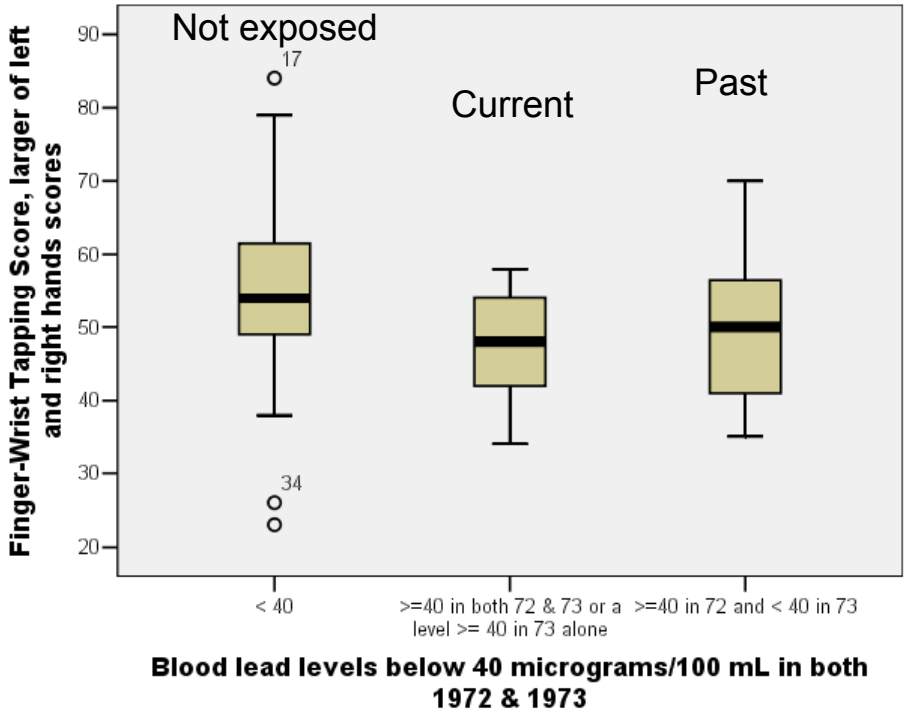


10.4 Performing One-way Analysis of Variance (ANOVA) in SPSS

Example 2 (Rosner page 582): Effects of lead exposure on Neurological and psychological function in children.

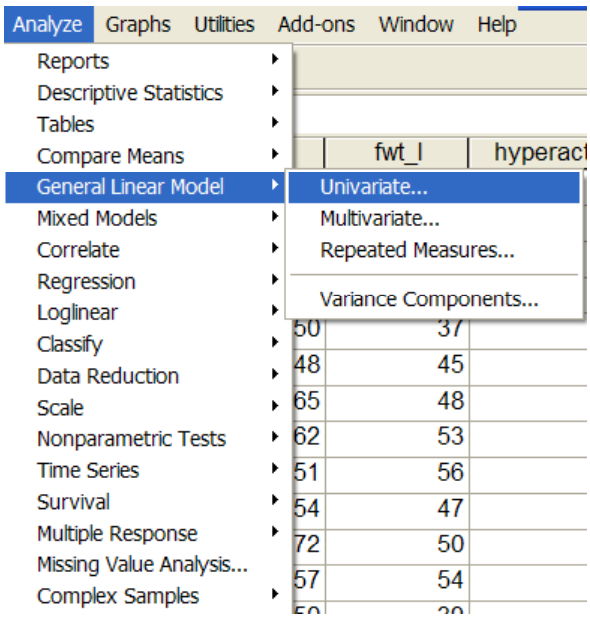
A group of children who lived near a lead smelter in El Paso, Texas, were identified and their blood levels of lead were measured. An exposed group of 46 children were identified who had blood-lead levels ≥40 mcg/mL in 1972 or 1973. A control group of 78 children was also identified who had blood-lead levels < 40 mcg/mL. Two important outcome variables that were studies were (1) the number of finger-wrist taps in the dominant hand and (2) the Wechsler full-scale IQ score.

Descriptive Analysis using Explore:



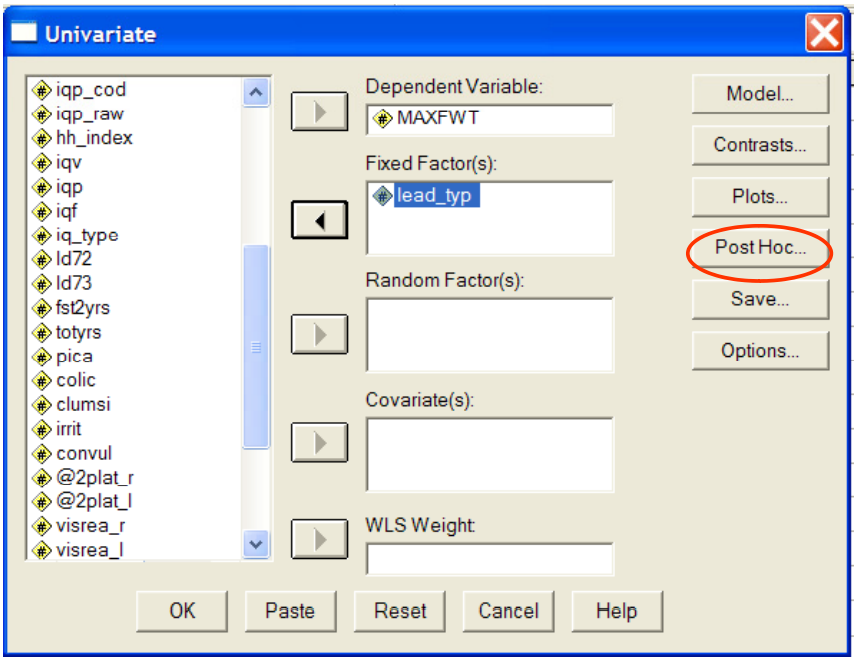
Perform one-way ANOVA (1)

Analyze
General linear model
Univariate
* Multivariate means when you have more than 1 dependent variables.



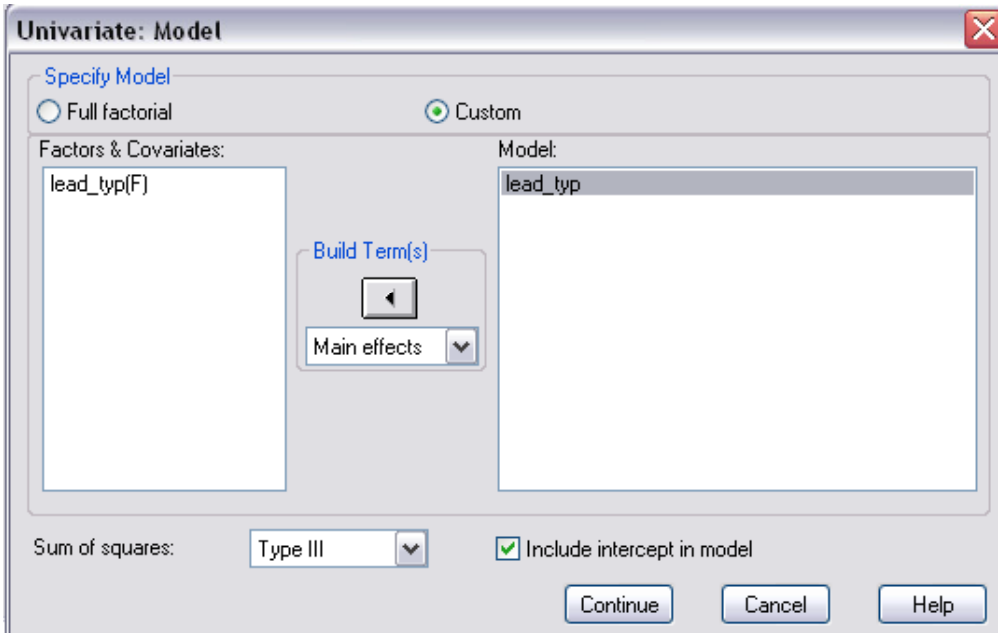
Perform one-way ANOVA (2)

In Univariate GLM dialog box,
Select MAXFWT as dependent, Lead_typ as Fixed Factor variables



Perform one-way ANOVA (4)

Click Models
Select Custom
Select lead_type into Model box, Continue
OK



SPSS outputs of One-way ANOVA (1)

Tests of Between-Subjects Effects

Dependent Variable: Finger-Wrist Tapping Score, larger of left and right hands scores

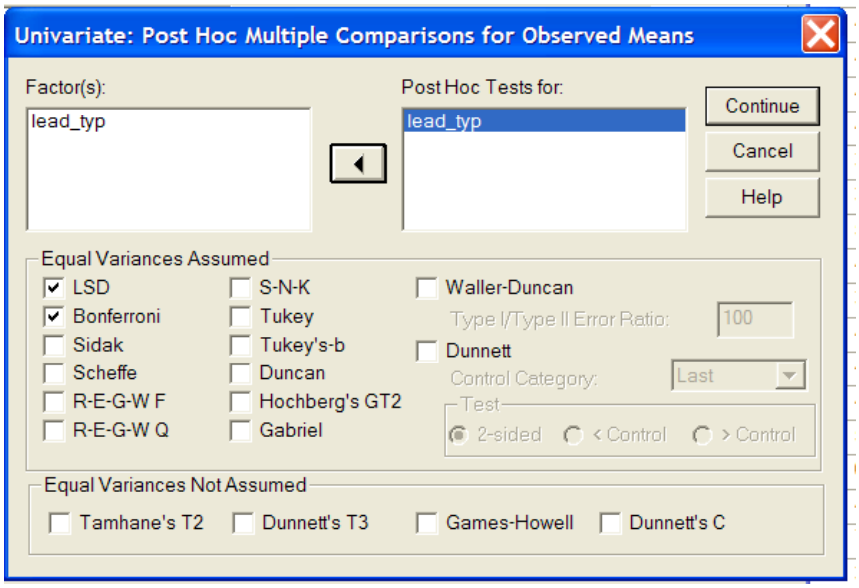
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	966.791 a	2	483.395	4.598	.012
Intercept	163616.705	1	163616.705	1556.458	.000
lead_typ	966.791	2	483.395	4.598	.012
Error	9671.146	92	105.121		
Total	276011.000	95			
Corrected Total	10637.937	94			

a. R Squared = .091 (Adjusted R Squared = .071)

The global test for the equality of the means indicating that there is a statistically significant association between type of lead exposure and finger-wrist tapping score with p=0.012.

Perform one-way ANOVA (5)

Click Post Hoc…
Select Lead_typ as Post Hoc Test for
Select LSD and Bonferroni
under the box of Equal Variances Assumed
Continue



SPSS outputs of One-way ANOVA (3)

Multiple Comparisons

Dependent Variable: Finger-Wrist Tapping Score, larger of left and right hands scores

	(I) Lead Exposure	(J) Lead Exposure	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	No Exposure	Current Exposure	7.51*	2.802	.009	1.94	13.07
		Past Exposure	5.70	2.946	.056	-.16	11.55
	Current Exposure	No Exposure	-7.51*	2.802	.009	-13.07	-1.94
		Past Exposure	-1.81	3.632	.619	-9.03	5.40
	Past Exposure	No Exposure	-5.70	2.946	.056	-11.55	.16
		Current Exposure	1.81	3.632	.619	-5.40	9.03
Bonferroni	No Exposure	Current Exposure	7.51*	2.802	.026	.67	14.34
		Past Exposure	5.70	2.946	.169	-1.49	12.88
	Current Exposure	No Exposure	-7.51*	2.802	.026	-14.34	-.67
		Past Exposure	-1.81	3.632	1.000	-10.67	7.05
	Past Exposure	No Exposure	-5.70	2.946	.169	-12.88	1.49
		Current Exposure	1.81	3.632	1.000	-7.05	10.67

Based on observed means.

*. The mean difference is significant at the .05 level.

Note: Bonferroni's p-values are 3 times larger than LSD p-values
Dunnet's p-values are 2 times larger than LSD p-values

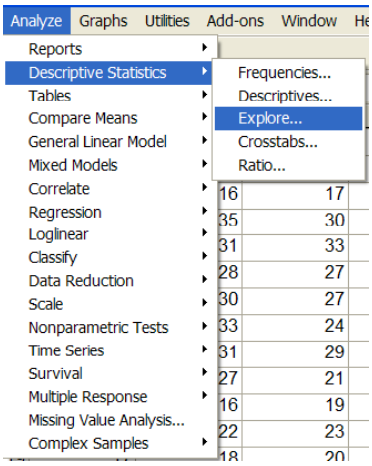
Interpretation of the results of the One-way ANOVA analysis

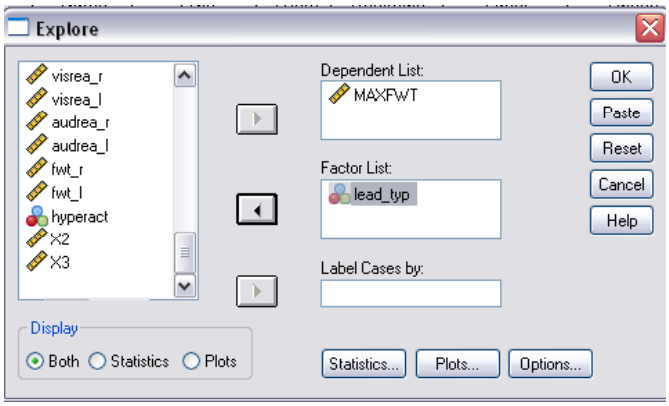
We see there is an overall significant difference among the mean MAXFWT scores in the three groups. The p-value is 0.012. Therefore, we proceeded to look at differences between each pair of groups. We did not adjust for p-value for multiple comparisons because ANOVA test was significant and the number of pair-wise comparisons were relatively small. We see that there is a significant difference between the mean MAXFWT score for the currently exposed group and the control group (p=0.009) with mean difference being 7.567, 95% CI = (1.9, 13.1). There is a strong trend toward a significant difference between the previously exposed group and the control group (p=0.056) with mean difference being 5.70, 95% CI = (-0.2, 11.5). There is clear no significant difference between the mean MAXFWT scores for the currently and previously exposed groups (p-value=0.619) with mean difference being 1.81 95% CI = (-9.0, 5.4).

10.5 Assumption of One-way Analysis of Variance (ANOVA)

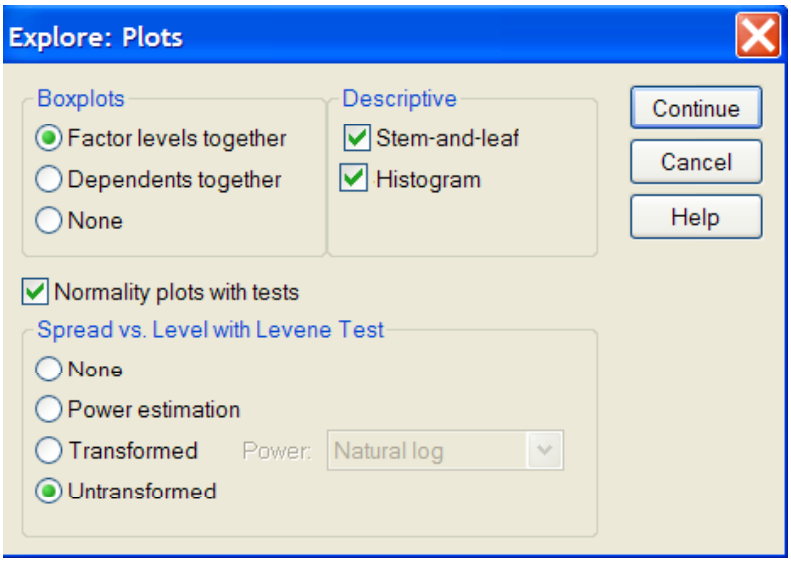
1. Observations are independent to each other.
2. Outcome variables are normally distributed within each group. [Spapiro-Wilk test]
3. Variance of outcome variables are the same across groups. [Levene test]

Testing for normality of residuals using Explore option in SPSS (1)

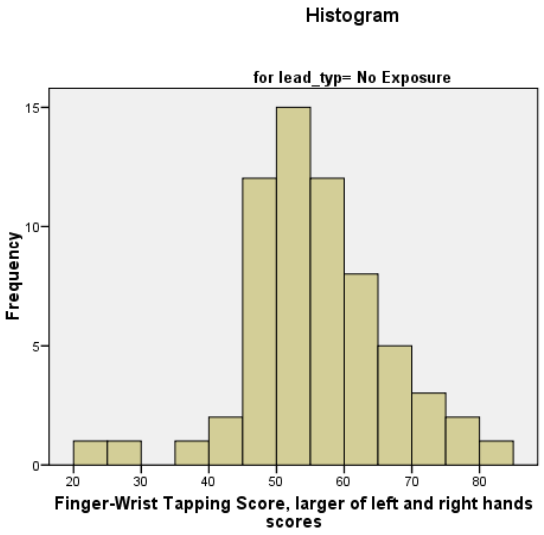


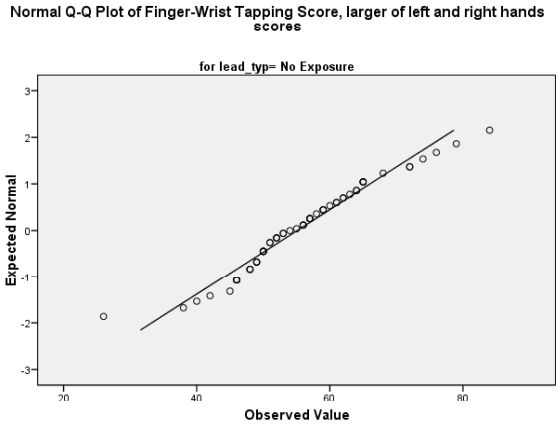


Testing for normality of residuals using Explore option in SPSS (2)

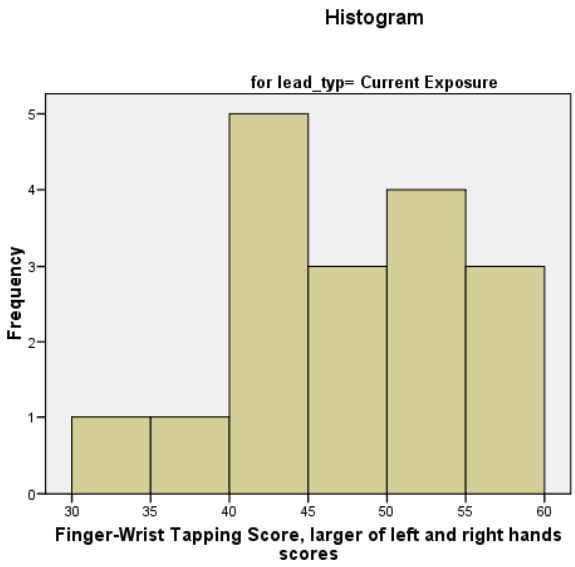


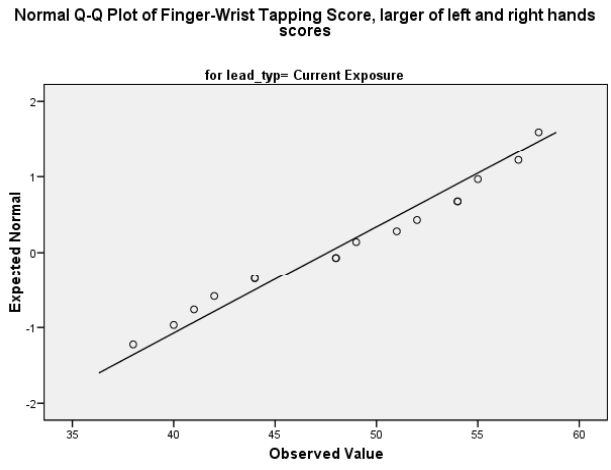
No Exposure Group



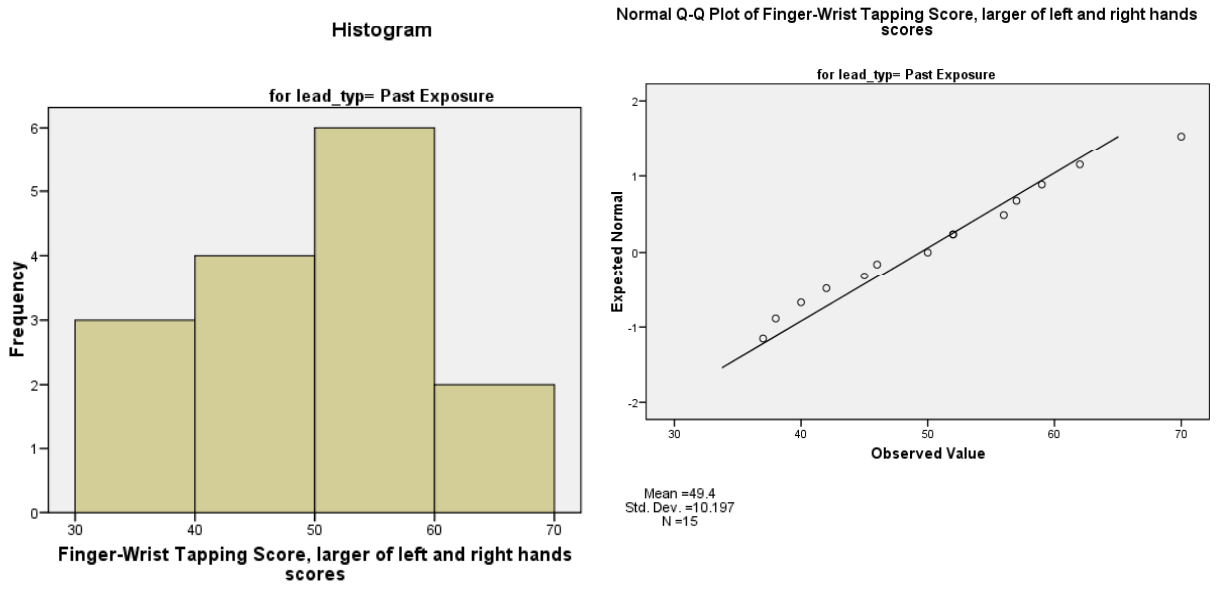


Current Exposure Group





Past Exposure Group



Testing for normality of outcome variable

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Lead Exposure	Statistic	df	Sig.	Statistic	df	Sig.
Finger-Wrist Tapping Score, larger of left and right hands scores	No Exposure	.108	63	.068	.966	63	.081
	Current Exposure	.112	17	.200*	.963	17	.692
	Past Exposure	.099	15	.200*	.966	15	.791

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

a. Lilliefors Significance Correction

Test of homogeneity of variances: Levene's test

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Finger-Wrist Tapping Score, larger of left and right hands scores	Based on Mean	.980	2	92	.379
	Based on Median	.986	2	92	.377
	Based on Median and with adjusted df	.986	2	81.751	.378
	Based on trimmed mean	1.013	2	92	.367

P> 0.05 failed to detect difference in variances, i.e, variances may be homogeneous.

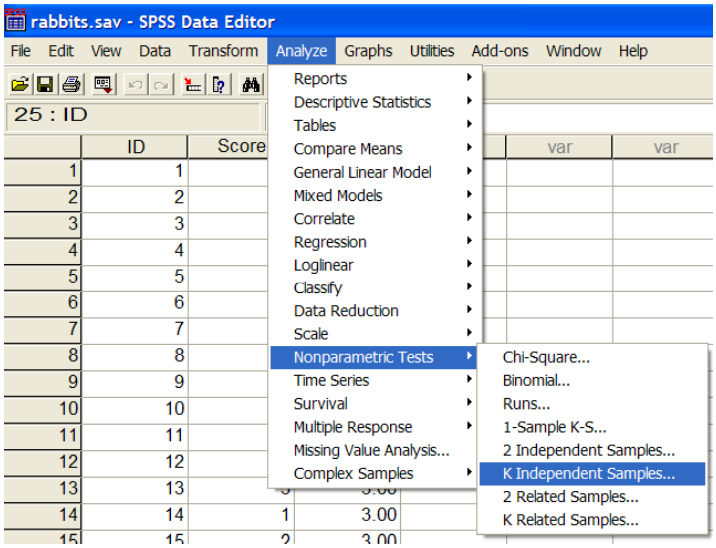
10.6 Non-parametric test corresponding One-Way ANOVA: Kruskal-Wallis Test

When assumption of one-way ANOVA (normality of outcome variable, homogeneity of variances) is not met, you may want to perform non-parametric tests.

As we learned previously in the chapter of comparing two means, Non-parametric tests are popular choice when you have variables which are believed non-normally distributed by nature, such as test scores, ranks.

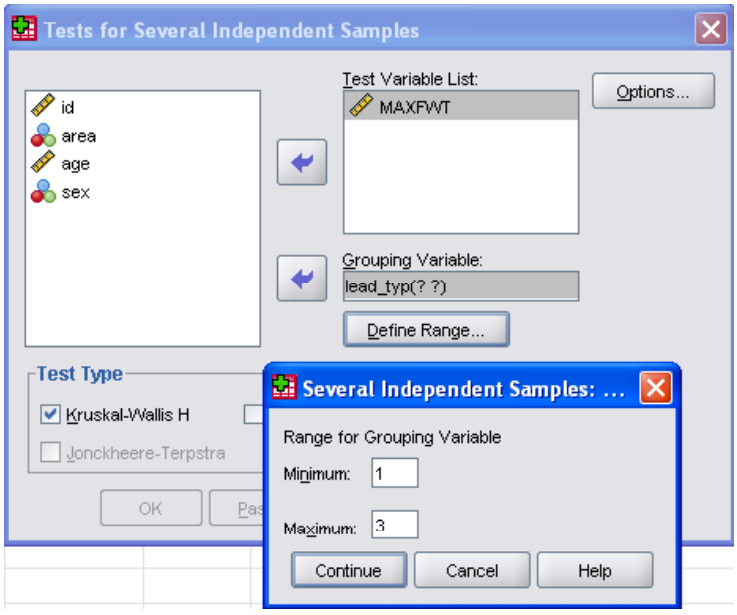
Performing Kruskal-Wallis test in SPSS (2)

Go to: Analyze, Nonparametric test, K independent samples

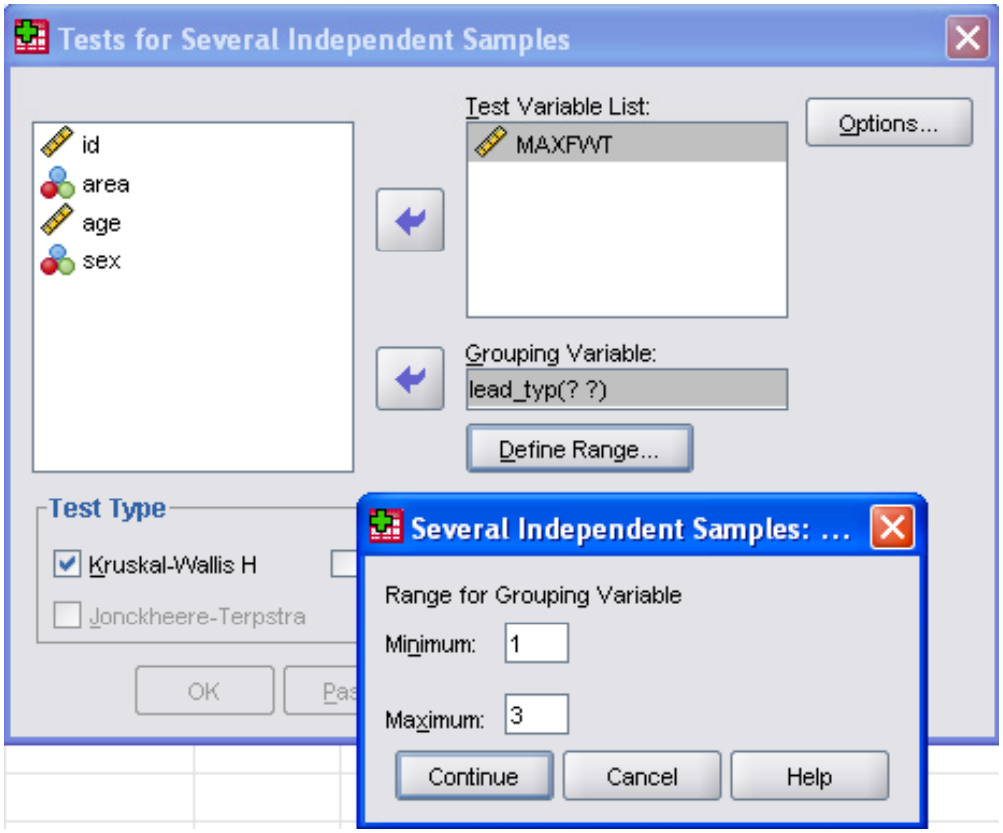


Performing Kruskal-Wallis test in SPSS (3)

In Test for Several Independent Samples box,
Select Score as dependent, Group as Grouping variable
Click on Define Range, and set 1 to 4 to minumun and maximum range.



Results of Kruskal-Wallis Test in SPSS



Test Statistics[a,b]

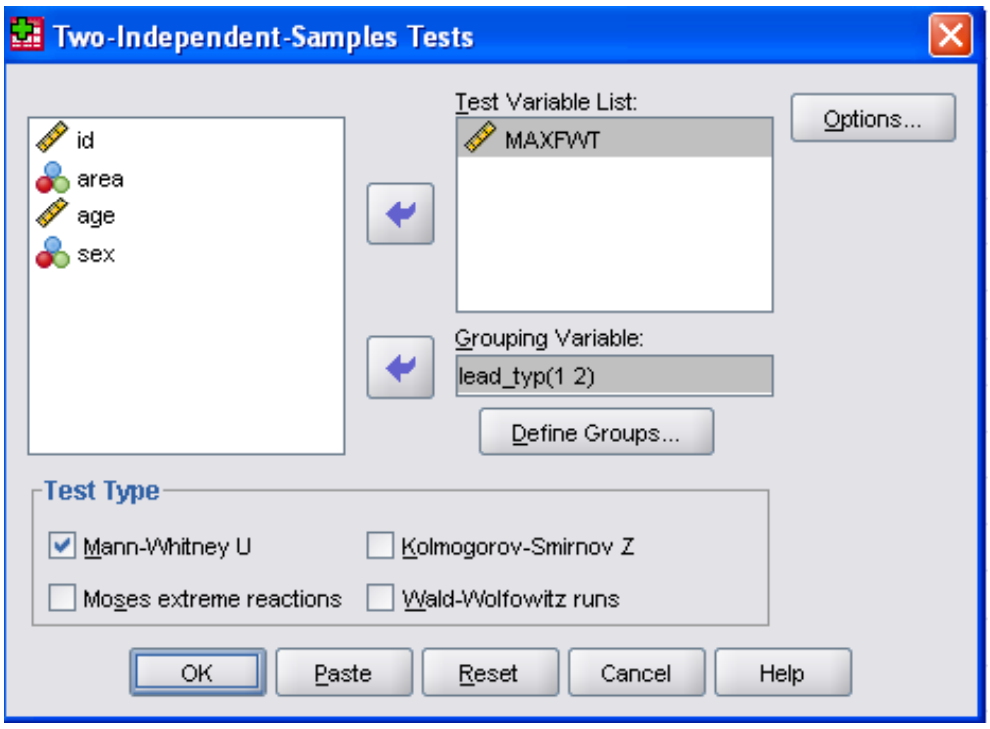
	Finger-Wrist Tapping Score, larger of left and right hands scores
Chi-Square	9.824
df	2
Asymp. Sig.	.007

a. Kruskal Wallis Test

b. Grouping Variable: Lead Exposure

P<0.05 thus this indicates that at least one group is different from others.

Non-parametric Pairwise comparisons with Man-whitney U tests



Results of Pairwise comparisons: with non-parametric approach

What test???

Groups		p-value
No exposure(1)	Current(2)	
No exposure(1)	Past(3)	
Current(3)	Past(2)	

Multiple comparison adjustment, Reject if p<0.016

Without multiple comparison adjustment, Reject if P<0.05

General procedure for comparing the means of k independent, non-normally distributed samples