

SPSS 3: COMPARING MEANS



UNIVERSITY OF GUELPH

LUCIA COSTANZO

lcostanz@uoguelph.ca

REVISED SEPTEMBER 2012

CONTENTS

SPSS availability	2
Goals of the workshop.....	2
Data for SPSS Sessions	3
Statistical Refresher	5
Types of Statistics	5
Types of Variables.....	6
Appropriate Statistics	6
Parametric Tests Versus Non-Parametric tests	7
Related versus Independent Samples	7
Choosing Statistical Test	8
Comparing two groups	8
Comparing more than two samples	8
T-test.....	9
Independent samples t-test.....	9
Paired samples t-test.....	13
One samples t-test.....	15
Analysis of Variance (ANOVA)	17

SPSS AVAILABILITY

Faculty, staff and students at the University of Guelph may access SPSS two different ways:

- 1. Library computers**

On the library computers, SPSS is installed on all machines.

- 2. Acquire a copy for your own computer**

If you are faculty, staff or a student at the University of Guelph, you may obtain the site-licensed standalone copy of SPSS at a cost. A free concurrent copy of SPSS is available to faculty, staff or graduate students at the University of Guelph. However, it may only be used while you are employed or a registered student at the University of Guelph. To obtain a copy, go to the CCS Software Distribution Site (www.uoguelph.ca/ccs/download).

GOALS OF THE WORKSHOP

This workshop builds on the skills and knowledge developed in “Getting your data into SPSS”. Participants are expected to have basic SPSS skills and statistical knowledge. Specific goals of this workshop are:

- Review statistical concepts: data types and associated tests
- Decide which test is the most appropriate one to use based on the study and data being analyzed
- Interpret three different t-tests analyses, the one-way ANOVA and the non-parametric equivalents
- Pull together what you learned about Normality, transformations and ANOVA to conduct an analysis and report results.

DATA FOR SPSS SESSIONS

DATASET: CANADIAN TOBACCO USE MONITORING SURVEY 2010 – PERSON FILE

This survey tracks changes in smoking status, especially for populations most at risk such as the 15- to 24-year-olds. It allows Health Canada to estimate smoking prevalence for the 15- to 24-year-old and the 25-and-older groups by province and by gender on a semi-annual basis.

The sample data used for this series of SAS workshops only includes respondents from the province of Quebec and only 14 of a possible 202 variables are being used.

To view the data, open the Excel spreadsheet entitled CTUMS_2010.xls

Variable Name	Label for Variable
PUMFID	Individual identification number
PROV	Province of the respondent
DVURBAN	Characteristic of the community
HHSIZE	Number of people in the household
HS_Q20	Number of people that smoke inside the house
DVAGE	Age of respondent
SEX	Respondent's sex
DVMARST	Grouped marital status of respondent
PS_Q30	Age smoked first cigarette
PS_Q40	Age begin smoking cigarettes daily
WP_Q10A	Number of cigarettes smoked – Monday
WP_Q10B	Number of cigarettes smoked – Tuesday
WP_Q10C	Number of cigarettes smoked – Wednesday
WP_Q10D	Number of cigarettes smoked – Thursday
WP_Q10E	Number of cigarettes smoked – Friday
WP_Q10F	Number of cigarettes smoked – Saturday
WP_Q10G	Number of cigarettes smoked – Sunday
SC_Q100	What was the main reason you began to smoke again?
WTPP	Person weight (survey weight variable)

Variable PROV : Province of the respondent

Values	Categories
10	N.L.
11	P.E.I.
12	Nova Scotia
13	N.B.
24	Quebec
35	Ontario
46	Manitoba
47	Saskatchewan
48	Alberta
59	B.C.

Variable HHSIZE : # of people in the household

Values	Categories
1	
2	
3	
4	
5	5 or more

Variable SC_Q100 : What was the main reason you began to smoke again?

Values	Categories
1	To control body weight
2	Stress, need to relax or to calm down
3	Boredom
4	Addiction / habit
5	Lack of support or information
6	Going out more (bars, parties)
7	Increased availability
8	No reason / felt like it
9	Family or friends smoke
10	Other
96	Valid skip
97	Don't know
98	Refusal
99	Not stated

Variable DVURBAN : Characteristic of community

Values	Categories
1	Urban
2	Rural
9	Not stated

Variable DVMARST : Grouped marital status of respondent

Values	Categories
1	Common-law/Married
2	Widow/Divorced/Separated
3	Single
9	Not stated

Variable PS_Q30 : Age smoked first cigarette

Variable PS_Q40 : Age begin smoking cigarettes daily

Variable HS_Q20 : # of people that smoke inside the home

Variable WP_Q10A : # of cigarettes smoked-Monday

Variable WP_Q10B : # of cigarettes smoked-Tuesday

Variable WP_Q10C : # of cigarettes smoked-Wednesday

Variable WP_Q10D : # of cigarettes smoked-Thursday

Variable WP_Q10E : # of cigarettes smoked-Friday

Variable WP_Q10F : # of cigarettes smoked-Saturday

Variable WP_Q10G : # of cigarettes smoked-Sunday

Values	Categories
96	Valid skip
97	Don't know
98	Refusal
99	Not stated

STATISTICAL REFRESHER

TYPES OF STATISTICS

Two broad types of statistics exist which are descriptive and inferential. Descriptive statistics describe the basic characteristics of the data in a study. Usually generated through an Exploratory Data Analysis (EDA), they provide simple numerical and graphical summaries about the sample and measures. Inferential statistics allow you to make conclusions regarding the data i.e. significant difference, relationships between variables, etc.

Here are some examples of descriptives and inferential statistics:

Descriptives	Inferential
<ul style="list-style-type: none">• Frequencies• Means• Standard Deviations• Ranges• Medians• Modes	<ul style="list-style-type: none">• t-tests• Chi-squares• ANOVA• Friedman

Which test to perform on your data largely depends on a number of factors including:

1. What type of data you are working with?
2. Are you samples related or independent?
3. How many samples are you comparing?

TYPES OF VARIABLES

Variable types can be distinguished by various levels of measurement which are nominal, ordinal, interval or ratio.

NOMINAL

Have data values that identify group membership. The only comparisons that can be made between variable values are equality and inequality. Examples of nominal measurements include gender, race religious affiliation, telephone area codes or country of residence.

ORDINAL

Have data values arranged in a rank ordering with an unknown difference between adjacent values. Comparisons of greater and less can be made and in addition to equality and inequality. Examples include: results of a horse race, level of educations or satisfaction/attitude questions.

INTERVAL

Are measured on a scale that a one-unit change represents the same difference throughout the scale. These variables do not have true zero points. Examples include: temperature in the Celsius or Fahrenheit scale, year date in a calendar or IQ test results.

RATIO

Have the same properties as interval variables plus the additional property of a true zero. Examples include: temperature measured in Kelvins, most physical quantities such as mass, length or energy, age, length of residence in a given place.

Interval and Ratio will be considered identical thus yielding three types of measurement scales.

APPROPRIATE STATISTICS

For each type of variable, a particular measure of central tendency is most appropriate. By central tendency, we mean one value that most effectively summarizes a variable's complete distribution.

Measurement Scale	Measure of Central Tendency
Nominal	Mode – Value that appears the most often in distribution.
Ordinal	Median – Value that divides the ordered distribution of responses into two equal size groups. (the values of the 50 th percentile)
Interval/Scale	Mean – The arithmetic average of a distribution.

PARAMETRIC TESTS VERSUS NON-PARAMETRIC TESTS

Parametric test are “techniques especially those involving continuous distributions, have stressed the underlying assumptions for which the techniques are valid. These techniques are for the estimation of parameters and for testing hypotheses concerning them.” (Steele et al., 1997)

Non-parametric test “a considerable amount of data the underlying distribution is not easily specified. To handle such data, we need distribution-free statistics; that is, we need procedures that are not dependent on a specific parent distribution. If we do not specify the nature of the parent distribution, then we will not ordinarily deal with the parameters. Non-parameters statistics compare distributions rather than parameters”. (Steele et al., 1997)

RELATED VERSUS INDEPENDENT SAMPLES

RELATED SAMPLES

- Measures taken on the same individual or responses given by the same individual
- For example:
 - Scores provided by a panel of judges on several products
 - Survey responses
 - Measures taken pre and post treatment

INDEPENDENT SAMPLES

- Measures taken on a number of individuals
- For example:
 - Scores provided by people in a mall on a product
 - Measures on different treatments

CHOOSING STATISTICAL TEST

COMPARING TWO GROUPS

Type of Variable	Independent	Related
Nominal	Chi-Square	McNemar Test
Ordinal	Mann-Whitney Wilcoxon Rank Sums	Sign Test Wilcoxon Test
Internal/Ratio/Scale/Continuous	Independent sample T-test	Related Sample T-test

COMPARING MORE THAN TWO SAMPLES

Type of Variable	Independent	Related	2 or more factors	Correlation
Nominal	Chi-Square	Cochran Q-test		
Ordinal	Kruskal-Wallis	Friedman 2-way ANOVA	Friedman 2-way ANOVA	Spearman's Rank Correlation
Internal/Ratio/Scale/Continuous	2+ factor ANOVA	2+ factor ANOVA	2+ factor ANOVA	Pearson's Product Correlation

PARAMETRIC TESTS

T-TEST

T-tests are used to test the null hypothesis that two population means are equal – that there are no differences between the two populations. There are three types of t-tests:

1. Independent sample t-test
2. Paired sample t-test
3. One sample t-test

INDEPENDENT SAMPLES T-TEST

The independent samples t-test. It is also referred to as unpaired or unrelated samples t-test. It allows for us to compare the means observed for one variable for two independent samples. When running this parametric test, SPSS generates descriptive statistics for each group, a Levene's test for equality of variance. It also reports the equal and unequal variance t-values and the 95% confidence interval for difference in means.

RESEARCH QUESTION

We are interested in determining whether there is a difference between the age at which females and males first smoked.

NULL HYPOTHESIS:

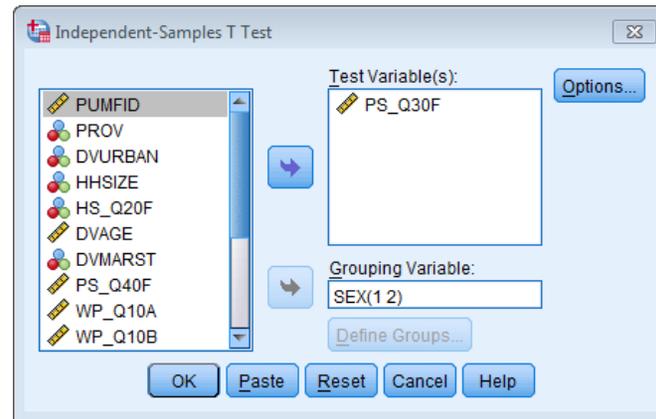
$$H_0 : \mu_{\text{females}} - \mu_{\text{males}} = 0$$

ALTERNATIVE HYPOTHESIS:

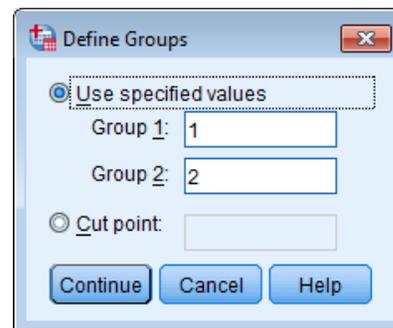
$$H_A : \mu_{\text{females}} - \mu_{\text{males}} \neq 0$$

STEPS ON RUNNING AN INDEPENDENT SAMPLES T-TEST

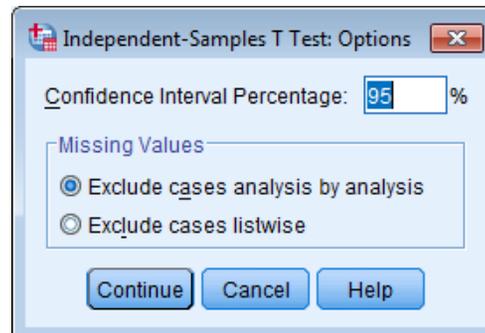
1. Click on Analyze>Compare Means>Independent Samples T-test. The following dialog box to the below appears.
2. The variables we are testing is the age first smoked a cigarette and sex is the grouping variable.



3. Clicking the Define Groups allows us to indicate which groups are to be compared. This can be done by actually entering in specific values for each group or specifying a cut-off point (cut point).



4. The Options buttons allows you to specify a value for the confidence interval.



RESULTS

Group Statistics

Respondent sex		N	Mean	Std. Deviation	Std. Error Mean
Age smoked first cigarette	Male	278	15.68	3.894	.234
	Female	262	15.88	4.442	.274

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age smoked first cigarette	Equal variances assumed	.005	.945	-.542	538	.588	-.194	.359	-.900	.511
	Equal variances not assumed			-.539	519.298	.590	-.194	.360	-.902	.514

Check this test first to determine which t-test to read.
 If the p-value is < 0.05 then the variances are not equal and you will need to read the unequal variances test results in the t-test table to the right.

Since Equal Variances are assumed, we will be looking at the first row. The p-value is greater than 0.05
 Looking at the Group Statistics table we can see that those people that undertook the exercise trial had lower cholesterol levels at the end of the programme than those that underwent a calorie-controlled diet.

Exercise

We are interested in determining whether there is a difference between the age at which females first smoked. Do we accept or reject the null hypothesis?

NULL HYPOTHESIS:

$$H_0 : \mu_{\text{females}} = \mu_{\text{males}}$$

ALTERNATIVE HYPOTHESIS:

$$H_A : \mu_{\text{females}} \neq \mu_{\text{males}}$$

PAIRED SAMPLES T-TEST

The paired samples t-test is also referred to as the dependent or related samples t-test. It is useful for testing if a significant difference occurs between the means of two variables that represent the same group at different times (before or after) or related groups (husband or wife). For example in medical research, a paired t-test is used to compare the means on a measure before (pre) and after (post) a treatment. Looking at market research, test could be used to compare the rating an individual gives a product they usually purchase and competing product on some characteristic.

RESEARCH QUESTION

We are interested in determining whether there is a difference between the age respondents first smoked and the age at which they began smoking cigarettes daily.

NULL HYPOTHESIS:

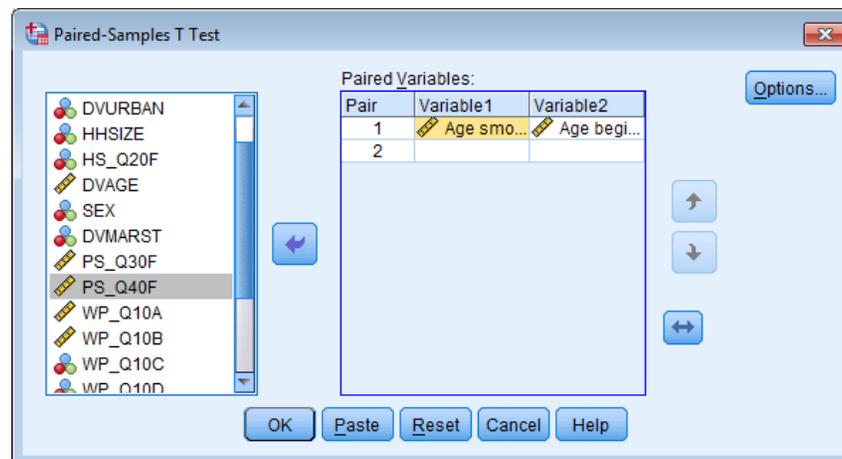
$$H_0 : \mu_{\text{firstsm}} - \mu_{\text{smokedly}} = 0$$

ALTERNATIVE HYPOTHESIS:

$$H_A : \mu_{\text{firstsm}} - \mu_{\text{smokedly}} \neq 0$$

STEPS ON RUNNING AN PAIRED SAMPLES T-TEST

1. Click on Analyze>Compare Means>Paired Samples T-test. The following dialog box to the below appears.
2. The pairs of variables we have chosen for the analysis is ps_q30 and ps_q40. Select these variables and move them to the Paired Variables window. Then click on the OK button to the right.



The Descriptives table displays the mean, sample size, standard deviation, and standard error for both groups.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Age smoked first cigarette	15.41	314	4.193	.237
	Age begin smoking cigarettes daily	17.28	314	4.513	.255

The Pearson correlation between the age smoked first cigarette and age began smoking cigarettes daily is 0.857, highly correlated and is significant.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Age smoked first cigarette & Age begin smoking cigarettes daily	314	.857	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Age smoked first cigarette - Age begin smoking cigarettes daily	-1.863	2.350	.133	-2.124	-1.602	-14.046	313	.000

This value (0.000) tells us we can reject the Null Hypothesis.

Since the significance value for change of cigarette smoking between the age first smoked and age began smoking daily is less than 0.05, you can conclude that the average time between first time smoked and age first began to smoke daily of 1.86 years is not due to chance variation and can be

ONE SAMPLES T-TEST

A one sample t-test procedure tests whether the mean of a single variable differs from a specified constant.

RESEARCH QUESTION

We are interested in determining whether the age that most people first smoked is 18 years of age.

NULL HYPOTHESIS:

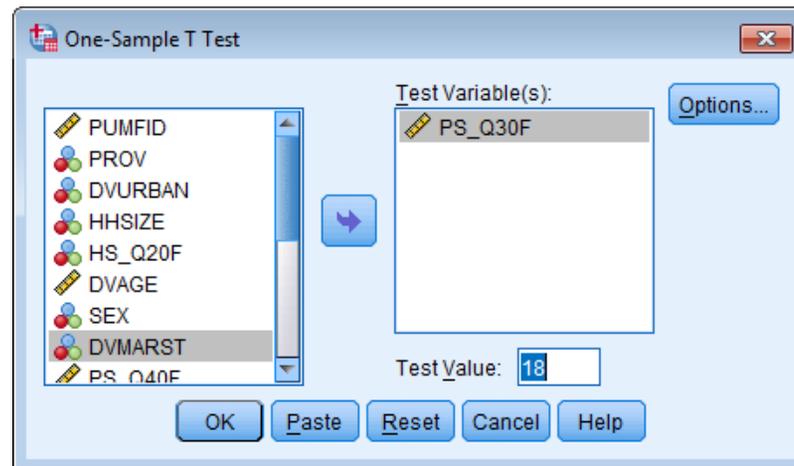
$$H_o : \mu_{\text{firstsm}} = 18$$

ALTERNATIVE HYPOTHESIS:

$$H_A : \mu_{\text{firstsm}} \neq 18$$

STEPS ON RUNNING A ONE-SAMPLE T-TEST

1. Click on the Analyze>Compare Means>One Sample test. The following dialog box in the right will appear.
2. The test variable is the PS_Q30. The test value for this case is 18.



T-

RESULTS

The Descriptives table displays the sample size, mean, standard deviation, and standard error.

	N	Mean	Std. Deviation	Std. Error Mean
Age smoked first cigarette	540	15.78	4.166	.179

	Test Value = 18					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Age smoked first cigarette	-12.395	539	.000	-2.222	-2.57	-1.87

Since their confidence intervals lie entirely below 0.0, you can safely say that age first smoked first cigarette is lower than 18 years of age on the average.

ANALYSIS OF VARIANCE (ANOVA)

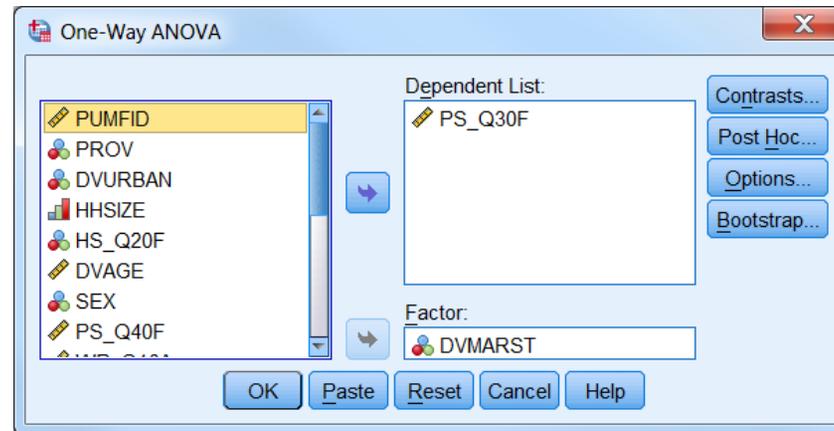
This procedure compares the means from several samples and tests whether they are all the same or whether one or more of them are significantly different. This is an extension of the t-test for datasets containing more than two samples.

RESEARCH QUESTION

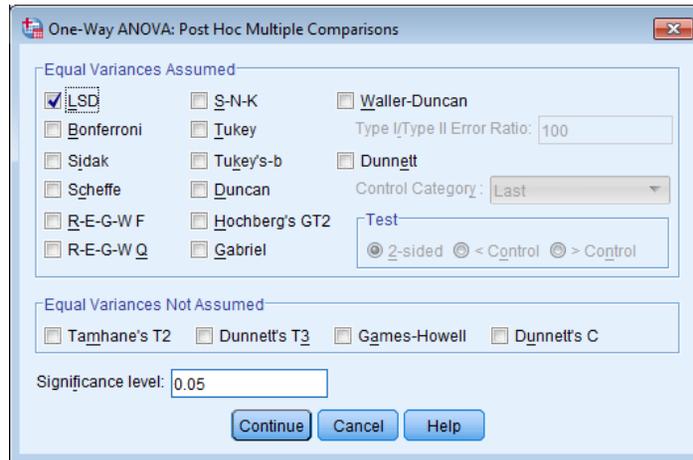
We are interested in determining whether there are differences between age that respondents first smoked and respondent age.

STEPS ON RUNNING AN ANALYSIS OF VARIANCE (ANOVA)

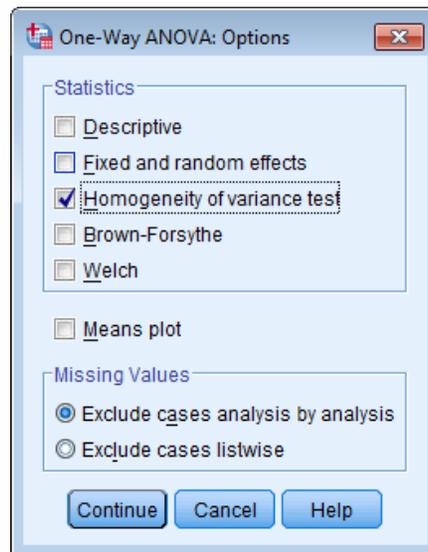
1. Click on Analyze>Compare Means>One-Way ANOVA. The following dialog box to the right will appear. The dependent variable is the PS_Q30F. For this example, the variable DVMARST is the factor which is the independent variable defining groups of cases.



2. Click on the Post Hoc button to test for pairwise multiple comparisons between means. Check the LSD checkbox and select Continue.



3. Click on the Options Button and the dialog box to the right will appear. Check off Homogeneity of Variance test and then click on the Continue button.



RESULTS

Test of Homogeneity of Variances

Age smoked first cigarette

Levene Statistic	df1	df2	Sig.
1.748	2	167	.177

An important first step in the analysis of variance is establishing the validity of assumptions. One assumption of ANOVA is that the variances of the groups are equivalent.

The Levene statistic accepts the null hypothesis that the group variances are equal. ANOVA is robust to this violation when the groups are of equal or near equal size; however, you may choose to transform the data or perform a nonparametric test that does not require this assumption.

ANOVA

Age smoked first cigarette

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	162.063	2	81.031	4.953	.008
Within Groups	2731.990	167	16.359		
Total	2894.053	169			

The significance value of the F test in the ANOVA table is less than 0.05. Thus, you must fail to accept the hypothesis that average age first smoked cigarette are equal across the different marital status groups.

Multiple Comparisons

Dependent Variable: Age smoked first cigarette

LSD

(I) Grouped marital status of respondent	(J) Grouped marital status of respondent	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Common-law/Married	Widow/Divorced/Separated	1.142	1.237	.357	-1.30	3.59
	Single	2.089*	.664	.002	.78	3.40
Widow/Divorced/Separated	Common-law/Married	-1.142	1.237	.357	-3.59	1.30
	Single	.947	1.195	.429	-1.41	3.31
Single	Common-law/Married	-2.089*	.664	.002	-3.40	-.78
	Widow/Divorced/Separated	-.947	1.195	.429	-3.31	1.41

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

NON-PARAMETRIC TESTS

KRUSKAL-WALLIS TEST

The Kruskal-Wallis Test is the nonparametric test equivalent to the one-way ANOVA and an extension of the Mann-Whitney Test to allow the comparison of more than two independent groups. It is used when we wish to compare three or more sets of scores that come from different groups.

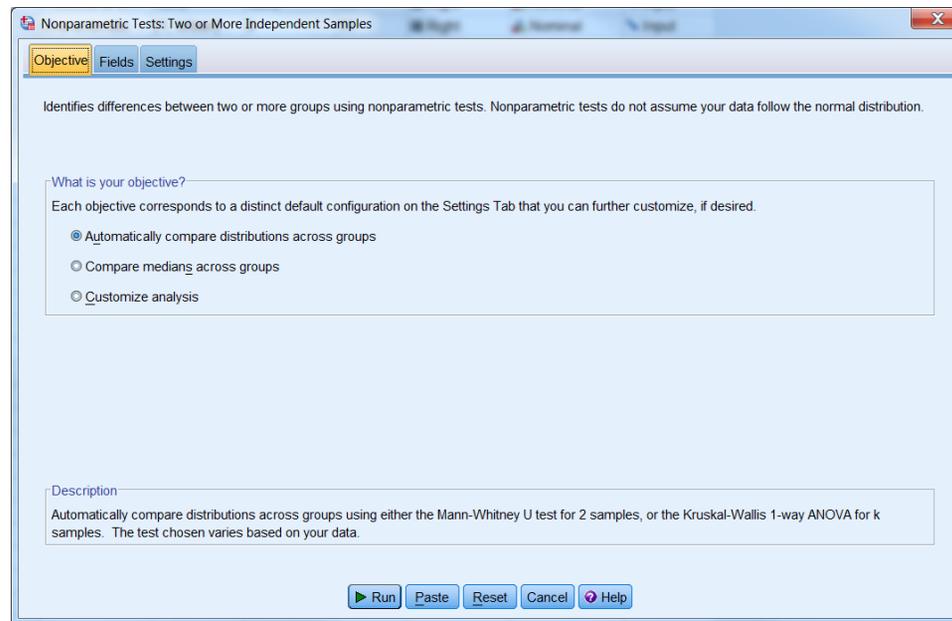
As the Kruskal-Wallis Test does not assume normality in the data and is much less sensitive to outliers it can be used when these assumptions have been violated and the use of the one-way ANOVA is inappropriate. In addition, if your data is ordinal then you cannot use a one-way ANOVA but you can use this test.

RESEARCH QUESTION

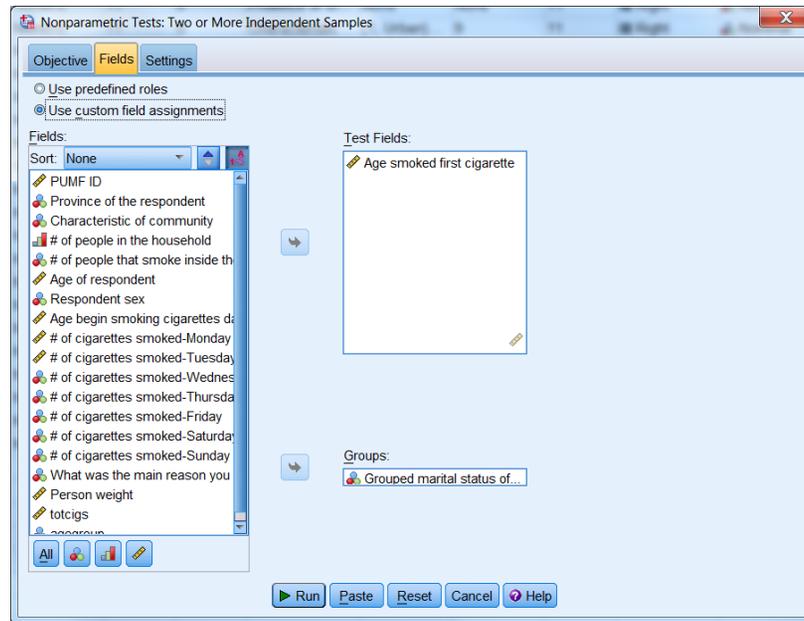
We are interested in determining whether there is a whether there are differences between age that respondents first smoked and respondent age.

STEPS ON RUNNING AN KRUSKAL-WALLIS TEST

1. Click on Analyze>Nonparametric Test>Independent Samples. The following dialog box to the below appears.



2. Click on the Fields Tab and move "Age smoked first cigarette" to Test Fields then move "agegroup" to Groups. Click Run.



RESULTS

Hypothesis Test Summary

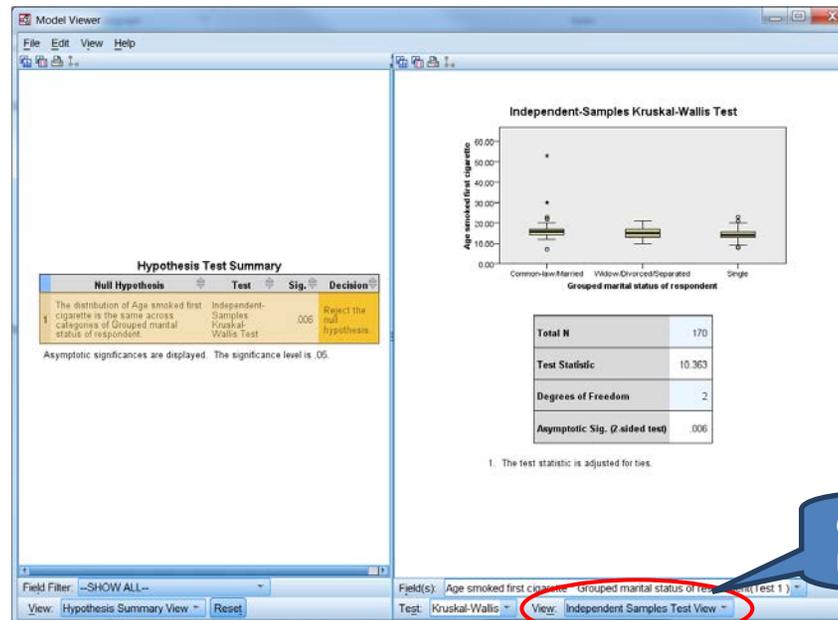
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Age smoked first cigarette is the same across categories of Grouped marital status of respondent.	Independent-Samples Kruskal-Wallis Test	.006	Reject the null hypothesis.

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

The test determines age first smoked cigarette are significantly different with a low "Sig." value of 0.005. This value simply states that the probability that the noted differences are attributable to chance alone is only 0.005.

The null hypothesis has been rejected, that the age first smoked cigarette is same for all age categories listed in agegroup.

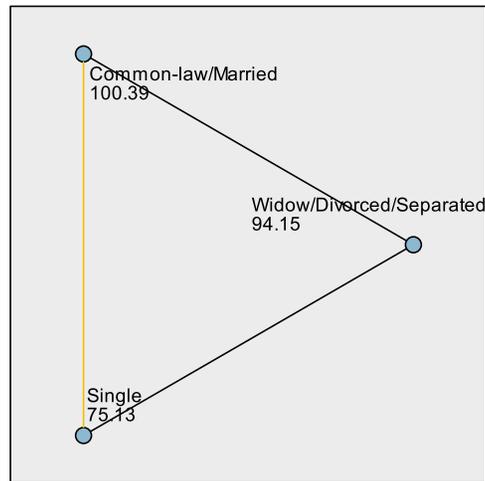
In the output window, double click on the Hypothesis Test Summary and the following appears.



Click here first and choose Pairwise Comparisons.

Click the arrow in the View line then click Pairwise Comparisons.

Pairwise Comparisons of Grouped marital status of respondent



Each node shows the sample average rank of Grouped marital status of respondent.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Single-Widow/Divorced/Separated	19.025	14.423	1.319	.187	.561
Single-Common-law/Married	25.263	8.020	3.150	.002	.005
Widow/Divorced/Separated-Common-law/Married	6.238	14.939	.418	.676	1.000

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.