

# SPSS program

2022

## What is SPSS?

SPSS means “Statistical Package for the Social Sciences”. SPSS is software for managing data and calculating a wide variety of statistics, and analyzing all sorts of data. SPSS can open all file formats that are commonly used for structured data such as

- ① spreadsheets from MS Excel or OpenOffice;
- ② plain text files (.txt or .csv);
- ③ relational (SQL) databases;
- ④ Stata and SAS.

## The SPSS Windows and Files

SPSS Statistics has three main windows, plus a menu bar at the top. The windows are

- ① Data Editor (.sav files):
- ② Output Viewer (.spv files)
- ③ Syntax Editor (.sps files)

# 1. Data Editor Window

a. Data View: After opening data, SPSS displays them in a spreadsheet. It is called data view. Changes you make to your data are not permanent until you save them (click File - Save or Save As). Data files are saved with a file type of .sav.

\*Employee data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Custom Utilities Add-ons Window Help

Visible: 10 of 10 Variables

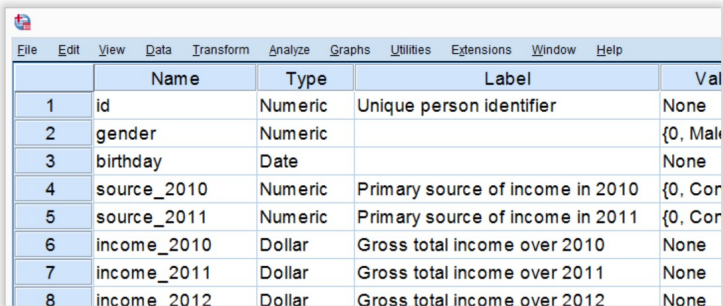
	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	p
1	1	Male	02/03/1952	15	Manager	\$57,000	\$27,000	98	
2	2	Male	05/23/1958	16	Clerical	\$40,200	\$18,750	98	
3	3	Female	07/26/1929	12	Clerical	\$21,450	\$12,000	98	
4	4	Female	04/15/1947	8	Clerical	\$21,900	\$13,200	98	
5	5	Male	02/09/1955	15	Clerical	\$45,000	\$21,000	98	
6	6	Male	08/22/1958	15	Clerical	\$32,100	\$13,500	98	
7	7	Male	04/26/1956	15	Clerical	\$36,000	\$18,750	98	
8	8	Female	05/06/1966	12	Clerical	\$21,900	\$9,750	98	
9	9	Female	01/23/1946	15	Clerical	\$27,900	\$12,750	98	
10	10	Female	02/13/1946	12	Clerical	\$24,000	\$13,500	98	
11	11	Female	02/07/1950	16	Clerical	\$30,300	\$16,500	98	
12	12	Male	01/11/1966	8	Clerical	\$28,350	\$12,000	98	
13	13	Male	07/17/1960	15	Clerical	\$27,750	\$14,250	98	
14	14	Female	02/26/1949	15	Clerical	\$35,100	\$16,800	98	

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode ON

# 1. Data Editor Window

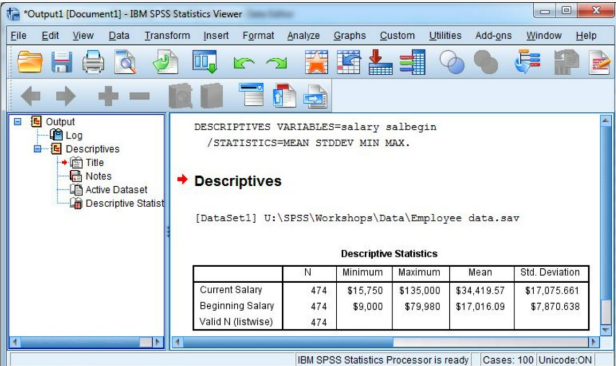
b. Variable View: It shows the meaning of variables, data types, and data values.



	Name	Type	Label	Val
1	id	Numeric	Unique person identifier	None
2	gender	Numeric		{0, Male
3	birthday	Date		None
4	source_2010	Numeric	Primary source of income in 2010	{0, Cor
5	source_2011	Numeric	Primary source of income in 2011	{0, Cor
6	income_2010	Dollar	Gross total income over 2010	None
7	income_2011	Dollar	Gross total income over 2011	None
8	income_2012	Dollar	Gross total income over 2012	None

## 2. Output Viewer Window

Statistical results will show up in the Output Viewer. The Output Viewer shows you tables of statistical output and any graphs you create. By default it also show you the programming language for the commands that you issued (called syntax in SPSS jargon), and most error messages will also appear here. The Output Viewer also allows you to edit and print your results. The tables of the Output Viewer are saved (click File - Save or Save As) with a file type of .spv, which can only be opened with SPSS software.



The screenshot shows the IBM SPSS Statistics Viewer window. The left pane displays a tree view of the output, with 'Descriptives' selected. The right pane shows the syntax command and the resulting descriptive statistics table.

```
DESCRIPTIVES VARIABLES=salary salbegin
/STATISTICS=MEAN STDDEV MIN MAX.
```

**Descriptives**

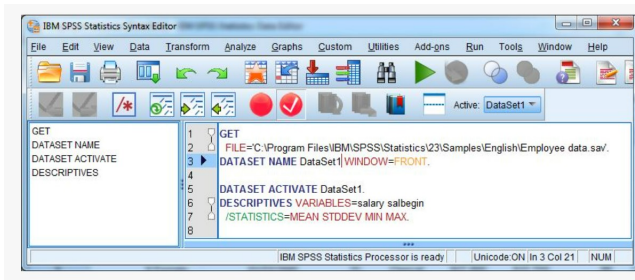
[DataSet1] U:\SPSS\Workshops\Data\Employee data.sav

	N	Minimum	Maximum	Mean	Std. Deviation
Current Salary	474	\$15,750	\$135,000	\$34,419.57	\$17,075.661
Beginning Salary	474	\$9,000	\$79,980	\$17,016.09	\$7,870.638
Valid N (listwise)	474				

IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode: ON

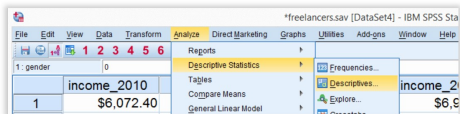
### 3. Syntax Editor Window

The Syntax Editor allows you to write, edit, and run commands in the SPSS programming language. If you are also using the menus and dialog boxes, the Paste button automatically writes the syntax for the command you have specified into the active Syntax Editor. These files are saved as plain text and almost any text editor can open them, but with a file extension of .sps.

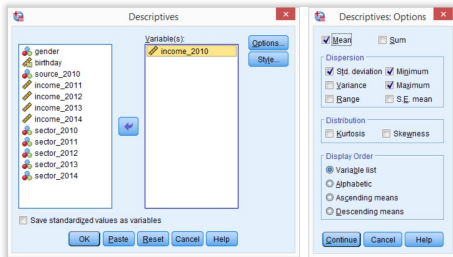


## Descriptive Statistics for quantitative variables

The data in freelancers.sav file contain a variable holding respondents' incomes over 2010, we can compute the average income by navigating to Descriptive Statistics as shown below

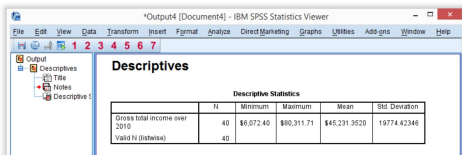


Doing so opens a dialog box in which we select one or many variables and one or several statistics we'd like to inspect.





After clicking Ok, a new window opens up: SPSS' output viewer window. It shows a table with all statistics on all variables we chose.

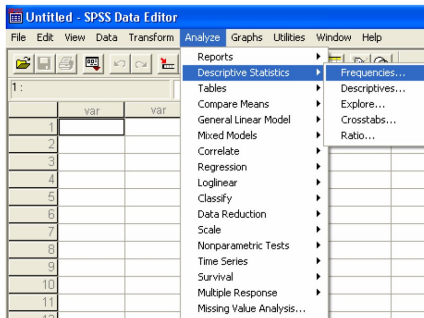


The screenshot shows the 'IBM SPSS Statistics Viewer' window titled '\*Output4 [Document4]'. The window has a menu bar (File, Edit, View, Data, Transform, Insert, Format, Analyze, Direct Marketing, Graphs, Utilities, Add-ons, Window, Help) and a toolbar with icons for navigation and document management. On the left, a tree view shows the hierarchy: Output > Descriptives > Title > Notes > Descriptive Statistics. The main area displays the 'Descriptives' table.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Gross total income over 2010	40	\$6,072.40	\$80,311.71	\$45,231.3520	19774.42346
Valid N (listwise)	40				

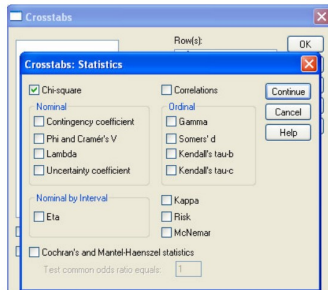
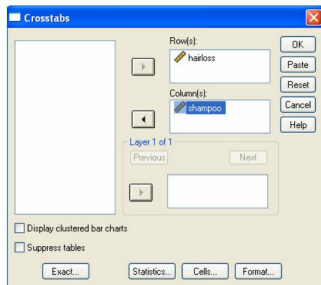
## Descriptive Statistics for qualitative variables

Analyze → Descriptive Statistics → Frequencies



## The Chi-Square test of association between two independent variables output:

Suppose we want to see if there is an association between brand of antidandruff shampoo ("Noflakes" and "Head and Shudders") and hair loss (totally bald versus no hair loss). In this case, we would have two columns. One would give the brand of shampoo that a participant used (coded 1 for "Noflakes" or "2" for "Head and Shudders") and the other would give the same participant's state of hairiness (coded with "1" for "bald" or "2" for "full head of hair"). The data is in excel file that is named "chi square data spss.xlsx".  $H_0$ : the hairloss and shampoo brand variables are independent vs  $H_1$ : not independent. To perform the Chi-Square analysis, go to "**Analyze**" and pick "**DescriptiveStatistics**" and then "**Crosstabs**". The output shows us that 17 people used "Noflakes" and 13 used "Head and Shudders". It also shows us the observed frequencies (how many users of each shampoo actually were bald and how many were actually hairy) and the expected frequencies (how many bald and hairy users of each shampoo we would expect to get if baldness and shampoo choice had nothing to do with each other). Hopefully you can see that the observed and expected frequencies are rather different from each other.



**Crosstabs: Cell Display**

**Counts**

☒ Observed  
☒ Expected

**Percentages**

☐ Row  
☐ Column  
☐ Total

**Residuals**

☐ Unstandardized  
☐ Standardized  
☐ Adjusted standardized

**Noninteger Weights**

☒ Round cell counts  
☐ Round case weights  
☐ Truncate cell counts  
☐ Truncate case weights  
☐ No adjustments

Continue  
Cancel  
Help

The SPSS output is given as bellow:

hairloss \* shampoo Crosstabulation

			shampoo		Total
			Noflakes	Head and Shudders	
hairloss	bald	Count	12	3	15
		Expected Count	8.5	6.5	15.0
	hairy	Count	5	10	15
		Expected Count	8.5	6.5	15.0
Total		Count	17	13	30
		Expected Count	17.0	13.0	30.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.652 <sup>a</sup>	1	.010		
Continuity Correction <sup>b</sup>	4.887	1	.027		
Likelihood Ratio	6.946	1	.008		
Fisher's Exact Test				.025	.013
Linear-by-Linear Association	6.430	1	.011		
N of Valid Cases	30				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.50.

the results of the Chi-Square test in the last table: Chi-Square is 6.65, with 1 degree of freedom, and this is significant at  $p = .01$  (i.e. there is a significant association (dependent) between shampoo choice and hair loss.

Pearson's Correlation Coefficient output: The following data concerns the blood haemoglobin (Hb) levels and packed cell volumes (PCV) of 14 female blood bank donors. It is of interest to know if there is a relationship between the two variables Hb and PCV when considered in the female population.

Hb	PCV
15.5	0.450
13.6	0.420
13.5	0.440
13.0	0.395
13.3	0.395
12.4	0.370
11.1	0.390
13.1	0.400
16.1	0.445
16.4	0.470
13.4	0.390
13.2	0.400
14.3	0.420
16.1	0.450

You can find the data in “corr data spss.xlsx” file.

SPSS produces the following correlation output:

Correlations			
		Hb	PCV
Hb	Pearson Correlation	1	.877**
	Sig. (2-tailed)		.000
	N	14	14
PCV	Pearson Correlation	.877**	1
	Sig. (2-tailed)	.000	
	N	14	14

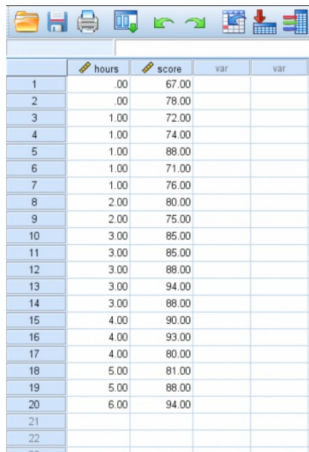
\*\* . Correlation is significant at the 0.01 level (2-tailed).

The Pearson correlation coefficient value of 0.877, i.e. there appears to be a positive correlation between the two variables.



## Linear Regression:

Suppose we have the following dataset that shows the number of hours studied and the exam score received by 20 students:



	hours	score	var	var
1	.00	67.00		
2	.00	78.00		
3	1.00	72.00		
4	1.00	74.00		
5	1.00	88.00		
6	1.00	71.00		
7	1.00	76.00		
8	2.00	80.00		
9	2.00	75.00		
10	3.00	85.00		
11	3.00	85.00		
12	3.00	88.00		
13	3.00	94.00		
14	3.00	88.00		
15	4.00	90.00		
16	4.00	93.00		
17	4.00	80.00		
18	5.00	81.00		
19	5.00	88.00		
20	6.00	94.00		
21				
22				
23				

first table of the output we're interested in is the one titled Model Summary:

Model Summary <sup>b</sup>				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.712 <sup>a</sup>	.506	.479	5.86100

a. Predictors: (Constant), hours

b. Dependent Variable: score

The next table we're interested in is titled Coefficients:

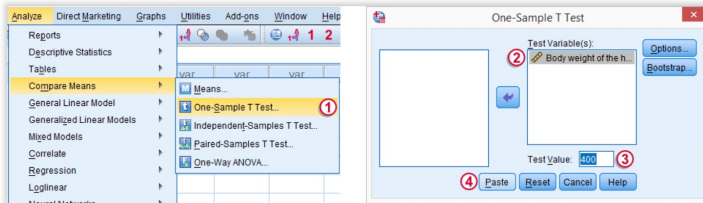
Coefficients <sup>a</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	t
1	(Constant)	73.662	2.410		30.571
	hours	3.342	.778	.712	4.297

a. Dependent Variable: score

The regression equation is Estimated exam score =  $73.662 + 3.342 \times (\text{hours})$ . Results showed that there was a statistically significant relationship between hours studied and exam score ( $t = 4.297$ ,  $p = 0.000$ ) and hours studied accounted for 50.6% of explained variability (variation) in exam score.

## T test for a Population Mean:

Example: A scientist from Greenpeace believes that herrings in the North Sea don't grow as large as they used to. It's well known that - on average - herrings should weigh 400 grams. The scientist catches and weighs 40 herrings, resulting in herrings.sav. Can we conclude from these data that the average herring weighs less than 400 grams?



One-Sample Test						
	Test Value = 400					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
body_weight	① -2.428	② 39	③ .020	④ -30.450	-55.81	-5.09

The actual t-test results are found in the **One-Sample Test** table.

① - ② The **t** value and its degrees of freedom (**df**) are not immediately interesting but we'll need them for reporting later on.

③ The p value, denoted by "**Sig. (2-tailed)**" is .02; if the population mean is exactly 400 grams, then there's only a 2% chance of finding the result we did. We usually reject the null hypothesis if  $p < .05$ . We thus conclude that herrings do not weight 400 grams (but probably less than that).

It's important to notice that the p value of .02 is **2-tailed**. This means that the p value consists of a 1% chance for finding a difference  $< -30$  grams and another 1% chance for finding a difference  $> 30$  grams.

④ The **Mean Difference** is simply the sample mean minus the hypothesized mean ( $369.55 - 400 = -30.45$ ). We could have calculated it ourselves from previously discussed results.

"we found that, on average, herrings weighed less than 400 grams;  $t(39) = -2.4$ ,  $p = .020$ ."

## Independent-Samples T-Test:

A two sample t-test is used to test whether or not the means of two populations are equal.

Example: Researchers want to know if a new fuel treatment leads to a change in the average miles per gallon of a certain car. To test this, they conduct an experiment in which 12 cars receive the new fuel treatment and 12 cars do not. The following screenshot shows the mpg for each car along with the group they belong to (0 = no fuel treatment, 1 = fuel treatment):

	mpg	group	var	var
1	20.00	.00		
2	23.00	.00		
3	21.00	.00		
4	25.00	.00		
5	18.00	.00		
6	17.00	.00		
7	18.00	.00		
8	24.00	.00		
9	20.00	.00		
10	24.00	.00		
11	23.00	.00		
12	19.00	.00		
13	24.00	1.00		
14	25.00	1.00		
15	21.00	1.00		
16	22.00	1.00		
17	23.00	1.00		
18	18.00	1.00		
19	17.00	1.00		
20	26.00	1.00		
21	24.00	1.00		
22	27.00	1.00		
23	21.00	1.00		
24	23.00	1.00		
25				
26				
27				

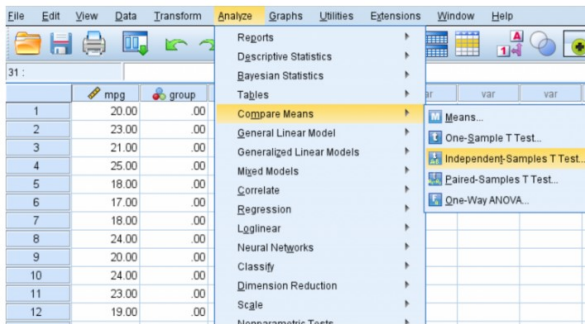
Use the following steps to perform a two sample t-test to determine if there is a difference in average mpg between these two groups, based on the following null and alternative hypotheses:

$H_0 : \mu_1 = \mu_2$  (average mpg between the two populations is equal)

$H_1 : \mu_1 \neq \mu_2$  (average mpg between the two populations is not equal)

Use a significance level of  $\alpha = 0.05$ .

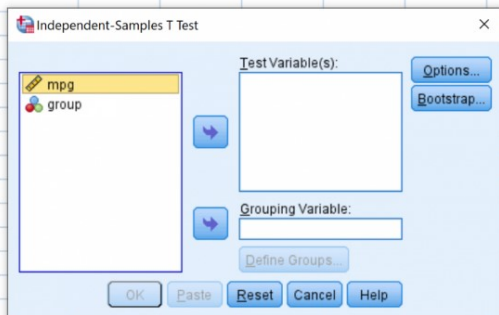
Step 1: Choose the Independent Samples T Test option.  
Click the **Analyze** tab, then Compare **Means**, then **Independent – Samples T Test**:



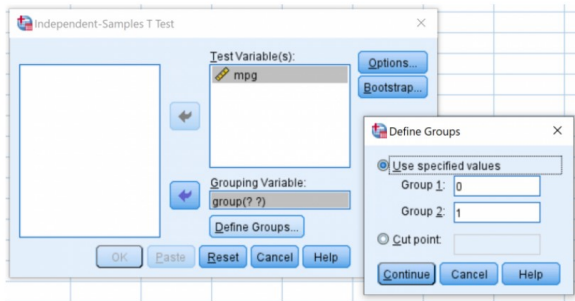


Step 2: Fill in the necessary values to perform the two sample t-test.

Once you click ***Independent — Samples T Test***, the following window will appear:



Drag the mpg into the box labelled Test Variable(s) and group into the box labelled Grouping Variable. Then click Define Groups and define Group 1 as the rows with value 0 and define Group 2 as the rows with value 1. Then click OK.



### Step 3: Interpret the results.

Once you click OK, the results of the two sample t-test will be displayed:

#### T-Test

Group Statistics

group	N	Mean	Std. Deviation	Std. Error Mean
mpg .00	12	21.0000	2.73030	.78817
1.00	12	22.7500	3.25087	.93845

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
mpg	Equal variances assumed	.034	.855	-1.428	22	.167	-1.75000	1.22552	-4.29157 .79157
	Equal variances not assumed			-1.428	21.362	.168	-1.75000	1.22552	-4.29597 .79597

The first table displays the following summary statistics for both groups:

N: The sample size

Mean: The mean mpg of cars in each group

Std. Deviation: The standard deviation of the mpg of cars in each group

Std. Error Mean: The standard error of the mean mpg, calculated as  $S/\sqrt{(n)}$

The second table displays the results of the two sample t-test. The first row shows the results of the test if you assume that the variance between the two groups is equal. The second row shows the results of the test if you don't make this assumption.

In this case, the two versions of the test produce nearly identical results. Thus, we will simply refer to the results of the first row:

t: The test statistic, found to be  $-1.428$

df: The degrees of freedom, calculated as

$$n_1 + n_2 - 2 = 12 + 12 - 2 = 22$$

Sig. (2-tailed): The two-sided p-value that corresponds to a t value of  $-1.428$  with  $df = 22$

Mean Difference: The difference between the two sample means

Std. Error Difference: The standard error of the mean difference

95% C.I. of the Difference: The 95% confidence interval for the true difference between the two population means

Since the p-value of the test (.167) is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the true mean mpg is different between cars that receive treatment and cars that don't.

## Paired-Samples T-Test

Example: A teacher developed 3 exams for the same course. He needs to know if they're equally difficult so he asks his students to complete all 3 exams in random order. Only 19 students volunteer. Their data -partly shown below- are in compare-exams.sav. They hold the number of correct answers for each student on all 3 exams.

The Hypotheses:

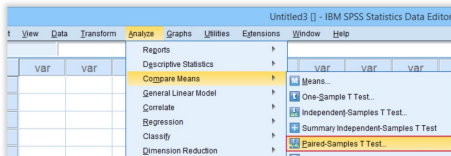
$$H_0 : \mu_{Exam1} - \mu_{Exam2} = \mu_d = 0, \text{ vs}$$

$$H_1 : \mu_{Exam1} - \mu_{Exam2} = \mu_d \neq 0.$$

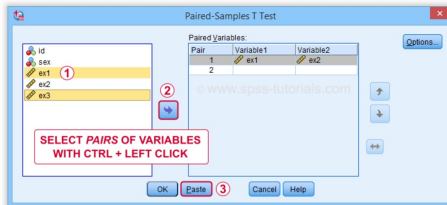
$$\text{and } H_0 : \mu_{Exam1} - \mu_{Exam3} = \mu_d = 0, \text{ vs}$$

$$H_1 : \mu_{Exam1} - \mu_{Exam3} = \mu_d \neq 0.$$

You find the paired samples t-test under **Analyze** SPSS Menu Arrow **Compare Mean** SPSS Menu Arrow **Paired Samples T Test** as shown below.



In the dialog below, ① select each pair of variables and ② move it to "Paired Variables". For 3 pairs of variables, you need to do this 3 times.



SPSS creates 3 output tables when running the test. The last one -Paired Samples Test- shows the actual test results.

www.spss-tutorials.com

		Paired Samples Test							
		Paired Differences		95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	Lower				Upper
Pair 1	Exam Version 1 - Correct Answers - Exam Version 2 - Correct Answers	- .579	2.524	.579	-1.795	.637	-1.000	18	.331

①                      ③                      ②

“The means of exams 1 and 2 did not differ,  $t(18) = 1.00$ ,  $p = 0.33$ .”

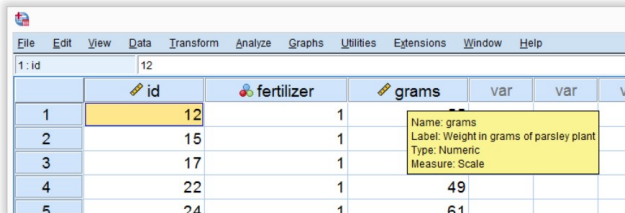
- ① SPSS reports the mean and standard deviation of the difference scores for each pair of variables. The mean is the difference between the sample means. It should be close to zero if the populations means are equal.
- ② The mean difference between exams 1 and 2 is **not statistically significant** at  $\alpha = 0.05$ . This is because 'Sig. (2-tailed)' or  $p > 0.05$ .
- ③ The 95% **confidence interval** includes zero: a zero mean difference is well within the range of likely population outcomes.

In a similar vein, the second test (not shown) indicates that the means for exams 1 and 3 *do* differ statistically significantly,  $t(18) = 2.46$ ,  $p = 0.025$ . The same goes for the final test between exams 2 and 3.



## One-Way-ANOVA:

Example: A farmer wants to know which fertilizer is best for his parsley plants. So he tries different fertilizers on different plants and weighs these plants after 6 weeks. The data -partly shown below- are in parsley.sav.



	id	fertilizer	grams	var	var	v
1	12	1				
2	15	1				
3	17	1				
4	22	1	49			
5	24	1	61			

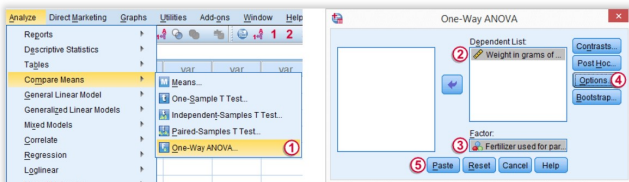
Name: grams  
 Label: Weight in grams of parsley plant  
 Type: Numeric  
 Measure: Scale

The Hypotheses:

$H_0$  : all population means are equal ( $\mu_1 = \mu_2 = \dots = \mu_k$ ), vs

$H_1$  : at least one mean is not equal ( $\mu_i \neq \mu_j, i \neq j$ ).

Now, we run a basic ANOVA from the menu. The screenshot below guides you through.



## SPSS One-Way ANOVA Output:

ANOVA

Weight in grams of parsley plant

© www.spss-tutorials.com

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	502.867	2	251.433	3.743	.028
Within Groups	5844.733	87	67.064		
Total	6347.600	89			

**"The means are significantly different,  $F(2,87) = 3.74$ ,  $p = 0.028$ ."**

So we reject the null hypothesis that all population means are equal.

Conclusion: different fertilizers perform differently. The differences between our mean weights are statistically significant.