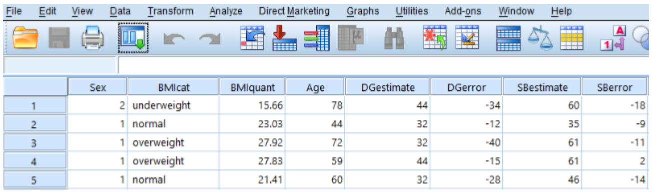
Compare two groups in SPSS

In this document we show how to compare 2 groups. We start with analysis of matched pairs and then show how make confidence intervals and testing for independent samples. We will show procedures for proportions and means of quantitative variables. For this document we need 'Skeleton' and 'Life Expectancy' data sets. It is assumed that you have managed to upload all these data into SPSS (please refer to 'Data sets import in SPSS' document for detailed explanation).

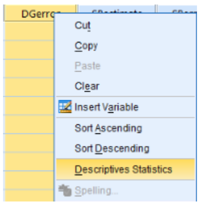
Matched Pairs

This section shows how to find confidence intervals and perform statistical testing for the difference between two dependent groups. Consider first the 'Skeleton' data set:



	Sex	BMIcat	BMIquant	Age	DGestimate	DGerror	SBestimate	SBerror
1	2	underweight	15.66	78	44	-34	60	-18
2	1	normal	23.03	44	32	-12	35	-9
3	1	overweight	27.92	72	32	-40	61	-11
4	1	overweight	27.83	59	44	-15	61	2
5	1	normal	21.41	60	32	-28	46	-14

We have two methods of age estimation here. It is the method of Di Gangi and Suchey-Brooks method. The error of estimation is captured in two variables 'DGerror' and 'SBerror' respectively. The goal is to understand if both methods give the same results or one method is more precise than the other. First let's look at 'DGerror' variable. To get basic summary statistics for this variable, **right click** on the header of this variable:



Select 'Descriptives Statistics' and the table of statistics is produced:

Statistics

Est. - Act. age using D (years)

N	Valid	400
	Missing	0
Mean		-14.15
Median		-13.00
Std. Deviation		14.126
Range		92
Minimum		-60
Maximum		32

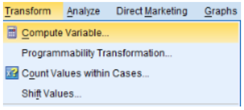
Similarly we do for the 'SBerror' variable:

Statistics

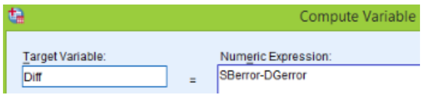
Est. - Act. age using SB (years)

N	Valid	398
	Missing	2
Mean		-7.26
Median		-6.00
Std. Deviation		10.498
Range		56
Minimum		-36
Maximum		20

Note that 2 observations are missing in the 'SBerror'. It seems that Suchey-Brooks' method is less biased than Di Gangi's method. We want to analyse the difference between these two variables Since these two variables are dependent (since two observations are taken from the same skeleton) we just want to find difference between 'SBerror' and 'DGerror'. To do that, go to **Transform** > **Compute Variable**



We call the new variable of differences 'Diff' and enter a simple expression:



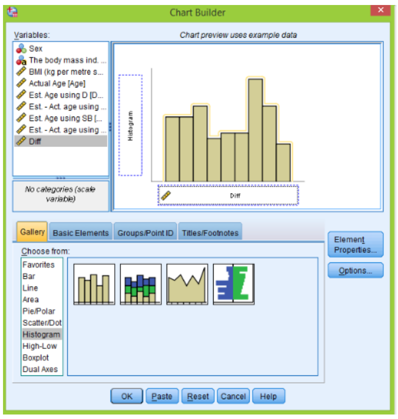
The new variable is calculated and we get summary statistics for this variable:

Statistics

Diff

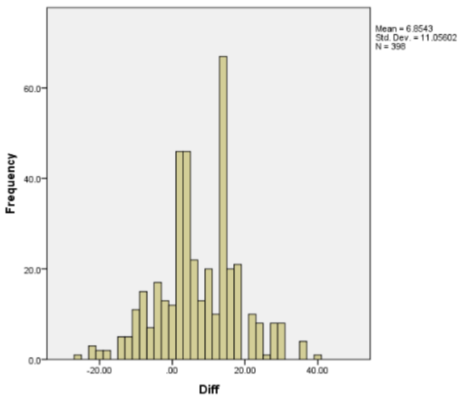
N	Valid	398
	Missing	2
Mean		6.8543
Median		6.0000
Std. Deviation		11.05602
Range		66.00
Minimum		-26.00
Maximum		40.00

Let's also make a histogram of this variable: **Graph** > **Chart Builder** > **Histogram** > **double click on Simple Histogram** then drag 'Diff' to horizontal axis:

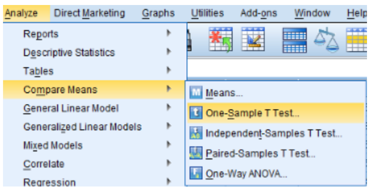


Click **OK** and the histogram is produced:

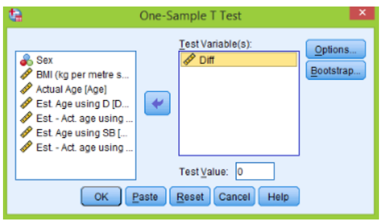
GGraph



The distribution of the differences looks nearly normal and therefore we can use one sample t-test to check if the true mean of difference is zero or not (two sided alternative). As we explained in the last document, open 'One-Sample T Test' **Analyze > Compare Means > One-sample T Test**



Send 'Diff' variable across using the arrow:



Click **OK** and results are produced:

T-Test

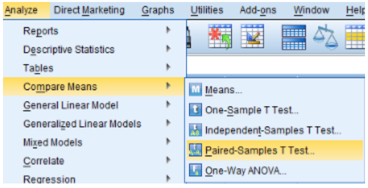
One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Diff	398	6.8543	11.05602	.55419

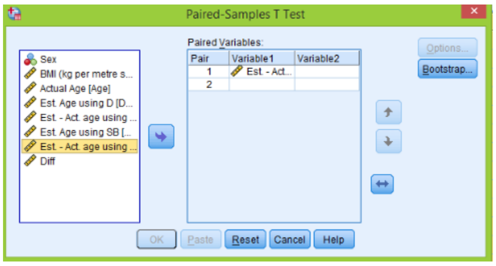
One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Diff	12.368	397	.000	6.85427	5.7648	7.9438

Here we can see the 95% confidence interval, p-value and other important information. Hence we are 95% sure that the true mean of difference of errors is between 5.76 and 7.94. The p-value is nearly zero and hence we reject the null hypothesis that two methods are the same (have the same average error) and conclude that there is a difference between them. Instead of using the long way that we have implemented above using 'Diff' variable, we can use another equivalent method. Go to **Analyze > Compare Means > Paired-Samples T Test**



Then put 'SBerror' in the first column and 'DGerror' into the second using arrow (we use only 'Pair 1')



Click **OK** to produce the next table:

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Est. - Act. age using SB (years)	-7.26	398	10.498	.526
	Est. - Act. age using D (years)	-14.11	398	14.142	.709

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Est. - Act. age using SB (years) & Est. - Act. age using D (years)	398	.633	.000

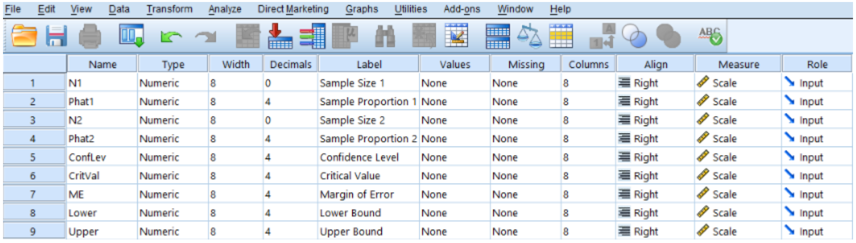
Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Est. - Act. age using SB (years) - Est. - Act. age using D (years)	6.854	11.056	.554	5.765	7.944	12.368	397	.000

Note that the confidence interval and the p-value is completely the same that we got before but this method is much quicker.

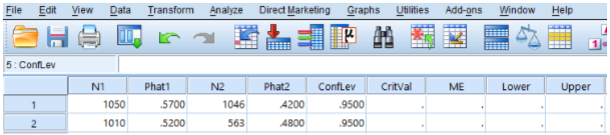
Comparing Two Proportions

In this section we show how two compare two independent proportions. We start with the 'Support for the Toronto mayor Rob Ford' example. We have two support surveys. In the first one the sample size was 1050 with sample proportion of support equals to 0.57 and another one with sample size of 1046 and sample proportion 0.42. The goal is to get 95% confidence interval for the difference in proportions. As usually for proportions, we first initialize new variables in the 'Variable View' section:



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	N1	Numeric	8	0	Sample Size 1	None	None	8	Right	Scale	Input
2	Phat1	Numeric	8	4	Sample Proportion 1	None	None	8	Right	Scale	Input
3	N2	Numeric	8	0	Sample Size 2	None	None	8	Right	Scale	Input
4	Phat2	Numeric	8	4	Sample Proportion 2	None	None	8	Right	Scale	Input
5	ConfLev	Numeric	8	4	Confidence Level	None	None	8	Right	Scale	Input
6	CritVal	Numeric	8	4	Critical Value	None	None	8	Right	Scale	Input
7	ME	Numeric	8	4	Margin of Error	None	None	8	Right	Scale	Input
8	Lower	Numeric	8	4	Lower Bound	None	None	8	Right	Scale	Input
9	Upper	Numeric	8	4	Upper Bound	None	None	8	Right	Scale	Input

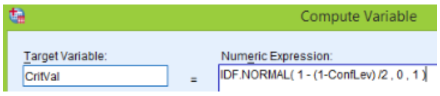
The table is filed automatically we only change 'Decimals' and 'Measure' and add some labels. Next in the 'Data View' section we enter summary statistics from the above problem in the first row. We also enter summary statistics for the 'Support for US president Obama' example in the second row to save some time (confidence level is fixed at 95% level).



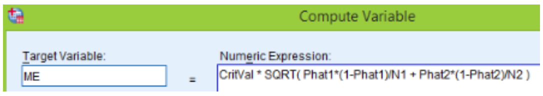
5 : ConfLev

	N1	Phat1	N2	Phat2	ConfLev	CritVal	ME	Lower	Upper
1	1050	.5700	1046	.4200	.9500	.	.	.	.
2	1010	.5200	563	.4800	.9500	.	.	.	.

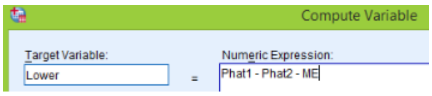
Now we use a calculator to compute 'Critical Value' (**Transform** > **Compute Variable**)

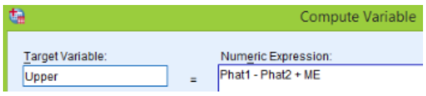


Next we use the formula for the margin of error for the difference in proportions:

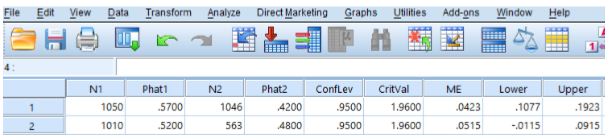


Finally we find lower and upper bounds for our confidence intervals:





The work is done, and we can observe the results:

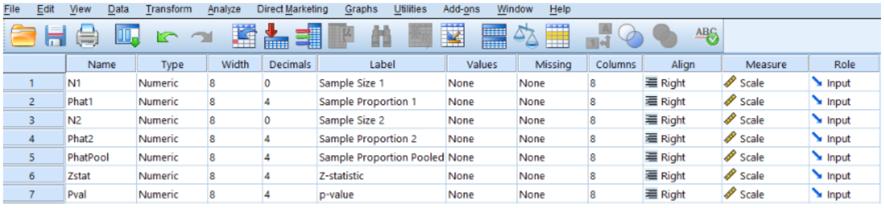


4 :

	N1	Phat1	N2	Phat2	ConfLev	CritVal	ME	Lower	Upper
1	1050	.5700	1046	.4200	.9500	1.9600	.0423	.1077	.1923
2	1010	.5200	563	.4800	.9500	1.9600	.0515	-.0115	.0915

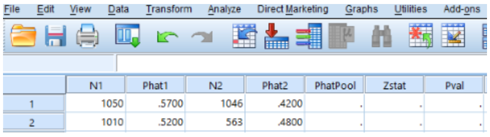
Hence we are 95% confident that the true proportion for 'Toronto Mayor' dropped from 0.10 to 0.19. In 'US President' example we are 95% sure that the true proportion dropped from -0.01 to 0.09, hence we cannot be sure that the support actually dropped since 0 is inside the confidence interval.

Next instead of finding confidence intervals we want to test equality of two proportions. In the 'Variable View' we change some variables:



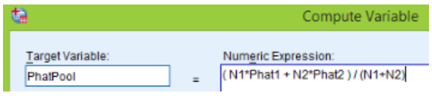
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	N1	Numeric	8	0	Sample Size 1	None	None	8	Right	Scale	Input
2	Phat1	Numeric	8	4	Sample Proportion 1	None	None	8	Right	Scale	Input
3	N2	Numeric	8	0	Sample Size 2	None	None	8	Right	Scale	Input
4	Phat2	Numeric	8	4	Sample Proportion 2	None	None	8	Right	Scale	Input
5	PhatPool	Numeric	8	4	Sample Proportion Pooled	None	None	8	Right	Scale	Input
6	Zstat	Numeric	8	4	Z-statistic	None	None	8	Right	Scale	Input
7	Pval	Numeric	8	4	p-value	None	None	8	Right	Scale	Input

Next we fill the table in the 'Data View' section (completely the same as before):

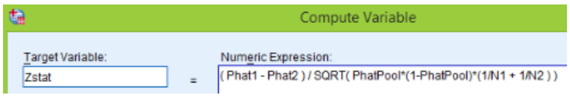


	N1	Phat1	N2	Phat2	PhatPool	Zstat	Pval
1	1050	.5700	1046	.4200	.	.	.
2	1010	.5200	563	.4800	.	.	.

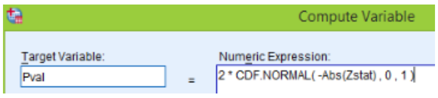
Now we start using the calculator function. First we compute the 'Pooled Sample Proportion' using the next expression:



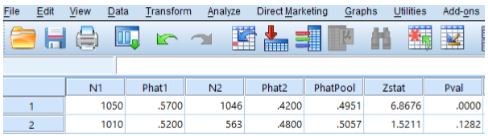
Afterwards we compute the z-statistic



Finish this process with two sided p-value calculation:



The results are displayed below:

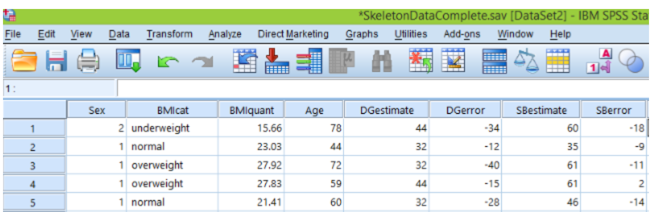


	N1	Phat1	N2	Phat2	PhatPool	Zstat	Pval
1	1050	.5700	1046	.4200	.4951	6.8676	.0000
2	1010	.5200	563	.4800	.5057	1.5211	.1282

So for the 'Toronto Mayor' example the p-value is almost zero and we reject null hypothesis that two proportions are the same and conclude that they are not the same. For the 'US President' example the p-value is 0.13 which is not significant and therefore we cannot reject that two proportions are the same.

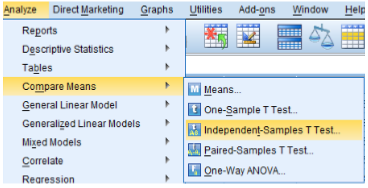
Comparing Two Means

In this section we show how to compare two independent quantitative groups. Consider first the 'Skeleton' data set.

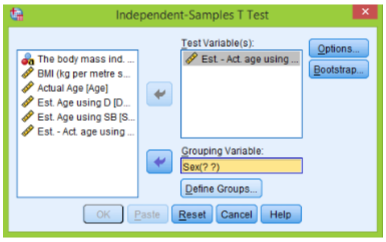


	Sex	BMIcat	BMIquant	Age	DGestimate	DGerror	SBestimate	SBerror
1	2	underweight	15.66	78	44	-34	60	-18
2	1	normal	23.03	44	32	-12	35	-9
3	1	overweight	27.92	72	32	-40	61	-11
4	1	overweight	27.83	59	44	-15	61	2
5	1	normal	21.41	60	32	-28	46	-14

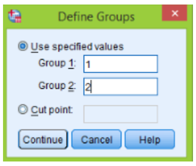
We want to find the 95% confidence interval for the difference between 'DGerror' for male and female and also test if the difference is zero or not. In SPSS it is very easy to do. Go to **Analyze > Compare Means > Independent-Samples T Test**:



Here move 'DGerror' to 'Test Variable(s)' and 'Sex' to 'Grouping Variable'.



Next click on 'Define Groups' button and enter 1 to 'Group 1' and 2 to 'Group 2':



Click **Continue > OK** and the following table is printed:

Group Statistics

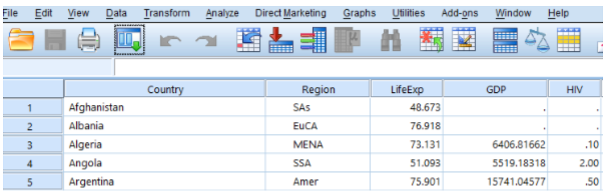
	Sex	N	Mean	Std. Deviation	Std. Error Mean
Est. - Act. age using D (years)	Male	281	-12.90	13.473	.804
	Female	119	-17.10	15.214	1.395

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Est. - Act. age using D (years)	Equal variances assumed	1.162	.282	2.741	398	.006	4.200	1.532	1.188	7.213
	Equal variances not assumed			2.610	200.095	.010	4.200	1.610	1.026	7.375

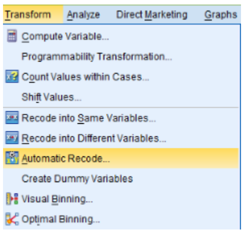
Note that we have two rows in the table. The first one assumes that two variances are the same, the second row does not have this assumption. We do not assume here that variables of 'DGerror' for male and female are the same and therefore focus on the second row. Based on the results we are 95% confident that error of estimation for male is from 1.03 to 7.37 larger than for female. Also the p-value is quite small and hence we reject null hypothesis that the true means of 'DGerror' for male and female are the same.

Consider next the 'Life Expectancy' data set:

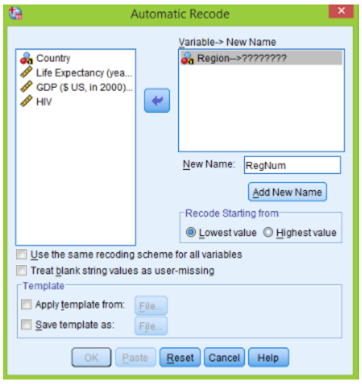


In this example we need 95% confidence interval for difference between 'LifeExp' for East Asia

& Pacific (EAP) and South Asia (SAs). Unfortunately we cannot use the ‘Region’ as ‘Grouping Variable’ as we did for ‘Sex’ variable because ‘Region’ consists of string observations rather than some numbers. To solve this problem, we will construct a new variable (mimicking the ‘Region’) but with numbers from 1 to 6 instead of string observations. To do that, go to **Transform** > **Automatic Recode**



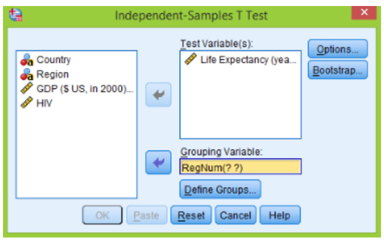
Move ‘Region’ to the right window and give a name for a new variable (‘RegNum’):



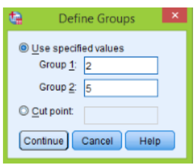
Click on the ‘Add New Name’ button and then **OK**, the column is produced:

Country	Region	LifeExp	GDP	HIV	RegNum
Afghanistan	SAs	48.673	.	.	5
Albania	EuCA	76.918	.	.	3
Algeria	MENA	73.131	6406.81662	.10	4
Angola	SSA	51.093	5519.18318	2.00	6
Argentina	Amer	75.901	15741.04577	.50	1
Armenia	EuCA	74.241	4748.92858	.10	3
Aruba	Amer	75.246	.	.	1
Australia	EAP	81.907	34642.38813	.10	2

We observe that ‘EAP’ region corresponds to 2 and ‘SAs’ to 5. Now as before go to **Analyze** > **Compare Means** > **Independent-Samples T Test**, move ‘Life Expectancy’ variable to the right window and ‘RegNum’ to ‘Grouping Variable’:



Click on the ‘Define Groups’ button and type 2 and 5 to ‘Group 1’ and ‘Group 2’ respectively:



Finish by clicking on **Continue** > **OK** to get the results:

Group Statistics

	RegNum	N	Mean	Std. Deviation	Std. Error Mean
Life Expectancy (years)	EAP	30	73.08603	6.220434	1.135691
	SAs	8	67.03263	8.517471	3.011381

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Life Expectancy (years)	Equal variances assumed	.014	.905	2.261	36	.030	6.053408	2.677455	.623277	11.483540
	Equal variances not assumed			1.881	9.088	.092	6.053408	3.218417	-1.216372	13.323188

So we are 95% sure that the true difference in averages of Life expectancy for these two regions is between −1.22 and 13.32. Also the p-value is 0.09 which is considered as large and hence there is no statistical evidence to reject the hypothesis that 'Life Expectancy' for two region are same. So they can be the same.