

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:

File	Edit	View	<u>D</u> ata	a <u>T</u> ransform	Analyze Di	rect <u>M</u>	arketing	<u>G</u> raphs <u>U</u> tiliti	es Add- <u>o</u> ns	Wir	ndow <u>H</u> elp	
					¥ 🎼			P H	*; 🖌		- <u>-</u>	A 14
		Sex		BMIcat	BMIqua	nt	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		14	-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:

St	ati	ist	ic	ş

Est Act. age using D (years)						
Ν	Valid	400				
	Missing	0				
Mean		-14.15				
Media	an	-13.00				
Std. D	Deviation	14.126				
Rang	e	92				
Minim	num	-60				
Maxin	num	32				

Similarly we do for the 'SBerror' variable:

Statistics

Est Act. age using SB (years)						
N	Valid	398				
	Missing	2				
Mean		-7.26				
Media	an	-6.00				
Std. D	Deviation	10.498				
Rang	e	56				
Minin	num	-36				
Maxir	num	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**

Transform	<u>A</u> nalyze	Direct Marketing	Graphs							
Compute Variable										
Program	Programmability Transformation									
Count Values within Cases										
Shi <u>f</u> t Valu	Shift Values									

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

ta		Compute Variable			
Target Variable: Diff	1 -	Numeric Expression: SBerror-DGerror			

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

Statistics							
Diff							
Ν	Valid	398					
	Missing	2					
Mean		6.8543					
Media	an	6.0000					
Std. D	Deviation	11.05602					
Rang	е	66.00					
Minin	num	-26.00					
Maxin	num	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:

1 8	Chart Builder	×				
Variables: Chart preview uses example data						
Sex The box mass ind This box mass ind Actual kge (kge) Actual kge (kge) Ext Age using D(D Ext Act arge using Ext Age using Ext Age using Dat Actual cge (kge) Actual cge (kge) Actual cge (kge) Actual cge)						
Gallery Basic Elements Choose from: Favorites Bar Line Area Pre/Polar Histogram High-Low Borglot Dual Ares	Element Properties					
OK Paste Reset Cancel Help						

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



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**

Analyze	Direct Marketing	Graphs	Utilities	Add- <u>o</u> ns	Win	dow	<u>H</u> elp
Repor	ts	*	*			47	
Descr	iptive Statistics		• ••••			-0	
Tables	5						
Co <u>m</u> p	are Means	- F.	Means				
<u>G</u> ener	al Linear Model		One-S	ample T T	est		
Gener	alized Linear Model	s 🕨	Indepe	ndent-Sa	mples 1	Test	
Mi <u>x</u> ed	Models	•	Paired	Samples	TTest		
Correl	ate			- ANOV	11030		
Reare	ssion	× .	<u>One-w</u>	ay ANOVA	.		

Send 'Diff' variable across using the arrow:

G	One-Sample T Test	~
Sox Mit(kg per metre s Katual Age (Age) Est Age using D [D Est Age using D [D Est Age using SB [Est Age using SB [Est Age using SB [Est Age using On Est Age using	Test Variable(s):	Options Bootstrap

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						
				Mean	95% Confidence Interval o Difference		
	t	df	Sig. (2-tailed)	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

Analyze	Direct Marketing	Graphs	Utilities	Add- <u>o</u> ns	Window	Help				
Repo	rts	•	*							
Desc	riptive Statistics									
Table	s	+								
Com	oare Means	- F.	Means							
<u>G</u> ene	ral Linear Model	•	One-Sample T Test							
Gene	ralized Linear Model	Is 🕨	Indepe	ndent-Sar	mples T Test					
Mi <u>x</u> ed	Models		Paired Pamples T Test							
Corre	late			- ANOVA	11000					
Reare	ession		Due-w	ay ANUVA	L					

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

Paired-Samples T Test	×
Concept Paired Variables - Feat Paired Variables - Feat Paired Variables - Feat Paired Variables - Variables Paired Variables - Variables - Variables Paired Variables - Variables - Variables Paired Variables - Var	

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

T-Test

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

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:

<u>F</u> ile <u>E</u> dit	<u>V</u> iew <u>D</u> ata	Transform	<u>A</u> nalyze E	Direct <u>M</u> arketi	ing <u>G</u> raphs <u>U</u> tiliti	es Add- <u>o</u> ns	<u>Window H</u>	elp			
🔁 H 🌒 💷 🖛 🛥 📓 🏪 💷 🃭 H 📓 🕱 📰 👫 🐼 🖷 👫 🐼 🧠 🤏											
	Name	Туре	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	滬 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).

File	Edit	View [<u>)</u> ata	Transform	m <u>A</u> nalyze	Direct Mar	keting <u>G</u> ra	phs <u>U</u> tilitie:	s Add- <u>o</u> ns	Window	<u>H</u> elp
				2	2	š 📥 :	씨	#1 🕺	ş 👱	- 4	1
5 : Cor	nfLev										
		N1		Phat1	N2	Phat2	ConfLev	CritVal	ME	Lower	Upper
1	1	10	50	.5700	1046	.4200	.9500				
2	2	10	10	.5200	563	.4800	.9500				

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

ta		Compute Variable
Target Variable: CritVal] =	Num <u>e</u> ric Expression: IDF.NORMAL(1-(1-ConfLev)/2,0,1)

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

t e		Compute Variable					
Target Variable: ME	=	Num <u>eric Expression:</u> CritVal * SORT(Phat1*(1-Phat1)/N1 + Phat2*(1-Phat2)/N2)					

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

6		Compute Variable
Target Variable:		Numeric Expression:
Lower	=	Phat1 - Phat2 - ME
5		Compute Variable
Target Variable:		Numeric Expression:
Upper	=	Phat1 - Phat2 + ME

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

File	Edit	View	<u>D</u> ata	Transform	m <u>A</u> nalyze	Direct Mar	keting <u>G</u> ra	phs <u>U</u> tilitie:	s Add- <u>o</u> ns	Window	Help	
				2	7	i 📥 i		H 🕈		- 4	1.	
4 :	4:											
		N1		Phat1	N2	Phat2	ConfLev	CritVal	ME	Lower	Upper	
1		1	050	.5700	1046	.4200	.9500	1.9600	.0423	.1077	.1923	
2		1	010	.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:

	<u>File</u> <u>E</u> dit	<u>V</u> iew <u>D</u> ata	Transform	Analyze	Direct <u>M</u> arketi	ing <u>G</u> raphs <u>U</u> tilities	Add- <u>o</u> ns <u>W</u> in	dow <u>H</u> elp				
😂 H 🖨 💷 🖛 🛥 📓 🏪 📰 🃭 M 🗏 🕱 🚍 🖧 🖽 🚜 🔕 🌑 🤲												
		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	ゝ 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):

Edit	<u>V</u> iew <u>D</u> at	a <u>T</u> ransfor	m <u>A</u> nalyze	Direct <u>M</u> ar	keting <u>G</u> raj	phs <u>U</u> tilities	Add- <u>o</u> ns
H			1	ä 🚣 🗉	Ч	H *	
	NI	Phat1	N2	Phat2	PhotPool	Zctat	Pyal
	1050	.5700	1046	.4200	rilatr o'o'	25101	Frai
2	1010	.5200	563	.4800			
	Edit	Edit View Dat N1 N1 1050 1010	Edit View Data Iransfor	Edit Yiew Data Transform Analyze Image: State S	Edit Yiew Data Transform Analyze Direct Mar N1 Phat1 N2 Phat2 P	Exit Yew Data Transform Analyze Dired Markeling Grav Image: State S	Edit Yiew Data Transform Analyze Direct Marketing Graphs Utilities Image: Strate Strate Strate Image: Strate Strate Strate Image: Strate Strate Strate Image: Strate Strate Strate Strate Image: Strate Strate Strate Strate Image: Strate Strate Strate Strate Strate Strate Strate Strate Image: Strate Image: Strate Image: Strate

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

6	Compute Variable	
Target Variable: PhatPool	=	Num <u>e</u> ric Expression: (N1*Phat1 + N2*Phat2) / (N1+N2)

Afterwards we compute the z-statistic

ta		Compute Variable
Target Variable: Zstat	-	Numeric Expression: (Phat1 - Phat2) / SQRT(PhatPool*(1-PhatPool)*(1/N1 + 1/N2))

Finish this process with two sided p-value calculation:

ta		Compute Variable
Target Variable: Pval	-	Numeric Expression: 2 * CDF.NORNAL(-Abs(Zstat) , 0 , 1)

The results are displayed below:

File	Edit	View	Data	Transfor	m <u>A</u> nalyze	Direct <u>M</u> ar	keting <u>G</u> raj	phs <u>U</u> tilitie	s Add- <u>o</u> ns
			0,	2	~	š 📥 E		H 🛎	i 🖬
		N1		Phat1	N2	Phat2	PhatPool	Zstat	Pval
1		1	1050	.5700	1046	.4200	.4951	6.8676	.0000
2		1	010	.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.

					*SkeletonDat	aComplete.sa	v [DataSet2] - I	IBM SPSS Sta
<u>F</u> ile <u>E</u> dit	<u>V</u> iew <u>D</u> at	a <u>T</u> ransform <u>A</u> r	nalyze Direct	Marketing	<u>G</u> raphs <u>U</u> tilities	Add-ons	<u>M</u> indow <u>H</u> elp	
😂 H		📮 🗠 계			М 🔠		4 <u>3</u>	A 14
1:								
	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:

A	Bin difference	Our star			Mr. dam	
Analyze	Direct Marketing	Graphs	Utilities	Add-ons	window	Help
Repo	orts	•	*			
Desc	criptive Statistics					
Table	es	•				
Com	pare Means	•	Means	š		
<u>G</u> ene	eral Linear Model	•	One-S	ample T Te	st	
Gene	eralized Linear Mode	ls ▶		endent-Sam	ples T Test.	
Mi <u>x</u> ec	d Models	•	Paired	-Samples 1	Test	
Corre	elate				1000	
Rear	ession	•	<u>One-v</u>	Vay ANOVA.		

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

ta Inde	pendent-Samples T Test
The body mass ind BMI dig per metre s # Actual Age (Age) # Est Age using D [D # Est Age using SB (B) # Est - Act age using	Test Variable(s): @ptions

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

ta	Define Groups ×
0 U	se specified values
	Group <u>1</u> : 1
	Group <u>2</u> : 2
Og	ut point:
Co	ntinue Cancel Help

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

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

	Independent Samples Test											
		Levene's Test Varia	for Equality of nces				t-test for Equality	of Means				
			Mean Std F						95% Confidence Interval of the Difference			
		F	Sig.	t	df	Sig. (2-tailed)	Difference	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:

Eile	Edit	View	<u>D</u> ata	Transform	Analyze	Dire	ct <u>M</u> arketing	Graphs	s <u>U</u> tilitie	s Add- <u>o</u> n	s <u>W</u> indow	Help
			0,	Ľ	~		╘	ų	H.	*, 👱	5	3 🛄 .
				Country			Regio	in	LifeEx	p	GDP	HIV
1	1	Afghan	istan				SAs		48	3.673		
2	2	Albania	1				EuCA		76	5.918		
3	3	Algeria					MENA		73	3.131	6406.8166	2.10
4	4	Angola					SSA		51	1.093	5519.1831	8 2.00
5	5	Argenti	ina				Amer		75	5.901	15741.0457	7.50

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'):

😭 Automatic Recode 💌							
Country Cutife Expectancy (yea GDP (\$ US, in 2000) HIV	Yariable≫ New Name						
	New Name: RegNum Add New Name Recode Starting from © Lowest value © Highest value						
Use the same recoding scheme for all variables Trat blank string values as user-missing Template Apply template from: Fine							
Save template as: File. OK Paste Reset Cancel Help							

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										
Leven		Levene's Test for Equality of Variances		t-test for Equality of Means						
							Mean Std. Error		95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Difference	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	<mark>-1.216372</mark>	<mark>13.323188</mark>

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.