

Summarizing Data: Relationships Between Variables in SPSS

In this document we show how to find relationship between two variables. Since quantitative and categorical variables must be treated differently we show which plots, tables and statistics are appropriate in different cases.

For this document we will use 'Skeleton' and 'Life Expectancy' data sets. It is assumed that you have managed to upload all these data into SPSS (please refer to the 'Data sets import in SPSS' document for detailed explanation).

Relationship between quantitative and categorical variables

We start this section with the 'Life Expectancy' data set. We know how to find descriptive statistics and get a boxplot of one variable, for example 'LifeExp'. Now we want to understand relationship between 'LifeExp' and 'Region'. Hence we need 'LifeExp' statistics for every region. To do that we first need to split the data, go to **Data** > **Split File**

Denters Synaktike Properties. Stylessement Level for Universe. Soft Measurement Level for Universe Sets. Soft Measurement Level for Universe Sets. Soft Variages.	Data	Transform	Analyze	Direct Marketing
2) Set Massurement (_vent for Unknown	🗇 De	fine <u>V</u> ariable F	roperties	
Bit Octo Data Propertes IF Here Custom Aftrightes IF United Data Second Sets IF Define Datas IF Datas	3 Se	t Measuremen	t Level for U	inknown
	<u>in c</u> a	py Data Prope	rties	
Bornes Judicies Conservations Setal	Ne Ne	w Custom Attr	ibute	
Dohne Mutpler Response Sets	🛗 Dg	fine Dates		
Vajdoon Validoon Validoo	🔡 De	fine <u>M</u> ultiple R	esponse Si	ets
iii Identify Digstate Cases ji Gonzy Datasets ji Gonzy Da	Va	ildation		+
E] Sterfor Unusual Cases	🔡 Ide	entify D <u>u</u> plicate	Cases	
R: Congrete Obtasets Image Parts	😋 Ide	entify Unusual	Cases	
© Bgd Cases Sold Vandges Mage Fields	📆 Co	m <u>p</u> are Datase	ts	
Boot Vandages Tragspoot Mage Fals > Mage Fals > Rest Works > Propensity Score Matching Case Control Matching Case Control Matching Soft Into Fries Ontgoornal Design Ontgoornal Design > Special Special Special Special Special	🔂 S <u>o</u>	rt Cases		
tall Tangeses	🐹 So	rt Varia <u>b</u> les		
Marge Fiels	Tr:	nspose		
∏Endestruture Rave Vivelynts Propensity Score Matching Case Control Matching Case Control Matching Sport ruto Fries Orthogonal Design Pr Cong Quaset Spece Conses A Weight Coses	M	arge Files		•
Rake Weghts Propersitely Gover Matching Case Coursel Matching 24 Agarquan 5 Spittion Fries Ortigonan Design	er Re	structure		
Propensil Score Matching Case Conrol Matching Splitto Fries Ordfogsmut Design ♪ Tip, Core Dataset Splitte Splitte Splitte	Ra	ike Weights		
Case Control Matching	Pr	opensity Score	Matching	
Approvation	Ca	ise Control Ma	tching	
Spitlinko Files Ottoponal Design ► % Copy Dataset Spitline. Spitline. Spitline. Spitline.	🔁 Ag	gregate		
Orthogonal Design ► R Copy Quitaset Split Fale Split Fale Split Cases Split Cases	Sp	lit into Files		
Re Copy Dataset Spill File Spill File Spillet Cases Weight Cases	Or	thogonal Desi	gn	• •
Spill File Select Cases 4 Weight Cases	R Co	py <u>D</u> ataset		
Select Cases 4 Weight Cases	E Sp	lit <u>F</u> ile		
4 Weight Cases	e Se	lect Cases		
	4 <u>₩</u>	eight Cases		

Here select 'Compare groups' then move 'Region' variable to the right window using arrow and also select 'Sort the file by grouping variables'



Click **OK** and you will see that data set is sorted by 'Region'. Next to get descriptive statistics (mean, median, quartiles etc.) by 'Region' just go to **Analyze** > **Descriptive Statistics** > **Frequencies**, move 'LifeExp' variable to the right, click on 'Statistics' button to select different statistics:

ta Frequencies: S	tatistics ×
Percentile Values Quarties Quarties Quarties Quarties Percentile(x) Add Change Remove	Central Tendency Gentral Tendency Gentral Median Gude Sum
	🔄 Vaļues are group midpoints
Dispersion	Distribution
🗹 Std. deviation 🗹 Minimum	Skewness
E Variance Maximum	E Kurtosis
Range S.E. mean	
Continue Cancel	Help

Click **Continue** > **OK** to produce the following tables:



See that statistics are calculated for each region. To get boxplots for each region, just go to **Chart Builder** > **Boxplot** > **double click on 1-D Boxplot** then drag 'LifeExp' to the 'X-axis'. Six separate boxplots are printed in the output window corresponding to each region. If we want to see all the boxplots in one plot, we have to return to the un-split data file. Hence go to **Data** > **Split File** and select 'Analyze all cases, do not create groups':

ta	Split File ×
Country Curry GP Life Expectancy (yea GDP (\$ US, in 2000) FIV	Analyze all cases, do not create groups Compare groups Grganize output by groups Groups Based on: A Region
Current Status: Compare:R	Sort the file by grouping variables Elle is already sorted
OK Pas	te <u>R</u> eset Cancel Help

Now when the file is not split go to ${\bf Chart}\ {\bf Builder} > {\bf Boxplot} > {\bf double\ click\ on\ Simple\ Boxplot}$



Drag 'Region' to the X-axis and 'LifeExp' to the Y-axis click **OK** and the plot is printed. Double click on the plot to open 'Chart Editor' to make some changes. Double click on any of the boxplots and 'Properties' window appears. To change color of the boxes click on 'Fill & Border', to change the width of the boxes click on 'Bar Options'



Next lets look at 'Skeleton' data set:

ta 🛛	SkeletonDataComplete.sav [DataSet2] - IBM SPSS Statistics Data Editor – 🖉 💌															
File Edit	jie Edit View Data Transform Analyze Direct.Marketing Graphs Ublities Add-gns Window Help															
			👔 📰 🛓	_ = 1	ih (A 🕴	9 12 1	- A 🗰			ABC						
	Visible: 8 of 8 Variables															
	Sex	BMIcat	BMIquant	Age	DGestimate	DGerror	SBestimate	SBerror	var	var	var	var	var	var	var	
1		2 underweight	15.66	78	44	-34	60	-18								-1
2		1 normal	23.03	44	32	-12	35	-9								- 1
3		1 overweight	27.92	72	32	-40	61	-11								- 1
4		1 overweight	27.83	59	44	-15	61	2								
5		1 normal	21,41	60	32	-28	46	-14								
- 6		1 underweight	13.65	34	25	-9	35	1								- 1
7		1 overweight	25.86	50	32	-18	35	-15								- 1
8		1 underweight	14.56	73	50	-23	61	-12								- 1
9		1 normal	22,44	70	39	-31	46	-24								- 1
10		1 normal	19.88	60	44	-16	46	-14								- 1
11		1 normal	23.24	58	32	-26	35	-23								- 11
12		1 overweight	25.09	61	32	-29	61	0								- 11
13		2 overweight	25.68	52	44	-8	48	-4								- 11
14		1 normal	24.97	67	44	-23	46	-21								- 11
15		1 normal	23.32	60	44	-16	46	-14								
16		1 normal	23.29	68	50	-18	61	-7								- 11
17		2 overweight	27.37	35	12	-23	38	3								
18		2 obese	34.82	81	39	-42	48	-33								
19		2 underweight	12.29	73	44	-29	60	-13								
20		1 normal	23.85	65	39	-26	46	-19								
21		1 normal	24.89	57	57	0	46	-11								
22		2 normal	24.69	67	32	-35	60	-7								
23		2 normal	23.18	60	44	-16	60	0								
24	_	1 normal	24.71	35	32	-3	35	0				_				-
_																-
Data View	Variable Vie	w														
										101	ODOC CRAME	ice Processor	in ready	Unicodes	00	

To get the relationship between 'DGerror' variable and 'Sex', we make the boxplots as described above: **Chart Builder** > **Boxplot** > **double click on Simple Boxplot**, drag 'Sex' variable to the X-axis and 'DGerror' to Y-axis, click **OK** and get:



Next we are interested in how 'DGerror' varies with 'BMIcat'. We get the boxplots, but observe that the order of 'BMIcat' is not logical, we want it to be 'underweight', 'normal', 'overweight' and 'obese'. It is very simple to change in SPSS. Double click on the plot then double click on any boxplot to open 'Properties' window and click on 'Categories' button



Now we put categories in the right order using arrows, to finish click **Apply** button. Also by double clicking on the 'fence' of any boxplot we can change its stile and color



Relationship between two categorical variables

In this section we show how to work with two categorical variables. We start with the 'Skeleton' data. The goal is to investigate the relationship between 'BMIcat' and 'Sex' variables (both of them are categorical). First we create a table of counts, go to **Analyze** > **Descriptive Statistics** > **Crosstabs**

<u>A</u> nalyze	Direct <u>M</u> arketing	Graphs	Utilities	Add- <u>o</u> ns	
Repo	orts		-1 ¥		ļ
Desc	criptive Statistics	- F	123 Frequ	encies	
Ta <u>b</u> le	es.	•	Desci	riptives	
Com	pare Means		A Explo	re	
<u>G</u> ene	eral Linear Model	•	Cross	stabs	
Gene	eralized Linear Mode	ls 🕨	TURF	Analysis	
Mi <u>x</u> eo	d Models	•	Ratio		
Corre	elate	•			
<u>R</u> egr	ession	•	P-P P	1015	
Logli	near		0-Q P	lots	

Then move 'BMIcat' variable to the 'Row(s)' window and 'Sex' to the 'Column(s)' section

6	Crosstabs	×
BMI (kg per metre squar Actual Age [Age] Est Age using D [DGest P Est - Age using 0 V P Est - Ada age using SB [SBest Est - Ad. age using SB (Rov(s): Column(s): Column(s): Layer 1 of 1 Proyous Net Column(s): Proyous Net	Erad. Statistics. Cells Format Style Bootstr <u>a</u> p
Display clustered <u>b</u> ar charts Suppress <u>tables OK I </u>	Display layer variables in table layers Paste Reset Cancel Help	5

Click \mathbf{OK} and a simple table of counts is produced:

Crosstabs

	Case Processing Summary												
Cases													
	Valid Missing Total												
	N	Percent	N	Percent	N	Percent							
The body mass ind. cat. * Sex	400	100.0%	0	0.0%	400	100.0%							

The body mass ind. cat. ' Sex Crosstabulation

Count				
		Se	ЭX	
		Male	Female	Total
The body mass ind. cat.	normal	166	59	225
	obese	10	10	20
	overweight	59	22	81
	underweight	46	28	74
Total		281	119	400

This table just shows how many observations are in each category. If we need joint or conditional distribution then as before **Analyze** > **Descriptive Statistics**> **Crosstabs** then click on 'Cells' button. If we need joint distribution then select 'Total' under 'Percentages' if we need for example conditional distribution given 'Sex' then select 'Column' option, here we need joint distribution:

ta Cro	osstabs: Cell Display
Counts © Observed Expected Hide small counts Less than 5	Compare column proportions Adjust p-values (Bonferroni method)
Percentages	Residuals Unstandardized Standardized Adjusted standardized
Noninteger Weights O Round cell counts Truncate cell counts No adjustments	 ○ Round case weights ○ Truncate case weights
Contin	ue Cancel Help

Click **Continue** > **OK** and a new table is printed:

Crosstabs

The

Case Processing Summary

		Cases									
	Va	lid	Mis	sing	Total						
	N	Percent	N	Percent	N	Percent					
The body mass ind. cat. * Sex	400	100.0%	0	0.0%	400	100.0%					

dy mass indicat 1 Sey Crosstabulatio

			Se	ЭX						
			Male	Female	Total					
The body mass ind. cat.	normal	Count	166	59	225					
		% of Total	41.5%	14.8%	56.3%					
	obese	Count	10	10	20					
		% of Total	2.5%	2.5%	5.0%					
	overweight	Count	59	22	81					
		% of Total	14.8%	5.5%	20.3%					
	underweight	Count	46	28	74					
		% of Total	11.5%	7.0%	18.5%					
Total		Count	281	119	400					
		% of Total	70.3%	29.8%	100.0%					

These percentages represent the joint distribution. To make a visual representation of the relationship we can use a bar-plot of counts for each 'BMIcat' and 'Sex' categories, go to **Chart Builder** > **Bar** > **double click on Clustered Bar**



Drag 'Sex' variable to the X-axis and 'BMIcat' to the 'Cluster on X: set color', click **OK** and the bar-plot is produced. Once again the order of categories of the 'BMIcat' is not good, hence double click on the plot, double click on any bar and then as usual click on 'Categories' button. Also to change colors of the bars, double click on the colored squares in the legend



If we need stacked bars then Chart Builder > Bar > double click on Stacked Bar



As before drag 'Sex' variable to X-axis and 'BMIcat' to the 'Stack: set color' and click \mathbf{OK}



These were plots of counts, but next we want to plot conditional distributions of 'BMIcat' given 'Sex' factors. Lets make a stacked bar plot. In the 'Chart Builder' section there is an 'Element Properties' window and under 'Statistic' select 'Percentage':



Then we click on the 'Set Parameters' button and select 'Total for each X-Axis Category'



Click Apply > OK to finish, the following plot is produced:



See that now the heights of the stacked bars are the same and equal to 100%.

Relationship between two quantitative variables

Let's start with 'Life Expectancy' data set.

ta 🛛			LifeExpComp	olete.sav [DataSet1]	- IBM SPSS	Statistic	s Data Editor						- 🛛 🗙
File Edit	View Data Transform Analyze D	irect Marketing Graph	ns Utilities (Add-ons Window	Help								
i 🔁 🖩	l 🖨 💷 🖛 🛥 🦉	* = P	M 👪	¥ 🖬 🖄		0	A						
1:												Visible	5 of 5 Variables
	Country	Region	LifeExp	GDP	HIV	var	var	var	var	var	var	var	var
1	Afghanistan	SAs	48.673										<u>~</u>
2	Albania	EUCA	76.918										
3	Algeria	MENA	73.131	6406.81662	.10								
- 4	Angola	SSA	51.093	5519.18318	2.00								
5	Argentina	Amer	75.901	15741.04577	.50								
6	Armenia	EUCA	74.241	4748.92858	.10								
7	Aruba	Amer	75.246										
8	Australia	EAP	81.907	34642.38813	.10								
9	Austria	EUCA	80.854	36871.07135	.30								
10	Azerbaijan	EUCA	70.739	9387.38950	.10								
- 11	Bahamas	Amer	75.620	21470.80420	3.10								
12	Bahrain	MENA	75.057										
13	Bangladesh	SAs	68.944	1669.87427	.06								
14	Barbados	Amer	76,835	16173.65009	1.40								
15	Belarus	EUCA	70.349	13224.02729	.30								
16	Belgium	EUCA	80.009	33047.96242	.20								
17	Belize	Amer	76.072	7255.00523	2.30								
18	Benin	SSA	56.081	1457.21565	1.20								
19	Bhutan	SAs	67.185	5640.90457	.20								
20	Bolivia	Amer	66.618	4171.36821	-20								
21	Bosnia_and_Herzegovina	EUCA	75.670										
22	Botswana	SSA	53.183	13625.11538	24.80								
23	81924	Amer	73,488	10373,40637	.45								
24	Brunei	EAP	78.005				_						
Data Mase	Variable View												
Data view	Paradic field												
								IBM SP	SS Statistics F	Processor is r	sady	Unicode:ON	

The goal is to investigate the relationship between 'GDP' and 'LifeExp' (both of them are quantitative). First we get basic statistics of 'GDP' variable in a usual way (Analyze > Descriptive Statistics > Frequencies):

Statistics						
GDP (\$ US, i	n 2000)					
N	Valid	147				
	Missing	50				
Mean		12325.31865				
Percentiles	25	2033.232880				
	50	6969.563610				
	75	17329.59545				

We can get boxplots of 'LifeExp' for four quarters of 'GDP' as we did in the first section of this document. Hence we need to construct a new categorical variable taking values 1,2,3,4 corresponding to which quarter, GDP observation belongs to. We will use the quantiles from the above table. Go to **Transform** > **Compute variable**, we call the target variable 'GDP.quart' and enter the following expression:

ta	Compute Variable				
Target Variable: GDP quart Type & LabeL_ Country Region Use Expectancy (yea.	Numgin: Dipression (COP = 2033) = 21(2033 COP) & (CDP+6960)(=31)(6666 - CDP) & (CDP+17329))+41(17329 - CDP) (4) Pandeo group: All				
CoPer (SUS in 2000)					
	CK Paste Reset Cancel Help				

Note that & means 'and'. Therefore if GDP is less than the first quartile we get 1 if it is bigger than the first quartile but less than the median we get 2 and so on. Click **OK**, and a new column appears:

ta 👘			*LifeExpCom	plete.sav [DataSet3]	- IBM SPS	S Statistics Da	ta Editor						🔊 🛛 🗶
Ele Edit	View Data Transform Analyze D	irect Marketing Graph	is <u>U</u> tilites (Add-gns Window	Help								
🗃 H	🖨 💷 🖛 🛥 🦉	📥 📰 🎼	M 👬	🖬 🔚 🐴 I	1	4 💊 🌑	Alle						
1: GDP.quart												Visible: 6 of	6 Variables
	Country	Region	LifeExp	GDP	HIV	GDP.quart	var	var	var	var	var	var	var
1	Afghanistan	SAs	48.673										-
2	Albania	EUCA	76.918										
3	Algeria	MENA	73.131	6406.81662	.10	2.00							
4	Angola	SSA	51.093	5519.18318	2.00	2.00							
5	Argentina	Amer	75.901	15741.04577	.50	3.00							
6	Armenia	EUCA	74.241	4748.92858	.10	2.00							
7	Aruba	Amer	75.246										
8	Australia	EAP	81.907	34642.38813	.10	4.00							
9	Austria	EUCA	80,854	36871.07135	.30	4.00							
10	Azerbaijan	EUCA	70.739	9387.38950	.10	3.00							
11	Bahamas	Amer	75.620	21470.80420	3.10	4.00							
12	Bahrain	MENIA	75.057										
13	Bangladesh	SAs	68.944	1669.87427	.06	1.00							
14	Barbados	Amer	76.835	16173.65009	1.40	3.00							
15	Belarus	EUCA	70.349	13224.02729	.30	3.00							
16	8elgium -	EUCA	80.009	33047.96242	.20	4.00							
17	Belize	Amer	76.072	7255.00523	2.30	3.00							
18	Benin	SSA	56.081	1457.21565	1.20	1.00							
19	Shutan	SAs	67.185	5640.90457	.20	2.00							
20	Bolivia	Amer	66.618	4171.36821	.20	2.00							
21	Bosnia_and_Herzegovina	EUCA	75.670										
22	Botswana	SSA	53.183	13625.11538	24.80	3.00							
23	Brazil	Amer	73,488	10373,40637	.45	3.00							
24	Brunei	EAP	78.005										-
Data View	variable view												
								IBM SPSS S	tatistics Proc	essor is ready	Un	lcode:ON	

Click on the 'Variable View' button and make sure that the 'Measure' for this new variable is 'Nominal' since we want SPSS to treat this variables as categorical.

	*LifeExpComplete.sav [DataSet3] - IBM SPSS Statistics Data Editor											
Ē	ile <u>E</u> dit	<u>V</u> iew <u>D</u> ata	Transform	<u>Analyze</u> I	Direct <u>M</u> arketi	ing <u>G</u> raphs <u>U</u>	tilities Add- <u>o</u> n	is <u>W</u> indow	Help			
	😂 H 🖨 💷 🖛 🛥 🗷 📥 💷 P 🖬 💷 🖾 📟 🖧 🚟 📲 🔕 🌑 🤲											
		Name	Туре	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
	1	Country	String	30	0		None	None	30	🎫 Left	🗞 Nominal	🔪 Input
	2	Region	String	15	0		None	None	15	📑 Left	🗞 Nominal	🔪 Input
	3	LifeExp	Numeric	10	3	Life Expectancy	None	None	10	🗃 Right	🛷 Scale	🔪 Input
	4	GDP	Numeric	15	5	GDP (\$ US, in 2	None	None	15	温 Right	🛷 Scale	🔪 Input
	5	HIV	Numeric	7	2		None	None	7	🕮 Right	🛷 Scale	🔪 Input
	6	GDP.quart	Numeric	8	2		None	None	11	🖷 Right	💰 Nominal 💌	🔪 Input
	7										🛷 Scale	
	8										d Ordinal	
	9										🕹 Nominal	

Now we can easily make a joint chart of boxplots for each quarter as explained in the first section.



However we can also make a simple scatterplot of two quantitative variables, go to Chart Builder > Scatter/Dot > double click on Simple Scatter



Afterwards drag 'GDP' to the X-axis and 'LifeExp' to Y-axis, click **OK** to produce the plot. If we want to add a line of best fit to the plot, then double click on the plot and click on 'Add Fit Line at Total'



Immediately the line appears with the equation of this line, we can remove this equation if we deselect the 'Attach label to line' option

Properties ×						
Chart Size Lines Fit Line Variables						
Display Spikes						
Fit Method						
💓 🖲 Linear 📈 🔘 Cubic						
Loess						
% of points to fit: 50						
Kernel: Epanechnikov 👻						
Confidence Intervals						
<u>N</u> one	l					
© M <u>e</u> an	l					
© Individual	l					
%: 95	l					
Attach label to line						

Also in the above 'Properties' window you can click on the 'Lines' button and make changes to this line (style and color). To get correlation statistic between 'GDP' and 'LifeExp' go to **Analyze** > **Correlate** > **Bivariate**

Analyze	Direct <u>M</u> arketing	Graphs	Utilities	Add- <u>o</u> ns
Repo	irts	•	¥.	
Desc	riptive Statistics			
Table	s	•		
Com	pare Means		LifeExp	
Gene	ral Linear Model	•	48.673	
Gene	ralized Linear Mode	ls 🕨	76.918	
Mi <u>x</u> eo	Models	•	73.131	
Corre	late	•	Bivariat	e
Regr	ession	+	Partial.	
Logli	near		Distanc	
Neur	al Natworks		Distance	· · · ·

Send these two variables across; make sure 'Pearson' option is selected

ta I	Bivariate Correlations	×			
✔ HIV ♣ GDP.quart	Variables:	Options Style Bootstrap			
Correlation Coefficients					
Test of Significance I jwo-tailed O One-tailed					
✓ Flag significant correlations OK Paste Reset Cancel Help					

Click **OK** and the table is printed in the output window:

	Correlations		
		Life Expectancy (years)	GDP per capita (\$ US, in 2000)
Life Expectancy (years)	Pearson Correlation	1	.635
	Sig. (2-tailed)		.000
	N	197	147
GDP per capita (\$ US, in	Pearson Correlation	.635	1
2000)	Sig. (2-tailed)	.000	
	N	147	147

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

We need only one number from this table, it is 0.635 which is correlation. Note: the correlation is positive and therefore the line of best fit increases. Similarly we make a scatterplot and find correlation for 'LifeExp' versus 'HIV':



In this case we observe negative association. Finally we move to the 'Skeleton' data set, as before we make a scatterplot (with the line of best fit) and find correlation for 'DGerror' versus 'BMIquant' variable:



In this example correlation is positive and therefore there is a positive association between 'BMIquant' and 'DGerror'.