

Summarizing Data in SPSS

This document describes how to find different descriptive statistics like mean, median, standard deviation and much more in SPSS. We also show how to make different kinds of plots for quantitative and categorical variables.

For this document we need 'Skeleton', 'Life Expectancy' and 'NY Red Bull Salaries' 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).

1 Five Number Summary

In this part we work with the 'Life Expectancy' data which shows for different countries the average life expectancy. These data contain five variables: Country, Region, LifeExp, GDP and HIV.

First we want to analyse 'LifeExp' variable. To make a very simple plot of this variable go to: Graphs > Chart Builder

ta 🛛			*LifeExpCorr	plete.sav [DataSet1]	- IBM SPS	S Statistics	Data Edito	r					- 🔿 🗙
Elle Edit	View Data Transform Analyze D	irect Marketing	Graphs Utilities	Add-ons Window	Help								
		🛃 🚽 🖬 🖬	👔 Chart Builder				ABC						
			Graphboard Tem	state Chooser									
1:LifeExp	48.673		Compare Subgro	adr								Visible	: 5 of 5 Variables
	Country	Region	Regression Varia	ble Plots	HIV	var	var	var	var	var	var	var	var
1	Afghanistan	SAs	Legacy Dislogs	+									<u> </u>
2	Albania	EUCA	191211										
3	Algeria	MENA	73.13	6406.81662	.10								
- 4	Angola	SSA	51.09	5519.18318	2.00								
5	Argentina	Amer	75.90	15741.04577	.50								
6	Armenia	EUCA	74.24	4748.92858	.10								
7	Aruba	Amer	75.24										
8	Australia	EAP	81.90	34642.38813	.10								
9	Austria	EUCA	80,85	36871.07135	.30								
10	Azerbaijan	EUCA	70.73	9387.38950	.10								
11	Bahamas	Amer	75.62	21470.80420	3.10								
12	Bahrain	MENA	75.05										
13	Bangladesh	SAs	68.94	1669,87427	.06								
14	Barbados	Amer	76.83	16173.65009	1.40								
15	Belarus	EUCA	70.34	13224.02729	.30								
16	Belgium	EUCA	80.00	33047.96242	.20								
17	Belize	Amer	76.07	7255.00523	2.30								
18	Benin	SSA	56.08	1457.21565	1.20								
19	Bhutan	SAs	67.18	5640.90457	-20								
20	Bolivia	Amer	66.61	4171.36821	.20								
21	Bosnia_and_Herzegovina	EUCA	75.67										
22	Botswana	SSA	53.18	13625.11538	24.80								
23	Brazil	Amer	73,48	10373.40637	.45								
24	Brunei	EAP	78.00					-		1	1	1	
	Control to Minure												
Data View	variable view												
Chart Builder								IBM SI	PSS Statistics	Processor is	ready	Unicode:ON	

Then choose Scatter/Dot > double click on Simple Scatter to get:



Then drag 'Life Expectancy' variable to the Y-axis and 'Country' to X-axis, click \mathbf{OK} to get the following plot in the Output window:



To make any changes to this plot like colors and other properties, **double click** on this plot leads to the 'Chart Editor'. To change the background color from grey to white (for example), **double click** on the background of the plot and get:



Choose white color, > Apply > close Chart Editor and get a new plot:



We can make this plot wider by dragging the right side of the plot to the right:



This was a very simple plot. Boxplots however produce more informative plots. From boxplot we can estimate the maximum, minimum, median, first quartile and third quartile. To draw the boxplot we start once again from 'Graphs' drop-down menu: **Graphs** > **Chart Builder** > **Boxplot** > **double click 1-D Boxplot**



As before drag 'Life Expectancy' to X-axis and click **OK**, the boxplot is produced. **Double click** on this plot opens 'Chart Editor'



If we want to find basic descriptive statistics of the 'LifeExp' variable (like minimum, maximum, mean, median and others), we can do that by right click on the header of this variable > Descriptives Statistics

Country	Region	LifeExp	GDP
Afghanistan	SAs	48-4	572
Albania	EuCA	7	Cut
Algeria	MENA	7.	Copy
Angola	SSA	5	Paste
Argentina	Amer	7	Clear
Armenia	EuCA	7.	Descriptives Statistics
Aruba	Amer	7	Grid Font
Australia	EAP	8	Spelling

Then get the following table in the 'Output' window:

Frequencies

	Statistics							
Life Ex	pectancy (ye	ars)						
N	Valid	197						
	Missing	0						
Mean		69.86282						
Media	n	73.23500						
Std. D	eviation	9.668736						
Range		35.600						
Minim	um	47.794						
Maxim	um	83.394						

It shows that total number of observations is 197, no missing values and other different quantities. Note however, this table does not show first and third quartiles. To get these measures go to Analyse > Descriptive Statistics > Frequencies

<u>A</u> nalyze	Direct Marketing	<u>G</u> raphs	Utilities	Add-ons y
Repo	rts		*	117 III
Desc	riptive Statistics	- F	122 Erequ	encies
Ta <u>b</u> le	s	•	Desc	riptives
Com	pare Means		A Explo	re
Gene	ral Linear Model		Cros.	stabs
Gene	ralized Linear Mode	ls 🕨	TUR	Analysis
Mixed	Models		Ratio	, and joint
Corre	late			
Regn	ession		P-PF	1015
Logli	near		0-Q F	Plots

Then select the 'Life Expectancy' variable and then click on the arrow to move this variable to the right window.



Then click on **Statistics**

End Frequencies: Sta	itistics 🗶
-Percentile Values Quartiee Quipoints for: 10 equal groups Bernove Bernove	Central Tendency
Dispersion ♥ Std. devlation ♥ Minimum ■ Yariance ♥ Magimum ♥ Range ■ S.E. mean	Vajues are group midpoints Distribution Skewness Kurtosis
Continue Cancel	Help

Here select all the measures that we want. You see that now we have 'Quartiles' option for the first quartile, median and third quartile. To finish click **Continue** > **OK** to get the next table:

Frequencies

5	Statistics							
Life Expectancy (years)								
N	Valid	197						
	Missing	0						
Mean		69.86282						
Median		73.23500						
Std. Deviation		9.668736						
Range		35.600						
Minimum		47.794						
Maximum		83.394						
Percentiles	25	64.44700						
	50	73.23500						
	75	76.74350						

Here 'Percentiles 25' is the first quartile, 'Percentiles 50' and 'Percentiles 75' correspond to the median and third quartile respectively.

2 Center of Data

This section we are going to start with the Skeleton data.

ta 🛛					SkeletonData	Complete.sav	/ [DataSet1] - IB	BM SPSS Stati:	stics Data	a Editor					- 0	×
Elle Edit	View Da	ta <u>T</u> ransform <u>A</u>	nalyze Direct	Aarketing	Graphs Utilities	Add-ons V	Vindow <u>H</u> elp									
			a 🛛 🖬 🕹	_	TRI del 🖗		🔤 🗠 🎹			ABS						
							·····	14 🕤		•						
														Vis	ible: 8 of 8 Va	iriables
	Sex	BMIcat	BMIquant	Age	DGestimate	DGerror	S8estimate	SBerror	var	var	var	var	var	var	var	
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								- 1
5		1 normal	21,41	60	32	-28	46	-14								
6		1 underweight	13.65	34	25	-9	35	1								
7		1 overweight	25.86	50	32	-18	35	-15								
8		1 underweight	14.56	73	50	-23	61	-12								- 11
9		1 normal	22,44	70	39	-31	46	-24								- 11
10		1 normal	19.88	60	44	-16	46	-14								
11		1 normal	23.24	58	32	-26	35	-23								
12		1 overweight	25.09	61	32	-29	61	0								- 11
13		2 overweight	25.68	52	44	-8	48	-4								
14		1 normal	24.97	67	44	-23	46	-21								- 11
15		1 normal	23.32	60	44	-16	46	-14								- 11
16		1 normal	23.29	68	50	-18	61	-7								
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.														
										IBI	I SPSS Statist	ics Processor	is ready	Unicode	ON	

These data have 400 observations of skeletons, and we are interested in the error variable ('DGerror') which measures the difference between estimated and actual age using the method of Di Gangi.

We want to plot the modified boxplot of the 'DGerror' variable (note that SPSS only produces modified boxplot that shows observations beyond fences), go to **Graphs** > **Chart Builder**



Then select **Boxplot** > double click 1-D Boxplot



As before drag the 'DGerror' variable into the X-axis of the boxplot. If you want to make a title, click on the **Titles/Footnotes** tab, then tick 'Title 1' and enter a title in the 'Content' window on the right side:

t a	Chart Builder	×	Element Properties	×
yarables:	Chartpreview uses example date		Egé Poportes of. Bert X-Alis (Bort) Title 1 Ten Type: Title 1 Gentent Modified Boxple(
Galley Basic Elements Checked Items add Miles Title 1 Subtile Ecothole 1 Foothole 2	Grouppafrant,D TBissFocholes and footholes b the chat Edit the ted within Properties OK Paste Beset Cancel Help	Element Properties_ Options	(450/) Cancel Help	

Click $\mathbf{Apply} > \mathbf{OK}$ to get the plot:



Note that now we have a title and some observations are beyond the fences.

As before to get some statistics of this variable go to $\mathbf{Analyze} > \mathbf{Descriptive Statistics} > \mathbf{Frequencies}$

Analyze	Direct Marketing	Graphs	Utilities	Add-ons	1
Repo	irts	•	4 🖁		E
Desc	riptive Statistics	•	122 Frequ	encies	
Table	15		Desc	riptives	٦
Com	pare Means		A Explo	re	
Gene	ral Linear Model	•	Cros:	stabs	
Gene	ralized Linear Mode	ls ▶	TUR	Apolysis	
Mixed	Models			Analysis	
Corre	elate		Matto		
Regr	ession		<u>Р</u> -Р Р	lots	
Logli	near		🛃 Q-Q F	Plots	

Select 'Est. - Act. using D (years)' variable and click on the arrow to send this variable across:



Click on the 'Statistics' button and choose all the necessary quantities:



Continue > **OK** to produce the next table of statistics:

Frequencies

	Statistics								
Est Act. us	Est Act. using D (years)								
N	Valid	400							
	Missing	0							
Mean		-14.15							
Range		92							
Minimum		-60							
Maximum		32							
Percentiles	25	-23.00							
	50	-13.00							
	75	-5.00							

Next open another data set which consists of salaries for NY Red Bulls players. These data consist of only two variables. First variable is the original data of salaries, the second one is the salaries divided by thousand to make graphs scale better. To make a good visual representation of these data we use boxplot as before. To make a different label to the vertical axis, click on X-Axis1(Box1) in the 'Element Properties' window:

G	Chart Builder	×	ta Ele	ment Prope	erties	×
Variables:	Chart preview uses example data		Egit Properties of: Box1 X-Axis1 (Box1)	ilaries in Tho	io ande	×
No categories (scale variable)			Scale Range Variable: PNY Minimum Magimum Major Incremen	Salaries in Tr Automatic	Custom	
Collery Choose from: Favorites Bar Line Area PiePolar Scate/Dot Histogram High-Low Berplot Dual Area		lemenţ roperties Options	Scale Type Type:	Linear Base: Exponent:	10	
	OK Paste Reset Cancel Help		Appl	Close	Help	

'Axis Label' now shows the usual label for this variable. You can easily change it to any other label; as usual Apply > OK to finish:



We immediately see two outliers that probably will affect the mean of this sample. To get some simple statistics of the 'NYSalary' variable right click of the header of this variable > Descriptives Statistics

NYSa	NIVS al in Th
33	Cuț
44	Copy
138	Paste
45	Clear
44	🖬 Insert Variable
141	Sort Ascending
292	Sort Descending
5600	Descriptives Statistics
103	
	= Spelling

And get the following table in the 'Output':

Frequencies

Statistics							
NY Red Bull Salaries							
N Valid	25						
Missing	0						
Mean	518311.6352						
Median	112495.5000						
Std. Deviation	1388822.106						
Range	5566250.00						
Minimum	33750.00						
Maximum	5600000.00						

As expected there is a big difference between mean and median. Hence we want to trim the data and find a new mean. To do that copy the first column by **right click on the header** > **Copy**

NYSal	NIVSal in Th
337	Cut
440	<u>С</u> ору
1381	Paste
455	Clear
440	Insert Variable
1416	Sort Ascending
2925	Sort Descending
56000	Gon Descending
1035	Descriptives Statistics
1900	Spelling

Copy it to the third column and name this column 'NYSalary.Trim'. To trim the variable by 8% we need to delete 2 largest and 2 smallest observations (since 25 times 8% is 2). Hence we need to sort the variable first by **right click on the header** > **Sort Ascending**

	NYSalary	NYSal.in.Th	NYSalary.	trim yor yor
1	33750.00	33.75	3375	Cuț
	44000.00	44.00	4400	Copy
	138188.00	138.19	13818	Paste
	45566.67	45.57	4556	Clear
	44000.00	44.00	4400	🔣 Insert Variable
	141666.67	141.67	14166	Sort Ascending
	292500.00	292.50	29250	Sort Descending
	560000.00	5600.00	560000	
	103500.00	103,50	10350	Descriptives Statistics
	100000.00	100.00	10000	Spelling
	190000.00	190.00	19000	

Then just remove two smallest and two largest observations manually:

th i	*NYRedBullsSalaries.av (DataSett) - IBM SPSS Statistics Data Editor – 🧳 💌																	
File	Edit	View	Data	Transform An	alyze Direct Mark	eting Grap	hs <u>U</u> tilite	s Add-ons	Window	Help								
6) 🖹 📥 🗄		H 🗄	s 🖌	-	III 1	101	► 4 6						
23 :																	Visible:	3 of 3 Variables
		NYSal	ary	NYSal.in.Th	NYSalary.Trim	var	var	var	var	var	var	var	var	var	var	var	var	var
	1	337	50.00	33.7	·5 .													-
	2	337	50.00	33.7	's .													
	3	337	50.00	33.7	5 33750.00													
	4	337	50.00	33.7	5 33750.00													
	5	440	00.00	44.0	44000.00													
	6	440	00.00	44.0	44000.00													
	7	440	00.00	44.0	44000.00													
	8	440	00.04	44.0	44000.04													
	9	455	66.67	45.5	45566.67													
	10	650	00.00	65.0	65000.00													
	11	950	00.00	95.0	95000.00													
	12	1035	00.00	103.5	103500.00													
	13	1124	95.50	112.5	112495.50													
	14	1381	88.00	138.1	9 138188.00													
	15	1410	66.67	141.6	141666.67													
	16	1815	00.00	181.5	181500.00													
	17	1850	00.00	185.0	185000.00													
	18	1900	00.00	190.0	190000.00													
	19	1943	75.00	194.3	194375.00													
	20	1950	00.00	195.0	195000.00													
	21	2050	00.00	205.0	205000.00													
	22	2921	00.00	292.5	0 292500.00													
	23	3019	99.00	302.0	0 301999.00													
	24	46000	00.00	4600.0	ю.													
	_									***								
Data	s View	Variable \	tew															
													IBM SPS	S Statistics P	rocessor is re	ady	Unicode:ON	

Now we get basic statistics for this trimmed variable:

Frequencies

Statistics							
NY Red Bull Salaries							
N Valid	21						
Missing	4						
Mean	128109.0895						
Median	112495.5000						
Std. Deviation	83990.81129						
Range	268249.00						
Minimum	33750.00						
Maximum	301999.00						

See how the mean changes when we trimmed the data.

3 Spread of the Data

We start with the 'Life Expectancy' data set. We will analyse again the 'LifeExp' variable which contains the mean life expectancies values. Lets plot the boxplot of this variable again. As usual go to **Graphs** > **Chart Builder** > **Boxplot**. After selecting the 'LifeExp' variable, if you want to change the scale of the vertical axis of the boxplot, then under 'Element Properties' select 'X-Axis1 (Box1):

6	Chart Builder	×	6	Element Prop	erties	×
Variables:	Chart preview uses example data		Edit Propert Box1	ies of:		×
Region			X-Axis1 (Bo	x1)		
A HIV	Acy (9e 48)		Axis Label:	Life Expectancy (ye	ars)	
	Eogea		Variable:	Life Expectancy	(years)	
			Minimun	n 🗹	0	
No calegories (scale variable)			Ma <u>x</u> imur Major Int	n 🗹 crement 🖌	0	
			Origin		0	
Choose from:	Groups/Point ID Titles/Footnotes	Element Properties	Scale Type Type	Linear		-
Favorites Bar		Options		Base:	10	
Area Pie/Polar Scatter/Dot Histogram High-Low Boxplot Dual Axes	ġ ₩₩₩ ► ₩					
	OK Paste Reset Cancel Help			Apply Close	Help	

Then deselect the Automatic option of Maximum and Minimum and enter the appropriate values (we want the vertical axis to be from 40 to 90):

t a	Chart Builder	×	ta	Element Prop	erties	×
Variables: ♣ Country ♣ Region ✔ Life Expectancy (yea. ✔ GDP ✔ HIV	Chair preview uses example data		Edit Proper Box1 X-Axis1 (Br	ties of: ax1) : Life Expectancy (ye	ars)	×
No calegories (scale variable)			-Scale Rai Variable: Minimur Magimu Major In <u>O</u> rigin	Automatic Mutomatic m crement Z	(years) Custom 40 90 0	
Gallery Basic Element Choose from: Favorites Bar Line Area Pie/Polar Scatter/Dot	CroupsPoint D TResFochores	Element Properties Options	Scale Type	e E <mark>Linear</mark> Base: Exponent:	10	
Histogram High-Low Boxplot Dual Axes	OK Paste Reset Cancel Help			(Apply) Cancel	Help	

Finishing you get the usual plot:



To get basic statistics of this variable, as before right click on the header of 'LifeExp' > Descriptives Statistics and get the table:

Frequencies

Statistics							
Life Expectancy (years)							
N Valid	197						
Missing	0						
Mean	69.86282						
Median	73.23500						
Std. Deviation	9.668736						
Range	35.600						
Minimum	47.794						
Maximum	83.394						

As you know we have several measures of the spread of quantitative data: range, IQR and standard deviation. The above table shows only the range and standard deviation, to get also IQR (interquartile range) we can do the following: go to **Analyze** > **Descriptive Statistics** > **Explore**

Analyze	Direct Marketing	Graphs	Utilities	Add-ons	1
Repo	rts		A 🐺		
Desc	riptive Statistics	- F	122 Erequ	encies	
Table	s		Desc	riptives	
Com	pare Means		A Explo	re	
Gene	ral Linear Model		Cross	stabs	
Gene	ralized Linear Mode	ls 🕨	TURF	Analysis	
Mixed	Models		Ratio		
Corre	fate	•	B.0.0	loto	
Regr	ession		E C-P P	1010	
Loglin	near		🛃 Q-Q F	Plots	

Select 'Life Expectancy' variable from the left window and click on the arrow to move it to the right:



In the Display option below select 'Statistics' since we do not need plots here, click OK to finish the process and get this table in the Output window:

Explore

Case Processing Summary								
Cases								
Va	lid	Miss	sing	Total				
N	Percent	N	Percent	N	Percent			
197	100.0%	0	0.0%	197	100.0%			
	Ca: Va N 197	Case Processi Valid N Percent 197 100.0%	Case Processing Summary Case Case Valid Mise N Percent N 197 100.0% 0	Case Processing Summary Cases Valid Missing N Percent N Percent 197 100.0% 0 0.0%	Case Processing Summary Cases Valid Missing To N Percent N Percent N 197 100.0% 0 0.0% 197			

	Descriptive	5		
			Statistic	Std. Error
Life Expectancy (years)	Mean		69.86282	.688869
	95% Confidence Interval	Lower Bound	68.50427	
	for Mean	Upper Bound	71.22136	
	5% Trimmed Mean		70.36401	
	Median		73.23500	
	Variance		93.484	
	Std. Deviation		9.668736	
	Minimum		47.794	
	Maximum		83.394	
	Range		35.600	
	Interquartile Range		12.296	
	Skewness		832	.173
	Kurtosis		399	.345

This table contains much more information than previous table. Here we have mean, 5% trimmed mean, median, range, standard deviation and IQR. Note that standard deviation is exactly the same as square root of the variance.

Next we want to compare different measures of data (statistics) for robustness using NY Red Bull salaries. Open this data file. From the last section we remember that this variable has two large outliers, so lets investigate which measures that we have learned change a lot in the presence of outliers and which are not (robust).

First copy the 'NYSalary' variable using right click on header of 'NYSalary' > Copy:



Then right click on 'var' of the third column > Paste:

NYSalary	NYSal.in.Th	var var	var
33750.00	33.75	Cut	
44000.00	44.00	Copy	
138188.00	138.19	Paste	
45566.67	45.57	Cl <u>e</u> ar	
44000.00	44.00	🔀 Insert Varia	ble
141666.67	141.67	Sort Ascent	ting
292500.00	292.50	Ret Desce	adiaa
560000.00	5600.00	Son Descer	iong
103500.00	103.50	Descriptive	s Statistics
190000.00	190.00	Spelling	

Then make a name for this column (we call it 'NYSalary.Trim'). Next sort this variable **right** click on header > Sort Ascending:

	NYSalary.	NYSal.in.Th	NYSalary
Cut	3375	33.75	33750.00
Copy	4400	44.00	44000.00
Paste	13818	138.19	138188.00
Clear	4556	45.57	45566.67
🔽 Insert V <u>a</u> riable	4400	44.00	44000.00
Sort Ascending	14166	141.67	141666.67
Sort Descending	29250	292.50	292500.00
Descriptives Statistics	560000	5600.00	560000.00
	10350	103.50	103500.00
Shenund"	19000	190.00	190000.00

Then manually delete two largest and two smallest values. Now we have a column of the trimmed data. To find all the necessary statistics for the original variable, as before go **Analyse** > **Descriptive Statistics** > **Explore**:

<u>A</u> nalyze	Direct Marketing	Graphs	Utilities	Add-ons
Repo	irts		*	
Desc	riptive Statistics	- F	121 Frequ	encies
Table	15		Desci	iptives
Com	pare Means		A Explo	re
Gene	ral Linear Model	•	Cross	tabs
Gene	ralized Linear Mode	is 🕨	TURE	Analysis
Mixed Models		,	Ratio	,
Corre	elate	•		
Regr	ession		P-P P	iots
Logli	near		0-Q P	lots

Select 'NYSalary' variable and send it to the 'Dependent List' using arrow and select 'Statistics' in the display option:

4	Explore	· · ·			
NY Salaries in Thou NY Red Bull Salane	Factor List Factor List Factor List Factor List Subset Gases by:	Statistics Plojs Options Bootstrap			
Display					
○ Both ● Statistics ○	Piots				
OK Paste Reset Cancel Help					

Click **OK** to get the table for the original variable:

Explore

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
NY Red Bull Salaries	25	100.0%	0	0.0%	25	100.0%

Descriptives

			Statistic	Std. Error
NY Red Bull Salaries	Mean		518311.6352	277764.4211
	95% Confidence Interval	Lower Bound	-54965.9540	
	for Mean	Upper Bound	1091589.224	
	5% Trimmed Mean		274026.8169	
	Median		112495.5000	
	Variance		1.929E+12	
	Std. Deviation		1388822.106	
	Minimum		33750.00	
	Maximum		5600000.00	
	Range		5566250.00	
	Interquartile Range		150687.50	
	Skewness		3.341	.464
	Kurtosis		10.182	.902

Now do exactly the same but for the trimmed variable, and get the following table:

Explore

case Processing Summary						
		Cases				
	Va	lid	Missing		Total	
	N	Percent	N	Percent	N	Percent
NY Red Bull Salaries Trimmed	21	84.0%	4	16.0%	25	100.0%

Descriptives					
			Statistic	Std. Error	
NY Red Bull Salaries	Mean		128109.0895	18328.29764	
Trimmed	rimmed 95% Confidence Interval	Lower Bound	89876.9306		
	for Mean	Upper Bound	166341.2484		
	5% Trimmed Mean		123715.8402		
	Median		112495.5000		
	Variance		7054456381		
	Std. Deviation		83990.81129		
	Minimum		33750.00		
	Maximum		301999.00		
	Range		268249.00		
	Interquartile Range		148187.48		
	Skewness		.596	.501	
	Kurtosis		557	.972	

See that this variable has only 21 valid values and 4 are considered as missing because we have deleted 4 values. Based on these two tables we can conclude that median and IQR almost did not change after trimming the data and therefore are robust to outliers, on the other hand mean, range and standard deviation changed a lot by trimming and hence are not robust to outliers.

4 Shape of the Data

In this section we introduce histograms and investigate shapes of different variables. Lets start with plotting the histogram of the life expectancy data: go to **Graphs** > **Chart Builder** > **Histogram** > **double click on Simple Histogram**



Then drag the 'Life Expectancy' variable to the horizontal axis. Also click on the 'Titles/Footnotes' tab to give a title for this plot (same as with boxplot):



Click **Apply** > **OK** to get the histogram of this variable; this histogram is noisy because it has many bins. To change the number of bins **double click on the plot** to open the 'Chart Editor', then **double click** on the bins opens the 'Properties' window, click on 'Binning'. Select the 'Custom' option and choose the number of intervals (for example 5):



Click **Apply** and close the 'Chart Editor' to get modified histogram:



Thus we can change the number of bins to make the plot more or less noisy. As before we can get the basic statistics of the Life Expectancy variable by right click on the header of the column > Descriptives Statistics and get the table:

Frequencies

	Statistics					
Life Ex	pectancy (ye	ars)	_			
Ν	Valid	197	l			
1	Missing	0	l			
Mean		69.86282	L			
Mediar	1	73.23500	l			
Std. De	eviation	9.668736	L			
Range		35.600	l			
Minimum		47.794	l			
Maxim	um	83.394				

Since the mean is smaller than median and based on the histogram we can conclude that these data are skewed to the left (has longer left tail).

Now lets consider 'Skeleton' data set and make the histogram of 'DGerror' variable. As explained earlier go to **Graph** > **Chart Builder** > **Histogram** > **Simple Histogram**, then drag the 'Est. - Act. using D (years)' variable to the horizontal axis. The horizontal label of the histogram would be the label of this variable, if we want to change the label then under 'Elements Properties' select 'X-Axis1 (Bar1)' and change the 'Axis Label':



Do not forget to click Apply > OK to produce the histogram:



Next we find descriptive statistics of the 'DGerror' variable as usual and get:

Frequencies

Statistics						
Est Act. using D ((years)					
N Valid	400					
Missing	0					
Mean	-14.15					
Median	-13.00					
Std. Deviation	14.126					
Range	92					
Minimum	-60					
Maximum	32					

Based on the plot and since mean and median are almost the same we see that the distribution of differences is symmetric. The last data set that we will analyse in this section is 'NY Red Bull Salaries'. Plotting the histogram of these data we again notice two large outliers:



To make a histogram without these two values, first sort the original data by right click on the header > Sort Ascending

NYSala	NVSal in Th	
337	Cuț	
440	Copy	
1381	Paste	
455	Clear	
440	🔣 Insert Variable	
1416	Sort Ascending	
2925	Sort Descending	
56000	Description Obtining	
1035	Descriptives statistics	
1900	🐞 Spelling	

After the sorting, just delete two largest values:

	NYSalary	NYSal.in.Th
11	95000.00	95.00
10	102500.00	103.50
12	112405 50	113.50
15	12495.50	112.50
14	136166.00	136.19
15	141000.07	141.67
16	181500.00	181.50
17	185000.00	185.00
18	190000.00	190.00
19	194375.00	194.38
20	195000.00	195.00
21	205000.00	205.00
22	292500.00	292.50
23	301999.00	302.00
24		
25		
26		

Now make a histogram of the second trimmed column (since it is in thousands) and get:



Then get descriptive statistics of the first column (trimmed NYSalaries):

Frequencies

Statist	ics
NY Red Bull Salari	es
N Valid	23
Missing	2
Mean	119903.9513
Median	103500.0000
Std. Deviation	84570.55807
Range	268249.00
Minimum	33750.00
Maximum	301999.00

This data set has mean which is larger than median and looking on the plot we say that these data are skewed to the right (or has longer right tail).

Now lets return to the 'DGerror' variable, and we want to check that it follows the empirical rule. Since the distribution of this variable is unimodal and symmetric we expect to get close to theoretical results. First we find mean and standard deviation of the error variable using descriptive statistics:

Frequencies

Statistics						
Est Act. using D	(years)					
N Valid	400					
Missing	0					
Mean	-14.15					
Median	-13.00					
Std. Deviation	14.126					
Range	92					
Minimum	-60					
Maximum	32					

Note that the mean is -14.15 and standard deviation is 14.126. Next we want to see which observations are between mean plus/minus one standard deviation and which are not. To do that go to **Transform** > **Compute variable**:



Here we construct a new 'Target variable' (which we name 'Between.1.sd'). In the numeric expression we enter the logical equation which takes two values: 1 if expression is true and 0 if it is false. We get 1 if 'DGerror' is greater than -14.15 - 14.126 (mean - one standard deviation) and (& means and) smaller than -14.15 + 14.126 (mean + one standard deviation) otherwise

zero. If both of these conditions hold then the value of the new variable is going to be 1 and these conditions are equivalent for 'DGerror' to be between mean and plus/minus one standard deviation.

ta	Compute Variable				
Jorde Vansbe Beheers 1.5d Tors 4.2bel.	Numgric Expression: DCarror = (-14.15 - 14.120) & (DCarror < (-14.15 + 14.120))				
(optional case selecti	Abo Any Any Asymotol Asimotol Attan				
OK Paste Reset Cancel Help					

Click on the \mathbf{OK} button and get a new column of 1's and 0's:

ta 🛛	"SkeletonDataComplete.sav (DataSet1) - IBM SPSS Statistics Data Editor – 🗸 🖉 💌																
Ele Edit	New	Data	Transform	Analyze Dire	t Marketing	Graphs U	titles	Add-gns	Window Help								
🔁 h		00	5	× 🖹	- = I I	r M	*,	<u>×</u>	- 🖧 🔜	1	AES						
1: Between.	Between 1 sd 0 Visible: 9 of 9 Variables																
	Sex		BMIcat	BMIquant	Age	DGestim	ate	DGerror	S8estimate	SBerror	Between 1.sd	var	var	var	var	var	var
1	1	2 L	inderweight	15.6	5 78	3	44	-3-	6	-18	.00						-
2		1.0	ormal	23.0	3 44		32	-1	3	5 -9	1.00						
3		1 0	werweight	27.9	2 72	2	32	-4	6	-11	.00						
- 4		1 0	werweight	27.8	3 59	•	44	-1	i 6	2	1.00						
5		1 1	iormal	21,4	60	0	32	-2	4	-14	1.00						
6		1.4	inderweight	13.6	5 34	1	25	-	3	5 1	1.00						
7		1 0	werweight	25.8	5 50)	32	-1	3	5 -15	1.00						
8		1 נ	inderweight	14.5	5 73	3	50	-2	6 6	-12	1.00						
9		1 1	iormal	22.4	4 70	0	39	-3	4	5 -24	.00						
10		1 r	iormal	19.8	3 60		44	-1	5 4	5 -14	1.00						
- 11		1 r	ormal	23.2	4 58	3	32	-2	3	5 -23	1.00						
12		1 0	werweight	25.0	9 61	1	32	-2	6 6	0	.00						
13		2 0	werweight	25.6	3 52	2	44	-	4	3 -4	1.00						
14		1 r	iormal	24.9	7 67	r	44	-2	4	-21	1.00						
15	_	1 1	iormal	23.3	2 60	0	44	-1	5 4	5 -14	1.00						
16		1 1	iormal	23.2	68	3	50	-1	6	-7	1.00						
17		2 0	werweight	27.3	7 35	5	12	-2	3	3 3	1.00						
18	_	2 0	ibese	34.8	2 81		39	-4	2 4	-33	.00						
19		2 L	inderweight	12.2	73	8	44	-2	6	-13	.00						
20		1 1	iormal	23.8	0 02	>	39	-0	4	-19	1.00						
21		1 1	iormal	24.8	9 57		57		4	5 -11	.00						
22		2 1	iormai	24.6	67		32	-3	6	-7	.00						
23		2 1	iormai	23.1	3 60		44	-1	6	0	1.00						-
24	4	1.1	ormai	24.7	35		32		3	, o	1.00						- N
Data View	Variable	view															
	_	_										IBM SPSS S	Statistics Proce	essor is ready	Uni	code:ON	

Then to find proportion of values between mean plus/minus one standard deviation we just find descriptive statistics of the new variable ('Between.1.sd'):

Frequencies

Statistics					
Betw	een.1.sd				
Ν	Valid	400			
	Missing	0			
Mode	9	1.00			
Rang	je	1.00			
Minin	num	.00			
Maxi	าานทา	1.00			

				sd		
			Frequency	Percent	Valid Percent	Cumulative Percent
7	Valid	.00	127	31.8	31.8	31.8
		1.00	273	68.3	68.3	100.0
		Total	400	100.0	100.0	

We see that 68.3% of observations are within this range. Similarly we do it for two standard deviations. Construct a new variable:

th	Compute Variable	×			
Turget Vanadas: Settive na. 2.ed Settive na. 2.ed Tree & Labet: The boly mass ind The boly mass ind The boly mass ind The boly mass ind Eat. Age using D (L.) Eat. Age using Set. Between 1.sd	Compute Variable Numget Egression: [] () Deares < (-14 15 - 2*14 128)) & (Deares < (-14 15 - 2*14 128))	×			
Any Any Applymodel Asia Asia Asia Asia Asia Asia Asia Asia					
OK Paste Reset Cancel Help					

Then get frequencies for this new variable:

Frequencies

Statistics						
Between.2.sd						
N	Valid	400				
	Missing	0				
Mode		1.00				
Rang	le	1.00				
Minin	num	.00				
Maxir	num	1.00				

	Between.2.sd						
		Frequency	Percent	Valid Percent	Cumulative Percent		
Valid	.00	20	5.0	5.0	5.0		
	1.00	380	95.0	95.0	100.0		
	Total	400	100.0	100.0			

We see that 95% are between mean and plus/minus two standard deviations. For the three standard deviations:

Frequencies

Statistics						
Betw	een.3.sd					
Ν	Valid	400				
	Missing	0				
Mode		1.00				
Rang	e	1.00				
Minin	num	.00				
Maxin	num	1.00				

	Between.3.sd							
		Frequency	Percent	Valid Percent	Cumulative Percent			
Valid	.00	3	.8	.8	.8			
1	1.00	397	99.3	99.3	100.0			
	Total	400	100.0	100.0				

Hence 99.3% are between mean plus/minus three standard deviations. The proportions that we get are very close to theoretical empirical rule!

5 Categorical Variables

In this section we focus on the visual representation of categorical variables. We start with the variable 'Region' from Life Expectancy data file that shows to which regions different countries belong. The variable of regions is of course categorical with six classes. First we make a table of counts and frequencies, **right click on Region header** > **Descriptives Statistics**

Country	Region	LifeEuro
Afghanistan	SAs	Cuț
Albania	EuCA	Copy
Algeria	MENA	Paste
Angola	SSA	Cl <u>e</u> ar
Argentina	Amer	Insert Variable
Armenia	EuCA	Sort Ascending
Aruba	Amer	Sort Descending
Australia	EAP	Descriptions Clatistics
Austria	EuCA	Descriptives Statistics
Azerbaijan	EuCA	5 Spelling

Then get the next table:

Frequencies

Statistics					
Regio	n				
N	Valid	197			
	Missing	0			

Region

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Amer	39	19.8	19.8	19.8
	EAP	30	15.2	15.2	35.0
	EuCA	50	25.4	25.4	60.4
	MENA	21	10.7	10.7	71.1
	SAs	8	4.1	4.1	75.1
	SSA	49	24.9	24.9	100.0
	Total	197	100.0	100.0	

The first column of 'Frequencies' shows how many observations are in each category while the second 'Percent' column shows the relative frequency which is just count divided by total number of observations.

Next we want to make a visual representation of this variable. We start with a bar-plot of counts: Graphs > Chart Builder > Bar > double click on Simple Bar:



Then drag 'Region' variable to the horizontal axis, 'Count' in vertical axis appears automatically:



As before you can add a title to this plot and then click **OK** to get the bar-plot of counts:



To make changes to the plot, **double click** on it to open 'Chart editor'. If you want to change the bar labels (now they are 'Amer', 'EAP',,) just **double click** on any of these labels and type other descriptions. To change color of the bars **double click** on any bar of the plot to open the 'Properties' window and select 'Fill & Border':



Now you can change the color then click **Apply** and close the 'Chart Editor'. To make a bar-plot of relative frequencies, as before **Graphs** > **Chart Builder** > **Bar** > **double click on Simple Bar**, drag 'Region' to the horizontal axis, open the 'Elements Properties' window and under 'Statistic' select 'Percentage' instead of 'Count':



Finish with $\mathbf{Apply} > \mathbf{OK}$ and get:



Now we see that vertical axis is 'Percent' which is just relative frequency.

Another useful plot for categorical variables is a pie chart. Go to **Graphs** > **Chart Builder** > **Pie/Polar** > **double click on Pie Chart**, then drag 'Region' variable to the horizontal axis:



Click **OK** to get the Pie chart in the output window. To change the color of each individual slice, **double click** on the plot to open the 'Chart Editor' then **double click** on the slice we want change to open the 'Properties' window then select 'Fill & Border' and change the color:



Now we return to the 'Skeleton' data and first analyse 'Sex' variable. Using the same procedure as before we make a bar-chart for counts. To make a title click on 'Titles/Footnotes' in the 'Chart Builder' window, select 'Title 1' and enter the title in the 'Content' window on the right:



Click Apply > OK to get the bar-chart for the 'Sex' variable:



Similarly we make a bar-chart of relative frequencies and a Pie chart:



Similarly we do for the 'BMIcat' variable from 'Skeleton' file. First get the bar-chart for counts:



We see however that order of these bars is not logical. We probably want them to be from left to right: 'underweight', 'normal', 'overweight' and 'obese'. To change the order **double click** on the plot to open the 'Chart Builder', then **double click** on any bar to open the 'Properties' window and select 'Categories':



Now under 'Order' select labels and move them with the arrows to right positions. Then click **Apply**, close the 'Chart Editor' and get a modified bar chart. Similarly we can do for the relative frequencies



Finally we get the Pie chart for this variable; to change the order of slices **double click** on the chart, **double click** on any slice and select 'Categories':



Change the order and get the following pie chart with better ordering:



