How can I analyze my data by categories?

Authored by

stats writer

June 30, 2024

Analyzing data by categories refers to the process of organizing and examining data based on specific characteristics or groups. This can be achieved by first identifying the relevant categories for the data, such as demographic information or product features. Then, the data can be sorted and grouped into these categories to gain a better understanding of trends, patterns, and relationships within the data. This method of analysis can provide valuable insights and inform decision-making in various fields, such as marketing, research, and business. To effectively analyze data by categories, it is important to have a clear understanding of the data and its categories, as well as the appropriate tools and techniques for data analysis.

How can I analyze my data by categories? | SPSS FAQ

Sometimes you may want to analyze your data based on
categories or a grouping variable. One way that you could do this is to split the data file into different data files and
conduct the same analyses on the two (or more) data sets. However, that is cumbersome and error prone. Several
commands in SPSS will allow you to do separate analyses by category, and we will consider them below.

Let's use the example data set below. You will notice that one
of the independent variables, iv1, is a string variable. We will use this variable as our grouping variable to demonstrate how
to use a string variable as the grouping variable. All of

the techniques that will be shown can be used with a numeric
categorical variable as well.

```
data list list / sub * iv1 (A) iv2 * dv1 dv2.
begin data
1 "1" 1 48 25
2 "1" 1 49 37
3 "1" 1 50 55
4 "2" 1 17 19
5 "2" 1 20 38
6 "2" 2 23 48
7 "2" 2 28 44
8 "3" 2 28 68
9 "3" 2 30 30
10 "3" 2 32 37
end data.
```

To begin with, suppose we wanted to find the mean and standard
deviation for dv1 for groups one, two and three in iv1. We can use
the means command to obtain simple descriptive statistics.

means tables= dv1 by iv1.

	Cases					
Included	Excluded	Total				
N	Percent	N	Percent	N	Percent	
DV1 * IV1	10	100.0%	0	.0%	10	100.0%

IV1	Mean	N	Std. Deviation
1	49.0000	3	1.00000
2	22.0000	4	4.69042
3	30.0000	3	2.00000
Total	32.5000	10	12.25878

You could also use the examine command, as shown below. We will use the plot = none subcommand to suppress the stem-and-leaf and boxplots.

examine dv1 by iv1
/plot = none.

	Cases					
Valid	Missing	Total				
N	Percent	N	Percent	N	Percent	
DV1	10	100.0%	0	.0%	10	100.0%

	Statistic	Std. Error	
DV1	Mean	32.5000	3.87657
95% Confidence Interval for Mean	Lower Bound	23.7306	

Upper Bound	41.2694		
5% Trimmed Mean	32.3889		
Median	29.0000		
Variance	150.278		
Std. Deviation	12.25878		
Minimum	17.00		
Maximum	50.00		
Range	33.00		
Interquartile Range	26.0000		
Skewness	.516	.687	
Kurtosis	-1.278	1.334	

	Cases						
Valid	**Missing**	**Total**					
	IV1	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**
DV1	**1**	3	100.0%	0	.0%	3	100.0%
2	4	100.0%	0	.0%	4	100.0%	
3	3	100.0%	0	.0%	3	100.0%	

	IV1	**Statistic**	**Std. Error**	
DV1	**1**	**Mean**	49.0000	.57735
95% Confidence Interval for Mean	**Lower Bound**	46.5159		
Upper Bound	51.4841			
5% Trimmed Mean	.			
Median	49.0000			
Variance	1.000			
Std. Deviation	1.00000			
Minimum	48.00			
Maximum	50.00			
Range	2.00			

Interquartile Range	.			
Skewness	.000	1.225		
Kurtosis	.	.		
2	**Mean**	22.0000	2.34521	
95% Confidence Interval for Mean	**Lower Bound**	14.5365		
Upper Bound	29.4635			
5% Trimmed Mean	21.9444			
Median	21.5000			
Variance	22.000			
Std. Deviation	4.69042			
Minimum	17.00			
Maximum	28.00			
Range	11.00			
Interquartile Range	9.0000			
Skewness	.543	1.014		
Kurtosis	-.153	2.619		
3	**Mean**	30.0000	1.15470	
95% Confidence Interval for Mean	**Lower Bound**	25.0317		
Upper Bound	34.9683			
5% Trimmed Mean	.			
Median	30.0000			
Variance	4.000			
Std. Deviation	2.00000			
Minimum	28.00			
Maximum	32.00			
Range	4.00			
Interquartile Range	.			
Skewness	.000	1.225		
Kurtosis	.	.		

**Now let's a technique that is more general and that can be
used with any type of analysis. First, we need to sort the data by by our grouping variable, in this case,
iv1. Then we split the file by the same variable. The split file command temporarily splits the file by the variable specified. All analyses will be grouped by this variable until the split file off command is issued, or until the data are resorted. Note that the
split file command can be used with numeric, short and long string variables. (Many SPSS commands will not work
with long string variables, but split file will.) Next, list the commands for the analyses that you would like. Finally, issue the
split file off command.**

```
sort cases by iv1.
split file by iv1.
correlations var = dv1 with dv2.
```

IV1	DV2		
1	DV1	Pearson Correlation	.993
Sig. (2-tailed)	.073		
N	3		

2	DV1	Pearson Correlation	.780
Sig. (2-tailed)	.220		
N	4		
3	DV1	Pearson Correlation	-.766
Sig. (2-tailed)	.444		
N	3		

split file off.

Note that you can use more than one variable to categorize your
analysis. To do so, list all of the variables by which you want the analysis categorized in the
sort cases command and in the split file command.

sort cases by iv1 iv2.
split file by iv1 iv2.
correlations var = dv1 with dv2.

IV1	IV2	DV2		
1	1.00	DV1	Pearson Correlation	.993
Sig. (2-tailed)	.073			
N	3			
2	1.00	DV1	Pearson Correlation	1.000
Sig. (2-tailed)	.			
N	2			
2.00	DV1	Pearson Correlation	-1.000	
Sig. (2-tailed)	.			

N	2			
3	**2.00**	**DV1**	**Pearson Correlation**	-.766
Sig. (2-tailed)	.444			
N	3			

split file off.