

Σ+ SPSS TUTORIALS

BASICS DATA ANALYSIS T-TEST ANOVA CHI-SQUARE TEST

Missing Values Tutorial

ezoic

You are here: [Home](#) > [SPSS System Missing Values](#)

[Basics](#) → [SPSS - Popular Tutorials](#) → [SPSS User Missing Values](#)

[SPSS Missing Values Tutorial](#) → [Setting User Missing Values](#)

[Inspecting Missing Values per Variable](#)

- [SPSS Data Analysis with Missing Values](#)

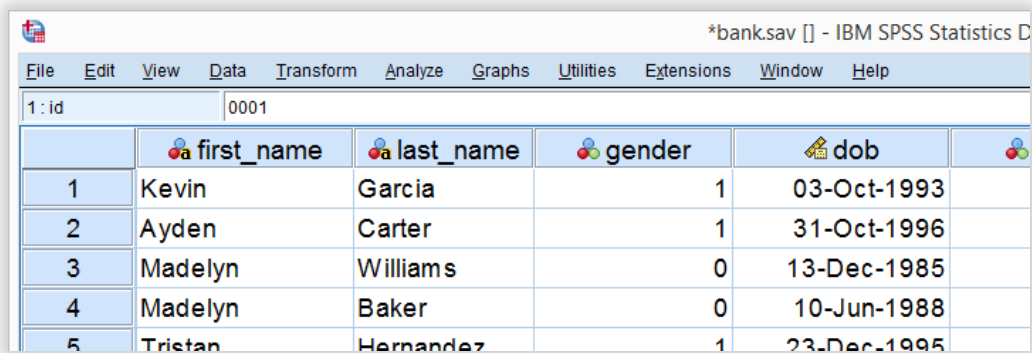
	Name	Label	Values	Missing	Columns
1	id	Questionnaire id...	None	None	10
2	completed	Date and time th...	None	None	22
3	first_name		None	None	11

What are “Missing Values” in SPSS?

In SPSS, “missing values” may refer to 2 things:

- **System missing values** are values that are completely absent from the data. They are shown as periods in **data view**.
- **User missing values** are values that are invisible while analyzing or editing data. The SPSS user specifies which values -if any- must be excluded.

This tutorial walks you through both. We'll use **bank.sav** -partly shown below- throughout. You'll get the most out of this tutorial if you try the examples for yourself after downloading and opening this file.



The screenshot shows the SPSS Data View for a file named *bank.sav. The variable list at the top includes id, first_name, last_name, gender, and dob. The data table below shows five rows of data.

	id	first_name	last_name	gender	dob
1	0001	Kevin	Garcia	1	03-Oct-1993
2		Ayden	Carter	1	31-Oct-1996
3		Madelyn	Williams	0	13-Dec-1985
4		Madelyn	Baker	0	10-Jun-1988
5		Tristan	Hernandez	1	23-Dec-1995

SPSS System Missing Values

System missing values are values that are completely absent from the data.

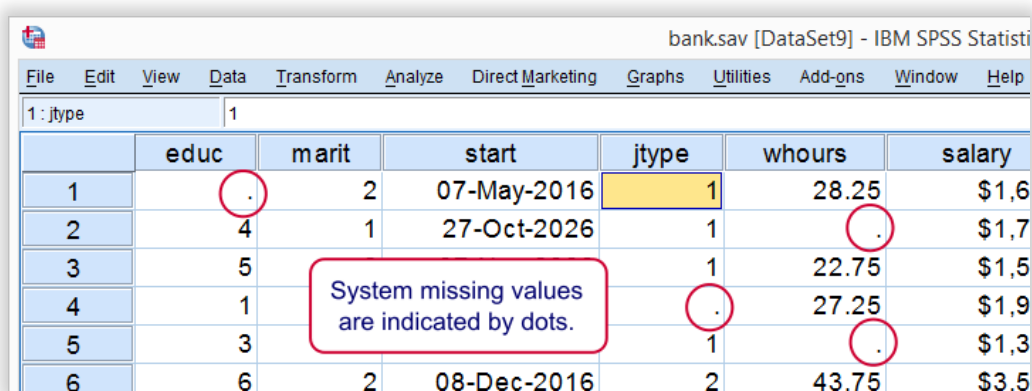
System missing values are shown as dots in data view as shown below.

AdChoices

Analyzing Data

Data Analysis Example

Missing Cases SPSS



The screenshot shows the SPSS Data View for a file named bank.sav [DataSet9]. The variable list at the top includes educ, marit, start, jtype, whours, and salary. The data table below shows six rows of data. System missing values are indicated by dots in the cells for educ, whours, and jtype in rows 1, 2, 4, and 5. A red box highlights the text "System missing values are indicated by dots." with arrows pointing to the dots in the data table.

	educ	marit	start	jtype	whours	salary
1	.	2	07-May-2016	1	28.25	\$1,6
2	4	1	27-Oct-2026	1	.	\$1,7
3	5			1	22.75	\$1,5
4	1			.	27.25	\$1,9
5	3			1	.	\$1,3
6	6	2	08-Dec-2016	2	43.75	\$3.5

System missing values are only found in numeric variables. **String variables** don't have system missing values. Data may contain system missing values for several **reasons**:

- some respondents **weren't asked** some questions due to the questionnaire routing;
- a respondent **skipped** some questions;
- something went wrong while converting or **editing** the data;
- some values weren't recorded due to equipment **failure**.

In some cases system missing values make perfect sense. For example, say I ask

“do you own a car?”

and somebody answers **“no”**. Well, then my survey software should skip the next question:

“what color is your car?”

In the data, we'll probably see system missing values on color for everyone who does *not* own a car. These missing values make perfect sense.

In other cases, however, it may not be clear why there's system missings in your data. Something may or may not have gone wrong. Therefore, you should try to

find out *why* some values are system missing

especially if there's many of them.

So how to detect and handle missing values in your data? We'll get to that after taking a look at the second type of missing values.



SPSS User Missing Values

User missing values are values that are excluded when analyzing or editing data.

“User” in user missing refers to the SPSS user. Hey, that's you! So it's *you* who may need to set some values as user missing. So which -if any- values must be excluded? Briefly,

- for **categorical variables**, answers such as “don't know” or “no answer” are typically excluded from analysis.
- For **metric variables**, unlikely values -a reaction time of 50ms or a monthly salary of € 9,999,999- are usually set as user missing.

For **bank.sav**, no user missing values have been set yet, as can be seen in variable view.

bank.sav [] - IBM SPSS Statistics D

	Name	Label	Values	Missing	Col
1	id	Questionnaire id...	None	None	10
2	completed	Date and time th...	None	None	22
3	first_name		None	None	11
4	last_name		None	None	22

Let's now see if any values should be set as user missing and how to do so.

User Missing Values for Categorical Variables

A quick way for inspecting categorical variables is running **frequency distributions** and corresponding **bar charts**. Make sure the output tables show both values and value labels. The easiest way for doing so is **running the syntax below**.

***Show both values and value labels in succeeding output**

set tnumbers both.

***Basic frequency table for q1.**

frequencies q1 to q9.

Result

q1 This company takes good care of its employees.

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 2	20	4.3	4.4	4.4
3	43	9.3	9.6	14.0
9	35	7.5	7.8	93.6
10 Totally agree	23	5.0	5.1	98.7
11 No answer				100.0
Total	450	97.0	100.0	
Missing System	14	3.0		
Total	464	100.0		

→ 11 MUST BE SET AS USER MISSING VALUE

First note that q1 is an **ordinal variable**: higher values indicate higher levels of agreement. However, this does not go for 11: "No answer" does *not* indicate more agreement than 10 - "Totally agree". Therefore, only values 1 through 10 make up an ordinal variable and 11 should be excluded.

The syntax below shows the right way to do so.

***Set 11 as user missing value for q1.**

missing values q1 to q9 (11).

***Rerun frequencies table.**

frequencies q1 to q9.

Result

q1 This company takes good care of its employees.

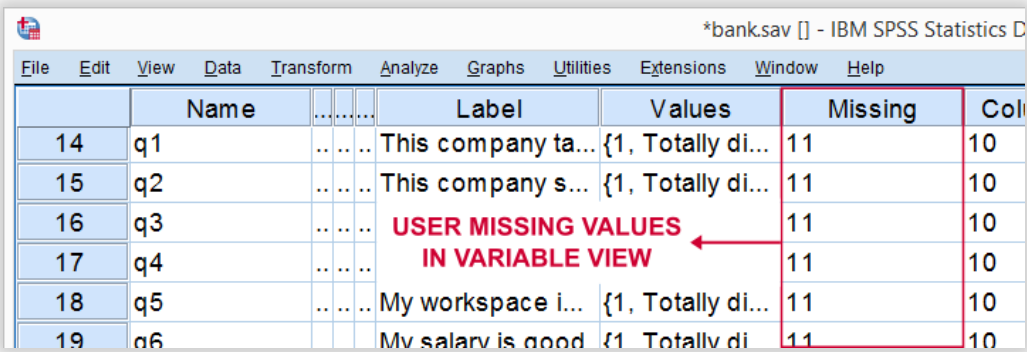
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 2	20	4.3	4.5	4.5
10 Totally agree	23	5.0	5.2	100.0
Total	444	95.7	100.0	
Missing 11 No answer	6			
System	14			
Total	20	4.3		
Total	464	100.0		

→ 11 = USER MISSING VALUE

→ SYSTEM MISSING VALUES

Note that 11 is shown among the missing values now. It occurs 6 times in q1 and there's also 14 system missing values. In variable view, we also

see that 11 is set as a user missing value for q1 through q9.



	Name	Label	Values	Missing	Col
14	q1	This company ta...	{1, Totally di...	11	10
15	q2	This company s...	{1, Totally di...	11	10
16	q3			11	10
17	q4			11	10
18	q5	My workspace i...	{1, Totally di...	11	10
19	q6	My salary is good	{1, Totally di...	11	10

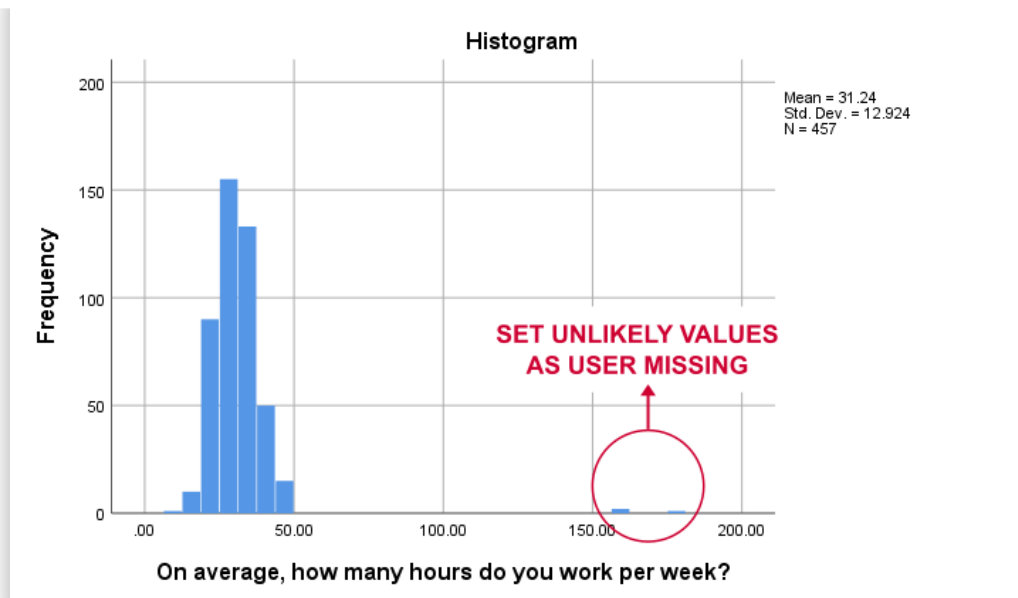
User Missing values for Metric Variables

The right way to inspect **metric variables** is running **histograms** over them.
The syntax below shows the easiest way to do so.

***Run basic histogram over working hours per week.**

```
frequencies whours
/format notable
/histogram.
```

Result



Some respondents report working over 150 hours per week. Perhaps these are their monthly -rather than weekly- hours. In any case, such values are not credible. We'll therefore set all values of 50 hours per week or more as user missing. After doing so, the distribution of the remaining values looks plausible.

***Set 50 hours per week or more as user missing.**

```
missing values whours (50 thru hi).
```

***Rerun histogram.**

```
frequencies whours
/format notable
/histogram.
```

Inspecting Missing Values per Variable

A super fast way to inspect (system and user) missing values per variable is running a basic DESCRIPTIVES table. Before doing so, make sure you don't have any **WEIGHT** or **FILTER** switched on. You can check this by running

```
SHOW WEIGHT FILTER N.
```


Also note that there's 464 cases in these data. So let's now inspect the descriptive statistics.

***Check missing values per variable.**

descriptives q1 to q9.

***Note: $(464 - N) = \text{number of missing values.}$**

Result

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
q1 This company takes good care of its employees.	444	2	10	5.99	2.106
q2 This company supports me in my work.	438				
q9 The cooperation with my colleagues is good.	440	3	10	6.97	1.799
Valid N (listwise)	309				

© 2018 www.spss-tutorials.com

The N column shows the number of non missing values per variable. Since we've 464 cases in total, $(464 - N)$ is the number of missing values per variable. If any variables have high percentages of missingness, you may want to exclude them from -especially- multivariate analyses. Importantly, note that **Valid N (listwise) = 309**. These are the cases without any missing values on all variables in this table. Some procedures will use only those 309 cases -known as **listwise exclusion** of missing values in SPSS.

Conclusion: none of our variables -columns of cells in data view- have huge percentages of missingness. Let's now see if any cases -rows of cells in data view- have many missing values.

Inspecting Missing Values per Case

For inspecting if any cases have many missing values, we'll create a new variable. This variable holds the number of missing values over a set of variables that we'd like to analyze together. In the example below, that'll be q1 to q9.

We'll use a short and simple variable name: `mis_1` is fine. Just make sure you add a description of what's in it -the number of missing...- as a variable label.

***Create new variable holding number of missing values**

```
count mis_1 = q1 to q9 (missing).
```

***Set description of `mis_1` as variable label.**

```
variable labels mis_1 'Missing values over q1 to q9'.
```

***Inspect frequency distribution missing values.**

```
frequencies mis_1.
```

Result

mis_1 Missing values over q1 to q9

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid .00	309	→ 309 CASES COMPLETE ON Q1 TO Q9		
1.00	125	26.9	26.9	93.5
2.00	22	4.7	4.7	98.3
8.00	1	.2	.2	99.6
9.00	2	.4	.4	100.0
Total	464	100.0		

© 2018 www.spss-tutorials.com

In this table, 0 means zero missing values over q1 to q9. This holds for **309** cases. This is the **Valid N (listwise)** we saw in the descriptives table earlier on.

Also note that 1 case has 8 missing values out of 9 variables. We may doubt if this respondent filled out the questionnaire seriously. Perhaps we'd better exclude it from the analyses over q1 to q9. The right way to do so is using a **FILTER**.

SPSS Data Analysis with Missing Values

So how does SPSS analyze data if they contain missing values? Well, in most situations,

SPSS runs each analysis on all cases it can use for it.

Right, now our data contain 464 cases. However, most analyses can't use all 464 because some may drop out due to missing values. Which cases drop out depends on which analysis we run on which variables.

Therefore, an important best practice is to

always inspect how many cases are actually used for each analysis you run.

This is not always what you might expect. Let's first take a look at **pairwise** exclusion of missing values.

Pairwise Exclusion of Missing Values

Let's inspect all (Pearson) correlations among q1 to q9. The simplest way for doing so is just running

```
correlations q1 to q9.
```

If we do so, we get the table shown below.

		q1	q2	q3	q4	q5	q6	q7	q8	q9
q1	Pearson Correlation	1	.566	.129	.168	.086	.110	.117	.073	.153
	Sig. (2-tailed)		.000	.008	.000	.080	.023	.016	.134	.002
	N	444	423	419	433	416	427	425	427	425
q2	Pearson Correlation	.566	1	.227	.138	.085	.076	.128	.069	.180
	Sig. (2-tailed)	.000		.000	.004	.086	.120	.009	.158	.000
	N	423	438	412	425	409	419	420	420	419
q3	Pearson Correlation	.129	.227	1	.243	.115	.140	.147	.284	.144

Note that each correlation is based on a different number of cases. Precisely, each correlation between a pair of variables uses all cases having valid values on these 2 variables. This is known as **pairwise exclusion** of missing values. Note that most correlations are based on some 410 up to 440 cases.

Listwise Exclusion of Missing Values

Let's now rerun the same correlations after adding a line to our minimal syntax:

```
correlations q1 to q9
/missing listwise.
```

After running it, we get a smaller correlation matrix as shown below. It no longer includes the number of cases per correlation.

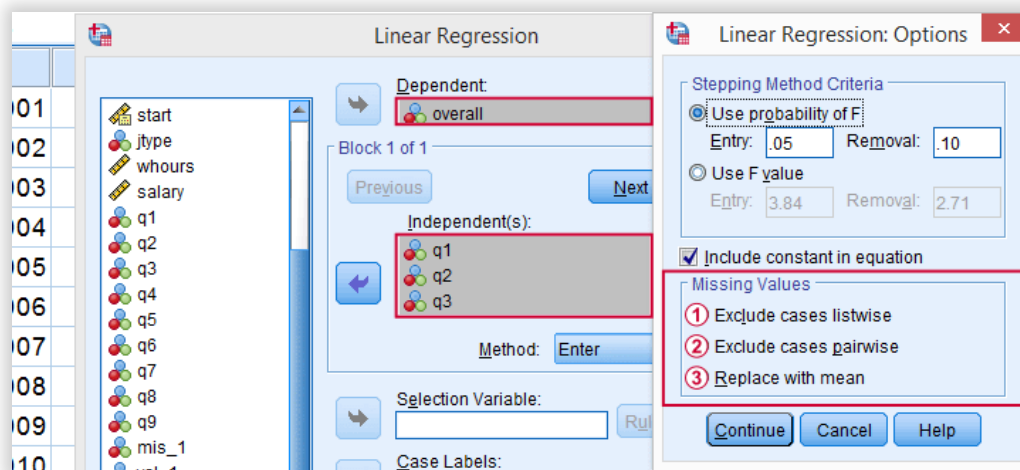
		q1	q2	q3	q4	q5	q6	q7	q8	q9
q1	Pearson Correlation	1	.584	.089	.153	.082	.104	.109	.050	.168
	Sig. (2-tailed)		.000	.116	.007	.149	.067	.055	.378	.003
q2	Pearson Correlation	.584	1	.168	.155	.105	.102	.106	.077	.182
	Sig. (2-tailed)	.003	.001	.002	.000	.006	.004	.024	.045	
a. Listwise N=309										

Each correlation is based on the same 309 cases, the **listwise N**. These are the cases without missing values on *all* variables in the table: q1 to q9. This is known as **listwise exclusion** of missing values.

Obviously, listwise exclusion often uses far fewer cases than pairwise exclusion. This is why we often recommend the latter: we want to use as many cases as possible. However, if many missing values are present, pairwise exclusion may cause computational issues. In any case, make sure you

know if your analysis uses
listwise or pairwise exclusion of missing values.

By default, **regression** and **factor analysis** use listwise exclusion and in most cases, that's *not* what you want.

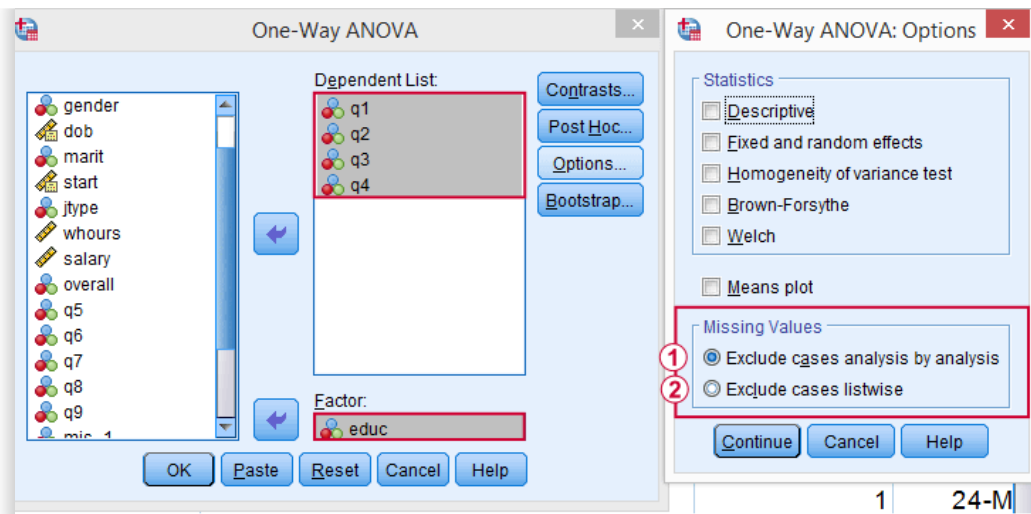


Exclude Missing Values Analysis by Analysis

Analyzing if 2 variables are associated is known as bivariate analysis. When doing so, SPSS can only use cases having valid values on both variables. Makes sense, right?

Now, if you run several bivariate analyses in one go, you can exclude cases **analysis by analysis**: each separate analysis uses all cases it can. Different analyses may use different subsets of cases.

If you don't want that, you can often choose **listwise** exclusion instead: each analysis uses only cases without missing values on *all* variables for *all* analyses. The figure below illustrates this for **ANOVA**.



- ① The test for q1 and educ uses all cases having valid values on q1 and educ, regardless of q2 to q4.
- ② All tests use only cases without missing values on q1 to q4 and educ.

We usually want to use as many cases as possible for each analysis. So we prefer to exclude cases analysis by analysis. But whichever you choose, make sure you

know how many cases are used for each analysis.

So check your **output** carefully. The **Kolmogorov-Smirnov test** is especially tricky in this respect: by default, one option excludes cases analysis by analysis and the other uses listwise exclusion.

Editing Data with Missing Values

Editing data with missing values can be tricky. Different commands and functions act differently in this case. Even something as basic as **computing means in SPSS** can go very wrong if you're unaware of this. The syntax below shows 3 ways we sometimes encounter. With missing values, however, 2 of those yield **incorrect results**.

***Right way to compute mean.**

```
compute mean_a = mean(q1 to q9).
```

***Compute mean - wrong way 1.**

```
compute mean_b = (q1 + q2 + q3 + q4 + q5 + q6 + q7 +
```

***Compute mean - wrong way 2.**

```
compute mean_c = sum(q1 to q9) / 9.
```

***Check results.**

```
descriptives mean_a to mean_c.
```

Result

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
mean_a	462			5.8961	1.02598
mean_b	309			5.8972	1.02899
mean_c	462			5.6190	1.10704
Valid N (listwise)	309				

Final Notes

In real world data, missing values are common. They don't usually cause a lot of trouble when analyzing or editing data but in some cases they do. A little extra care often suffices if missingness is limited. Double check your results and know what you're doing.

Thanks for reading.

Let me know what you think!

Done!

**Required field. Your comment will show up after approval from a moderator.*

This tutorial has 28 comments

By Yang Chia Heng on October 29th, 2019

after so much slides read, i still dont understand how to find the missing data



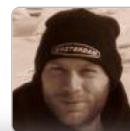
By neema on June 5th, 2019

Good explanation.



By **Ruben Geert van den Berg** on April 10th, 2019

Good question. I tried it (see below) but it crashes.



You can, however, **RECODE** system
huge number like 999999999 and

Expand comment | **all comments**

1 ... 6

Get In Touch!

Ruben Geert van den Berg

 **LinkedIn**

 **Facebook**

SPSS Help (Netherlands)

Sigma Plus Statistiek

www.sigma-plus-statistiek.nl

info@sigma-plus-statistiek.nl

SPSS Help (International)

SPSS tutorials

www.spss-tutorials.com

info@spss-tutorials.com

 **SPSS tutorials**

© Copyright Protected 2019 by **Sigma Plus Statistiek**

Disclaim