

What is the difference between long and wide data?

Authored by
stats writer

July 1, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the difference between long and wide data?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=165264>

Long data refers to a format where each row represents a unique observation and each column represents a specific variable. This format is commonly used for statistical analysis and allows for easy manipulation and analysis of data.

On the other hand, wide data refers to a format where each row represents multiple observations and each column represents a specific variable. This format is often used for data visualization and allows for a more compact representation of data.

In summary, the main difference between long and wide data is the way in which the data is organized and presented. Long data is more suitable for statistical analysis, while wide data is more convenient for data visualization.

Long vs. Wide Data: What's the Difference?

A dataset can be written in two different formats: wide and long.

A wide format contains values that *do not* repeat in the first column.

A long format contains values that *do* repeat in the first column.

For example, consider the following two datasets that contain the exact same data expressed in different formats:

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Notice that in the wide dataset, each value in the first column is unique.

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

Each value is
unique in first
column

By contrast, in the long dataset the values in the first column repeat.

Long Format

	Team	Variable	Value
The values in the first column repeat	A	Points	88
	A	Assists	12
	A	Rebounds	22
	B	Points	91
	B	Assists	17
	B	Rebounds	28
	C	Points	99
	C	Assists	24
	C	Rebounds	30
	D	Points	94
	D	Assists	28
	D	Rebounds	31

Both datasets contain the exact same information about the teams, but they're simply expressed in different formats.

When to Use Wide vs. Long Data

Depending on what you want to do with your data, it may make more sense to have it in a wide or long format.

When to Use Wide Format

As a rule of thumb, if you're analyzing data then you typically will use a wide data format.

For example, if you want to find the average points, assists, and rebounds scored per team then it's often easier to have the data in a wide format:

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31
Average	93	20.25	27.75

Most datasets that you encounter in the real world will also be recorded in a wide format because it's easier for our brains to interpret.

When to Use Long Format

As a rule of thumb, if you're visualizing multiple variables in a plot using statistical software such as you typically must convert your data to a long format in order for the software to create the plot.

For actual examples of this, check out these tutorials in R in which the data must be in a long format to create certain types of plots:

Occasionally you may need to reshape your data into a different format if you're using as well.

The following tutorials explain how to reshape data frames in Python:

Additional Resources

The following tutorials provide information about other commonly used statistical terms:

ARABPSYCHOLOGY.COM