

Data Verification Guidelines

When collecting data from a survey, you want the responses to be accurate. Unfortunately, respondents often give incorrect answers to questions. Some errors are impossible to detect, and those that you can detect may not be possible to correct. However, whenever possible, you should attempt to correct those that you can.

If the mistake is due an incorrect unit, especially when non-traditional units are requested, you might be able to correct it. For example, if the question asks “How much do you weigh (in kilograms)?”, some respondents will not catch the reference to kilograms, or may not know the correct conversion rate. So a response such as “160 pounds” can be corrected by converting it into kilograms.

In other cases, you might suspect that there is a mistake but you can’t be 100% certain. Suppose the respondent answers “250”. Although it is possible that an outlier could weigh 250 kilograms (= 550 pounds) it does seem unlikely. Since changing a response is a delicate matter, you should try to cross-check with other related questions. For example, if this person’s primary vehicle is a Mazda Miata, it is *really* unlikely that they weigh 250 kilograms! This is the hardest type of situation to deal with because you really have to decide whether the answer is a mistake or if it is an outlier. If you decide that it is not an outlier, you need to explain how you dealt with it.

Other answers are clearly wrong, but you might not be able to correct them. Suppose the respondent answers “10”. There is almost no way to determine what the person intended: it could have been 100 or 110 or 120 or almost anything else.

Respondents sometimes give ranges of answers instead of a single numerical value. For example, someone might answer “85-90” for their weight. Although this is probably a more accurate answer since weight varies from day-to-day and week-to-week, but it is impossible to do statistical calculations with ranges of value. Since there is no single “best” way to deal with these types of responses, you need to settle on a method and then apply it reasonably consistently on the data.

Categorical variables given as freeform answers will almost always have errors and inconsistencies. For example birth state, make of first vehicle, and cell phone provider were not selected from a list or drop-down menu, and have differences in capitalization, spelling and abbreviations. Explain how you chose to make the answers accurate and consistent.

You will also need to address data type mismatch. This is particular common when a variable that should be numerical is given as text, for example answering “None” instead of “0”.

Make sure to follow the guiding principle of data validation: The prime directive of data validation is to have clean, consistent and correct data, without sacrificing too much data. Salvage your data wherever possible. Only remove data as an absolute last resort!