

What to look for in Descriptive Statistics

1st Pass

Ensure accurate data that matches your mental image.

2nd Pass

Identify variables of interest to focus on.

Attempt to identify errors in data entry, problems with the codebook, problematic observations, or values representing missing data (e.g., should be set as User Missing or recoded).

Goal

Think ahead to avoid wasting time on variables that are inappropriate for analyses you know. Each analysis has assumptions that must be verified.

for all variables of possible interest

Univariate

for initial set of variables selected

<ul style="list-style-type: none"> <input type="checkbox"/> Consider whether each var should have <i>any</i> missing values <input type="checkbox"/> Consider whether the <i>amount</i> of missing values is plausible 	<p>Missing Values: Frequencies codebook, c</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Look for any patterns of missing values <input type="checkbox"/> Consider whether the missing values are “at random”
<ul style="list-style-type: none"> <input type="checkbox"/> Confirm that each valid value is labeled <input type="checkbox"/> Confirm there is variation (< 90% cases in every category) 	<p>1 Categorical: Frequencies tab1</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Consider grouping when variables have > 5 categories <input type="checkbox"/> Group to avoid categories with < 10% of the obs
<ul style="list-style-type: none"> <input type="checkbox"/> Check for implausible values, considering the population: minimum, maximum, mean, median, and mode 	<p>1 Numeric: Descriptives sum</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Look for a flat or normal distribution / histogram <input type="checkbox"/> Check for extreme values using min/max or z-scores

only for variables of some interest

Bivariate

only for selected variables.

<ul style="list-style-type: none"> <input type="checkbox"/> If a survey had skips, confirm presence of missing values <input type="checkbox"/> Look for at least 5 observations in each cell 	<p>2 Categorical: Crosstabs tab2</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Consider whether % differences are substantial <input type="checkbox"/> Look for similar patterns across levels of the other var
<ul style="list-style-type: none"> <input type="checkbox"/> Check the n’s for disproportionate missing values in a var <input type="checkbox"/> Look for equal variance and std deviation across groups 	<p>1 Each: Compare Means tab, sum</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Do histograms in each category to see distributions <input type="checkbox"/> Consider whether mean differences are substantial
<ul style="list-style-type: none"> <input type="checkbox"/> Check for observations that are isolated from the others <input type="checkbox"/> Check for unexpected strong or perfect relationships 	<p>2 Numeric: Scatterplot scatter</p>	<ul style="list-style-type: none"> <input type="checkbox"/> Check for any patterns or relationships (e.g., linear) <input type="checkbox"/> Look for homoscedasticity