# Checking Dataset Outliers

## Purpose

To describe the process of taking a dataset, separating it into population and sample, and checking for outliers and implausible values in given variables.

## Scope

This SOP is intended for use by any clinical research staff who may be involved in the collection and analysis of any dataset.

## Responsibilities

#### Data Manager

* To provide guidance to Research Assistant(s) working on a dataset
* To adjudicate any potential outliers or implausible data that are found in a dataset

#### Research Assistant

* To identify and verify or note any outliers or implausible data in the dataset

## Procedure

1. Check dataset based on population

* Data Manager or Research Assistant will compare variable values to literature, using programs such as SAS and Spout, in order to identify any extreme values while taking into consideration confounding factors, such as age and gender
  + If extreme values are identified, Data Manager or Research Assistant should manually check individual values to confirm if the value is valid

1. Check dataset based on sample

* Data Manager or Research Assistant will run an univariate distribution in order to determine the mean and standard deviation
* Data Manager or Research Assistant will compare values that are greater than “x” standard deviation away from the mean
  + “x” values are commonly 3 and/or 5
* If extreme outliers are identified, Data Manager or Research Assistant should manually check individual values to confirm if the value is valid

1. Implausible values

* Data Manager or Research Assistant will review variable values for any negative values when not possible

## References

=============== [Back to README](README.md)