|  |  |
| --- | --- |
| USAID Logo | USAID GLOBAL HEALTH SUPPLY CHAIN PROGRAM  Procurement and Supply Management |

|  |
| --- |
| CONSUMPTION ANOMALIES TOOL  User Guide  2022 |

The USAID Global Health Supply Chain Program-Procurement and Supply Management (GHSC-PSM) project is funded under USAID Contract No. AID-OAA-I-15-0004.  GHSC-PSM connects technical solutions and proven commercial processes to promote efficient and cost-effective health supply chains worldwide. Our goal is to ensure uninterrupted supplies of health commodities to save lives and create a healthier future for all. The project purchases and delivers health commodities, offers comprehensive technical assistance to strengthen national supply chain systems, and provides global supply chain leadership.

GHSC-PSM is implemented by Chemonics International, in collaboration with Arbola Inc., Axios International Inc., IDA Foundation, IBM, IntraHealth International, Kuehne + Nagel Inc., McKinsey & Company, Panagora Group, Population Services International, SGS Nederland B.V., and University Research Co., LLC. To learn more, visit [ghsupplychain.org](http://www.ghsupplychain.org/)

DISCLAIMER:  
The views expressed in this publication do not necessarily reflect the views of the U.S. Agency for International Development or the U.S. government.

Contents

[Contents 1](#_Toc119057745)

[Acronyms 2](#_Toc119057746)

[Consumption Anomalies Tool User Guide 3](#_Toc119057747)

[**Background and Purpose** 3](#_Toc119057748)

[Anomaly Detection Method 3](#_Toc119057749)

[**Requirements** 3](#_Toc119057750)

[Data Requirements 3](#_Toc119057751)

[**Setup** 4](#_Toc119057752)

[Running the Tool 5](#_Toc119057753)

[**Excel Dashboard (optional)** 7](#_Toc119057754)

[Annex 1. Anomaly Detection Methodology 10](#_Toc119057755)

[Overview 10](#_Toc119057756)

[Control Charts 10](#_Toc119057757)

[Data Quality 11](#_Toc119057758)

[Test Rules 12](#_Toc119057759)

[Appendix A: Control Chart Constants 13](#_Toc119057760)

Acronyms

|  |  |
| --- | --- |
| eLMIS | Electronic Logistics Management Information System |
| GHSC-PSM | Global Health Supply Chain – Procurement and Supply Management |
|  |  |

# Consumption Anomalies Tool User Guide

## **Background and Purpose**

Background

The Consumption Anomalies tool is designed to detect anomalous activity in consumption quantities of stock data. The tool was originally developed and used in Zambia. The solution was later refactored to be used on generalized stock data in ordered to be used by a broader set of countries. A simple user interface was added so users can run the tool from start to finish without needing to interact directly with the code.

Using historical consumption data, the tool detects anomalies in consumption at the facility and product level and flags these combinations to the user for review.

Anomaly Detection Method

The Consumption Anomaly Detection tool uses a method called Statistical Process Control (SPC) to detect anomalies. SPC is a standard practice in the manufacturing industry. It detects anomalies in data by calculating a moving range, with an upper and lower bound based on points of data over time.

This method is useful because it accounts for some reasonable increases and decreases in values over time, which will not be flagged as anomalies, while being able to detect “unexpected” deviations in the data. For the application of facility consumption data, this means that a facility can may have a gradual decrease or increase in consumption of a product over time without it being flagged as an anomaly.

Find details on the mathematical specifics of the SPC implemented in this tool below in Annex 1.

## **Requirements**

* Python 3

Data Requirements

Data Fields Required

* Facility
* Product
* Consumption
* Date

Historical Data Requirements

The recommended amount of historical data to run anomaly detection is 24 periods (24 months if data is monthly, or 24 quarters if data is quarterly) but the tool can be run with as few as 12 periods, though results will be more accurate with 24 periods or more.

The tool includes some requirements for the completeness of data. By default, 75% of records in the historical data must be present for a given facility and product in order to perform anomaly detection, as well as 4 of the last 6 periods must be present. If this is not the case, the tool will automatically exclude that facility and product combination from the resulting analysis and show only combinations where it had enough data points to perform anomaly detection. While these are the recommended thresholds, they can be adjusted by the user in the tool if appropriate.

Monthly vs. Quarterly Data

The Consumption Anomaly Detection Tool can be run on both data collected on a monthly or quarterly reporting basis. Monthly data refers to one data point on consumption per product and facility per month. Similarly, quarterly data refers to one data point on consumption per product and facility per quarter. There is an option in the interface to select whether the data is monthly or quarterly.

As of this time the tool is not set up to work on transactional data, though such data could simply be aggregated to the monthly level and then run through this tool.

## **Setup**

Setup only needs to be performed once when setting up a computer to run the Consumption Anomalies tool. Once setup is completed successfully once the tool will be able to be run repeatedly.

Install Python 3

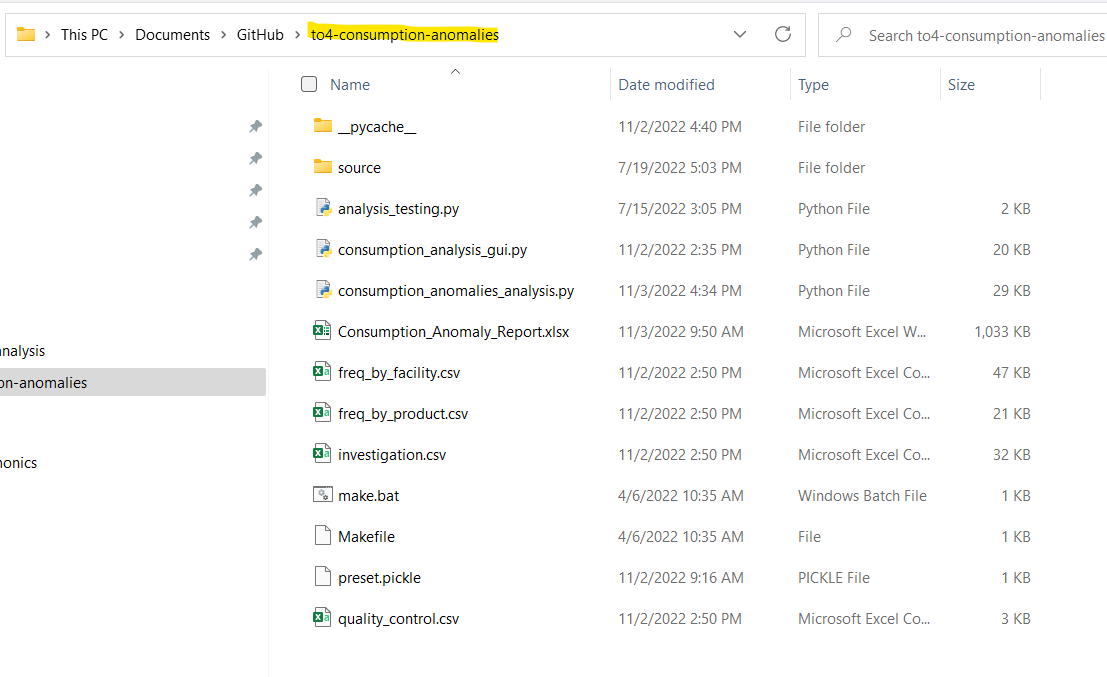
Python can be installed by going to this webpage <https://www.python.org/downloads/> and downloading the latest version of python that’s appropriate for your computer. Go through the installation prompts on your computer as directed.

Install Python Packages

In the folder that contains the extraction tool you will see a document called **requirements.txt**. This document helps to quickly install all the libraries needed to run this program.

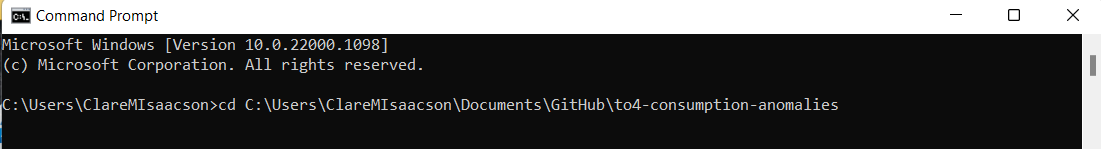
To do this, open the **command prompt** on your PC and navigate to the to4-consumption-anomalies folder

Copy the path to the folder – Right click on the folder name (highlighted below) and select Copy address.

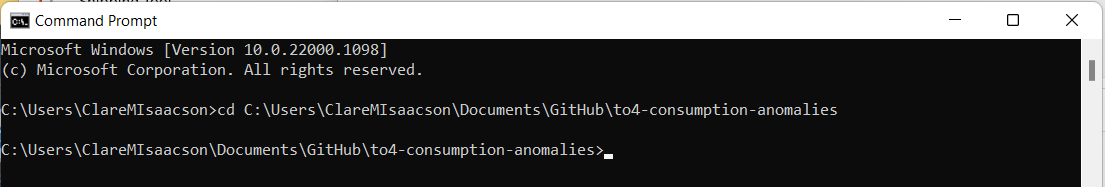


To navigate to the program folder in the command prompt ender the following code after the >

cd *paste*/*path/to/the/extract/folder*

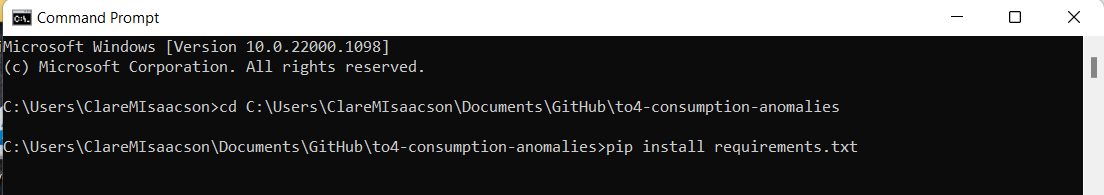


Press Enter. The folder path should now appear before the >



Enter the following text into the command prompt:

pip install -r requirements.txt



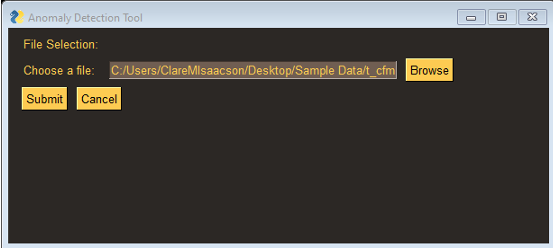
You will see text running in the command prompt, which is installing each package. Once all libraries are installed you should return to showing the line with your current folder followed by >.

Package installation is now complete

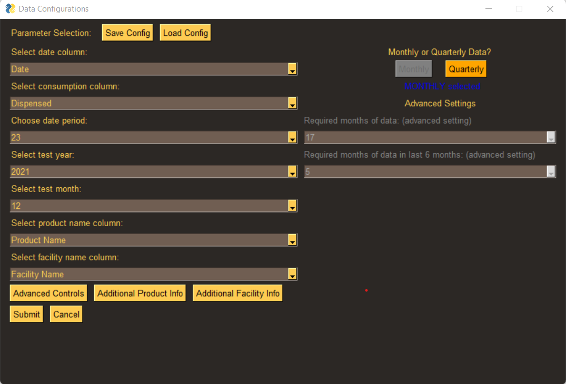
Running the Tool

Startup

The tool will launch the first window to prompt the user to select their data file. Select **Browse** and navigate to the file in the file system. When finished the path to the file will appear in box. Select **Submit**.



The second window will open prompting the user to fill in information about their specific dataset. Fill in each field. Some fields dropdown menus will be pre-populated with the column headers of the selected dataset. See below for specifics on each field.



1. **Date column**: select which column contains the date
2. **Consumption column**: select which column contains consumption
3. **Date period**: How many periods of *historical* data does this data set contain? Exclude the last period (e.g. if there are 24 months of data total, the last month will be checked for anomalies, leaving 23 months of historical data. Enter 23 in the field).
4. **Test year**: Which year does the testing period occur in (this should be the latest period in the dataset).
5. **Test month**: Which month is the testing period (this should be the latest period in the dataset). In combination with the test year, this should indicate the testing period. Note: for quarterly data select the month that corresponds to the quarter in the dataset (e.g. if the date is the first of the quarter 1 2021, select 1 for the month and 2021 for year).
6. **Product name column**: select which column contains the unique product name
7. **Facility name columns**: select which column contains the unique facility name
8. **Monthly or quarterly data**: select whether data is reported monthly or quarterly
9. **Required months of data**: select how many periods of data out of the total date period must be present in order to perform anomaly detection. The tool defaults to 75% (shown automatically, and recommended) this can be changed by clicking the “Advanced Controls” button to unlock the field
10. **Required months of data last 6 months**: select how many periods of data out of the last 6 periods must be present in order to perform anomaly detection. The tool defaults to 5 out of 6 (shown automatically, and recommended) this can be changed by clicking the “Advanced Controls” button to unlock the field.
11. **Additional product info (optional):** select fields related to the product to be included in the final report. Include Product ID, Product Category, or Pack Size. Leave fields blank that don’t have a corresponding column header
12. **Additional facility info (optional):** select fields related to the facility to be included in the final report. Include Facility ID, Region, or District. Leave fields blank that don’t have a corresponding column header

Saving and Loading Configurations

If the user plans to run data with a similar structure time after time (e.g. a data extract which always contains the same column headers) it can be useful to save configurations to cut down on the number of fields that need to be populated.

To save, once all fields are populated, click “Save Config” at the top of the panel.

To load previous configurations, click “Load Config” at the top of the panel. The fields will populate with the saved configurations then fields that need to be changed (such as testing month and year) can be changed as needed.

Running

Once finished populating fields, click “Submit”.

The analysis will take the given inputs and begin running. Information on the analytics process will be printed to the command prompt.

The analysis may take some time to run, depending on the amount of data provided to the tool. The output of the python scrips is a series of csv tables which will be updated in the file system.

1. **investigation.csv**: contains the list of anomalies detected and information on the anomaly
2. **freq\_by\_product.csv**: contains high level details on the anomalies occurring per product
3. **freq\_by\_facility.csv**: contains high level details on the anomalies occurring per facility

These csvs can be used on their own, read into an existing system (such as an eLMIS), used to create a dashboard to visualize the anomalies, or used with the included excel dashboard.

Once finished, the tool will open a window prompting the user to open an excel report. Click “Open Excel Report”.

## **Excel Dashboard (optional)**

This section provides an overview of what information is shown in the provided excel report that displays the information in the output csv files from the analysis.

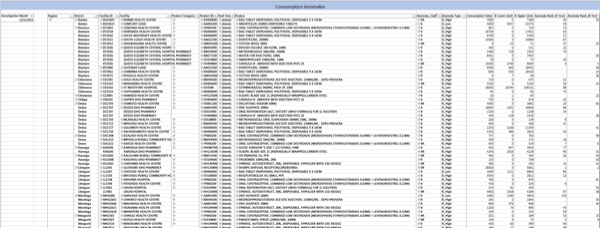
Startup

Upon opening the report go to the **Data** tab on the top ribbon and select the “Refresh All” button to load in the data from the csv files.

Anomalies

The Anomalies tab lists out each anomaly detected for the testing period. The testing period is displayed in the upper left corner.

The **Anomaly Code** shows the type of anomaly that was detected (X, R or both, see appendix for additional information). The **Anomaly Type** shows whether the consumption anomaly is higher or lower than the calculated bounds. The **Consumption Value** shows the value that was determined to be an anomaly, and the **X Lower Limit** and **X Upper Limit** show the control limits calculated for the X test. Furthest to the right shows the ranking of each anomaly by X and R types. The lower the ranking, the greater the anomaly.



Anomalies by Product

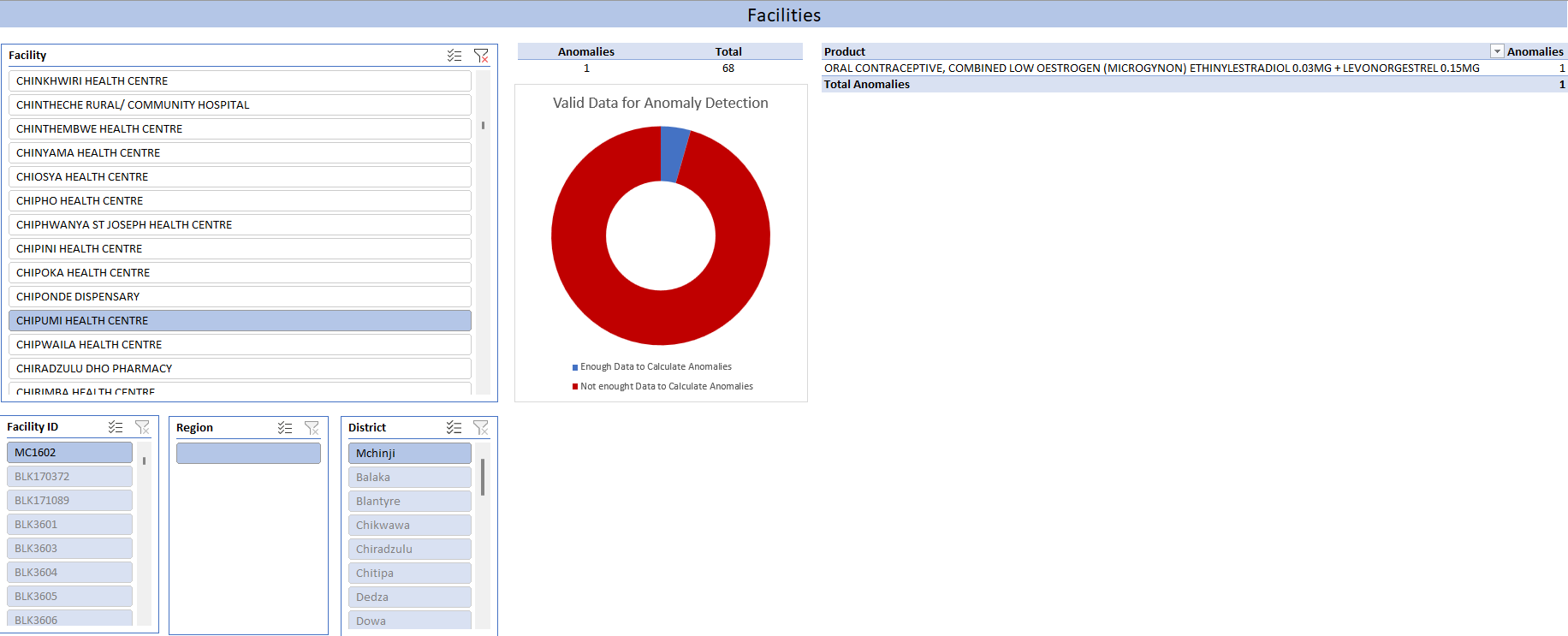
The Product tab visualizes anomalies for each product. Selecting a product from the **Product** slicer displays all the facilities that had an anomaly for that product. The center of the page displays the number of anomalies detected for that product out of the total facilities with data on that product. The donut chart displays how many of those facilities had enough data to perform anomaly detection per the thresholds established on the user interface input.

Graphical user interface, application

Description automatically generated

Anomalies by Facility

The Facility tab visualizes anomalies for each product. Selecting a product from the **Facility** slicer displays all the products that had anomalies at that facility. The center of the page displays the number of anomalies detected at that facility out of all products for which that facility had data. The donut chart displays how many of those products had enough data to perform anomaly detection per the thresholds established on the user interface input.



Annex 1. Anomaly Detection Methodology

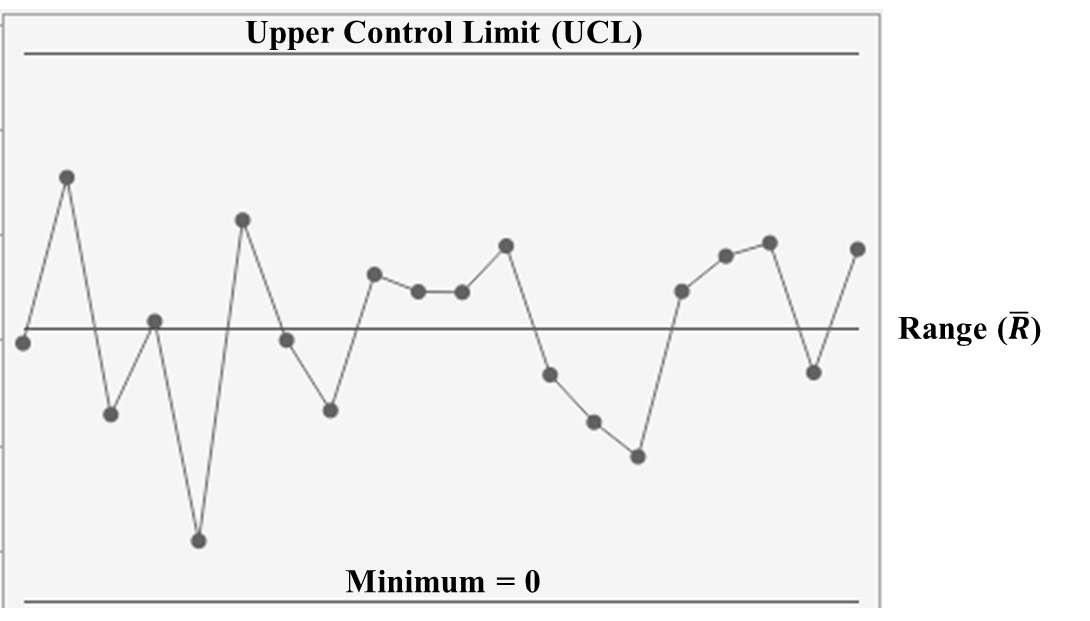
Overview

Detection of consumption outliers requires a statistical procedure known as Statistical Process Control (SPC). SPC is an industry-standard methodology in supply chain management using control charts to see if any part of the process is not under control. It is achieved by taking new data samples from the process and plotting these sample points on a chart, to see if the process is within statistical control limits. The control limits are calculated using historical data over time once the process is consistent and predictable and requires a minimum sample size of 20 for each consumed product.

Control Charts

There are two types of control charts used in SPC. These are the “Moving Range” chart (FIGURE 1), which is used to analyze the spread and variability of the sample’s data points, and the “Individuals” chart (FIGURE 2), which is used to analyze the center of the sample’s data points to detect shifts and trends. These two charts are used in tandem and are known as the I-MR chart.

FIGURE : Moving Range Chart

****

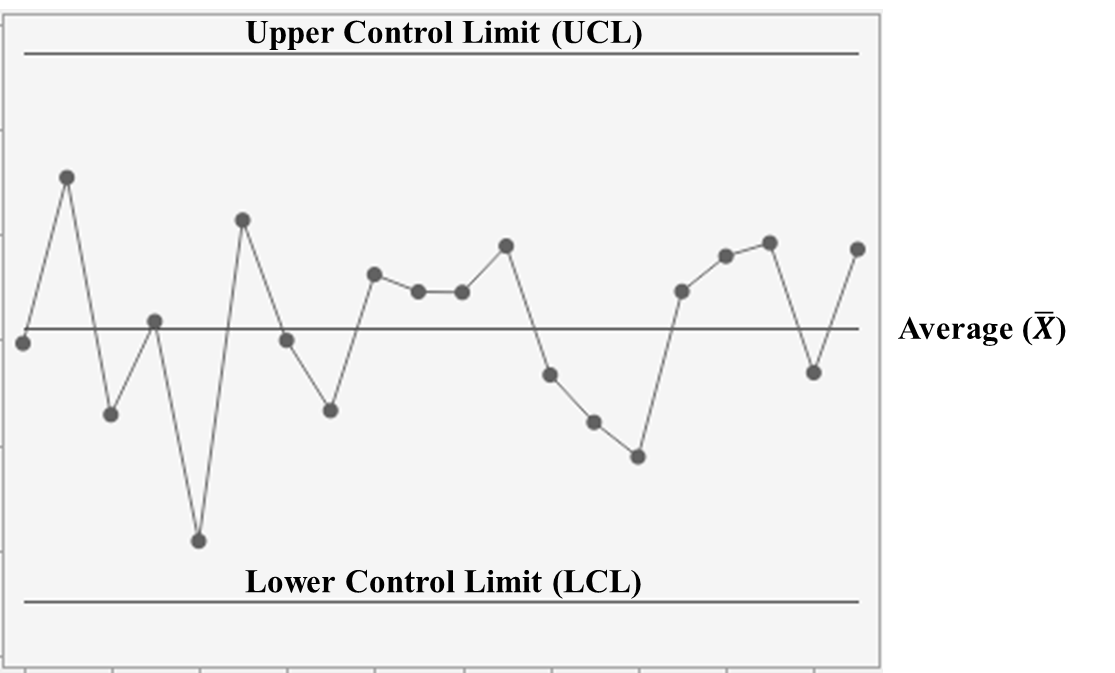
Because the moving range is calculated only with positive values (no negative consumption), the lower limit is always zero. For the Upper Control Limit (UCL) and the Range () of the Moving Range Chart, the following equations are used:

Where:

D4 is pulled from the Control Chart Constant table depending on the sample size of the moving range subgroup (See Appendix A)

Xmax and Xmin are the maximum and minimum data points, respectively, of subgroup k.

FIGURE : Individuals Chart



The following equations are used to calculate the Upper Control Limit (UCL), the Lower Control Limit (LCL), and the Average () of the Individuals Chart:

Where:

d2 is pulled from the Control Chart Constant table depending on the sample size of the moving range subgroup (See Appendix A).

Data Quality

Seasonality Adjustment

The historical data of seasonal products will need to be adjusted for seasonality to eliminate peaks and valleys that could be misidentified as out-of-control conditions. If historical data is not adjusted for seasonality, fewer irregularities will be identified, and these irregularities will be of the extreme variant.

Outliers

Outliers in each product’s historical data will need to be removed. If outliers are not removed, the calculations for the statistical process control limits may introduce bias and be inaccurate when identifying irregularities with new data samples.

To identify outliers, the interquartile range rule will be used. This rule states that a data point is considered an outlier if the following conditions are met:

or

Where:

Q1 is the first quartile

Q3 is the third quartile

Stock on Hand Considerations

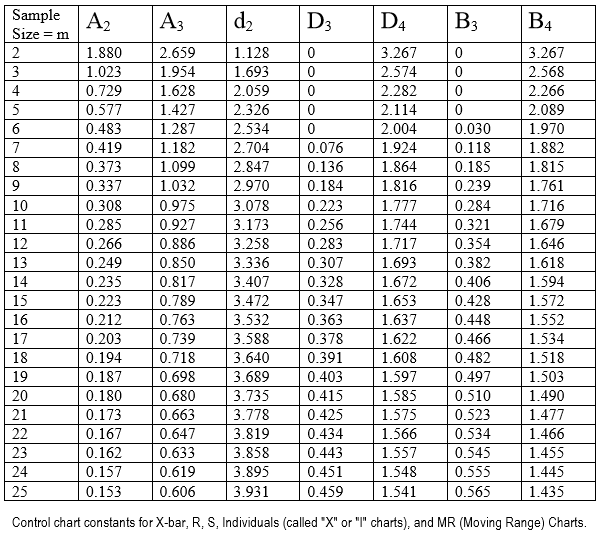
If stock on hand for a product at the beginning of the current month or at the end of the previous month is equal to zero, consumption data will be ignored as there is no more product to be consumed.

Test Rules

Once the statistical control limits have been established, rules can be used to identify irregularities or out-of-control conditions in new data samples from the process, such as:

* If one or more points falls outside of the UCL or LCL.
* 7 or more consecutive points on one side of the Average
* 7 consecutive points trending up or trending down
* 14 consecutive points alternating up and down

Appendix A: Control Chart Constants



*Source: http://web.mit.edu/2.810/www/files/readings/ControlChartConstantsAndFormulae.pdf*