KitCheck\_Outliers\_Analysis

Matthew Yeseta (Master of Science in Data Science, Indiana University)

11/23/2020

### Kit Check Outliers Analysis

This implements the Kit Check Outlier exploratory data discovery analysis for detecting when feature Amount (Medication) Dispensed has any outlier risk for potential discrepancy for balance computation. This study can serve as the base foundation framework which can be offered as a starting point for building a data quality feedback loop to integrate together into doctor’s data practices. For such a quality feedback loop to work effectly, there must be a full complete analysis which can accurately identify outliers that may skew critical computations.

Following is the data specification as put forth in the algorithm for Kit Check’s computed sum of Amount Administered plus Amount Wasted plus Amount Returned in order to audit for the computed balance to match the Amount Dispensed.

Instilling this data outlier risk analysis permits the identification of outlier data that may be shared with the doctor practice which then can further be investigated by the doctor’s practice in order to manage data audits and/or data accurcy corrections to deliver improved accuracy/quality on all data calcuations and audits. Hence the need for to strive for both an accurate outlier detection and a credible feedback loop for sustainable doctor data quality practices.

### Data Acquisition

bcs\_data <- suppressWarnings(suppressMessages(read\_csv("bcs\_data\_homework.csv")))

### Internal Outlier boxplot funcations

discrepency\_func <- function(df, c, g){  
 boxplot(as.formula(paste0(c, ' ~ ', g)), data = df, plot = FALSE)  
}  
  
plot\_discrepency\_func <- function(df, c, g, x, y, t){  
 boxplot(as.formula(paste0(c, ' ~ ', g)), data = df, plot = TRUE, col = "lightblue", xlab=x, ylab=y, main=t)  
}

### Computation Filter

The computation filter is designed to select only data records which do not balance. A balance discrepency occurs when the feature Amount Dispensed does not balance with the computed sum of Amount Administered + Amount Wasted + Amount Returned.

bcs\_data.discrepency <- bcs\_data %>%   
 filter(amount\_dispensed != amount\_administered + amount\_wasted + amount\_returned)

### Outlier Boxplox

This analysis begins outlier analysis using BoxPlots. This forms the exploratory analysis section to visually determine whether outlier anomalies exist in the filtered records for which they do not balance.

Each doctor is grouped into a single boxplot, positioned to display outliers that may be the cause of balance discrepency. These outliers may be the root cause for the sum computation not matching correctly with the algorithm sum of Amount Administered plus Amount Wasted plus Amount Returned.

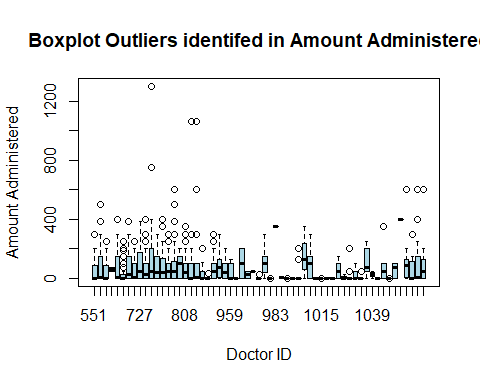
#### Boxplot for Feature Amount Administered

The boxlot of outliers is identifed visually thourgh the boxplot for the filtered data which do not balance. The feature under investigation is the Amount Administered.

Visible on the horizontal x-axis is the doctor ID. The grouping for this display is by Doctor ID. Starting at x-axis zero, in the boxplots, a collection of small boxplots provides a summary view of each detected high value Amount Administered outliers within the respective doctor practice.

In this boxplot analysis, clear evidence is found for outliers within the feature Amount Administered in several of the doctor’s practices. The nature of these outliers is represented by high values charted in the positive range above zero. A significantly large number of outliers exist in this analysis of feature Amount Administered. Chart analysis shows evidence of serveral Amount Administered high value outliers, which skew the normal distribution shape within many of the doctor groups data when charted on a normal distribution bell curve. These outliers give clear evidence in data skewness and will negatively impact computed balance computation for matching the feature data Amount Medication Dispensed.

plot\_discrepency\_func(bcs\_data.discrepency, 'amount\_administered', 'doctor\_id', "Doctor ID", "Amount Administered", "Boxplot Outliers identifed in Amount Administered")



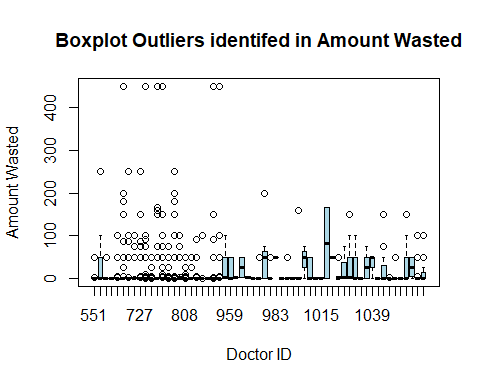
#### Boxplot for Feature Amount Wasted

The boxlot of outliers identifed visually thourgh the boxplot for the filtered data which do not balance. The feature under investigation is the Amount Wasted.

Visible on the horizontal x-axis is the doctor ID. The grouping for this display is by Doctor ID. Starting at x-axis zero in the boxplots, a collection of small boxplots provides a summary view of each detected high value Amount Wasted outliers within the respective doctor practice.

In this boxplot analysis, clear evidence is found for outliers within the feature Amount Wasted in several of the doctor’s practices. The nature of these outliers is represented by high values charted in the positive range above zero. A significantly large number of outliers exist in the analysis of feature Amount Wasted. Chart analysis shows evidence of serveral Amount Wasted high value outliers, which skew the normal distribution shape within many of the doctor groups data when charted on a normal distribution bell curve. These outliers give clear evidence in data skewness and will negatively impact computed balance computation for matching the feature data Amount Medication Dispensed.

plot\_discrepency\_func(bcs\_data.discrepency, 'amount\_wasted', 'doctor\_id', "Doctor ID", "Amount Wasted", "Boxplot Outliers identifed in Amount Wasted")



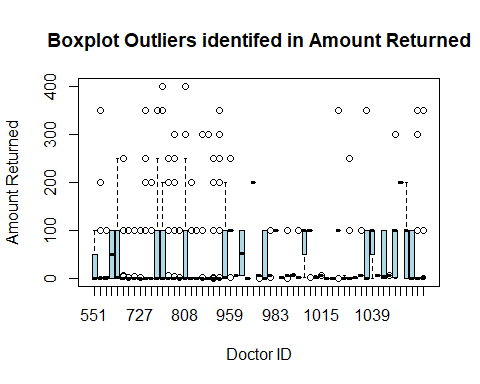
#### Boxplot for Feature Amount Returned

The boxlot of outliers identifed visually thourgh the boxplot for the filtered data which do not balance. The feature under investigation is the Amount Returned is displayed.

Visible on the horizontal x-axis is the doctor ID. The grouping for this display is by Doctor ID. Starting at x-axis zero in the boxplots, a collection of small boxplots provides a summary view of each detected high value Amount Returned outliers within the respective doctor practice.

In this boxplot analysis, clear evidence is found for outliers within the feature Returned Wasted in several of the doctor’s practices. The nature of these outliers is represented by high values charted in the positive range above zero. A significantly large number of outliers exist in the analysis of feature Amount Returned. Chart analysis shows evidence of serveral Amount Returned high value outliers, which skew the normal distribution shape within many of the doctor groups data when charted on a normal distribution bell curve. These outliers give clear evidence in data skewness and will negatively impact computed balance computation for matching the feature data Amount Medication Dispensed.

plot\_discrepency\_func(bcs\_data.discrepency, 'amount\_returned', 'doctor\_id', "Doctor ID", "Amount Returned", "Boxplot Outliers identifed in Amount Returned")



### Statistical Tests.

The statistical tests and techniques for outlier detection utilize sample size and confidence levels, and to enlist formal statistical tests with the Grubbs one-sided and two-sided tests and the Rosner multiple outlier detection tests.

The assumption for normal distribution is necessary to establsih. Therefore, in this study, evidence for statistical normal distribution is verified using the Shapiro-Wilk normality test. For each feature data, namely, Amount Administered, Amount Wasted, Amount Returned, and Amount Dispensed, the Shapiro-Wilk normality test will be run for a verification for normal distribution assessment.

After all Shapiro-Wilk normality tests are run, it can be observed (below) that each test did in fact verify the normal distribution assumption with evidence from the threshold p-value measure, e.g., 2.2e-16, as it was less then the alpha signficanct confidence level at 0.05. This establishes the assumption for normal distribution, and now permits the next step to continue where the analysis can henceforth move forward into formal outlier statistical detection tests (Grubbs, Rosner).

bcs\_data.sample <- sample\_n(bcs\_data, 2500)   
shapiro.test(bcs\_data.sample$amount\_administered)

##   
## Shapiro-Wilk normality test  
##   
## data: bcs\_data.sample$amount\_administered  
## W = 0.59611, p-value < 2.2e-16

shapiro.test(bcs\_data.sample$amount\_wasted)

##   
## Shapiro-Wilk normality test  
##   
## data: bcs\_data.sample$amount\_wasted  
## W = 0.22533, p-value < 2.2e-16

shapiro.test(bcs\_data.sample$amount\_returned)

##   
## Shapiro-Wilk normality test  
##   
## data: bcs\_data.sample$amount\_returned  
## W = 0.47548, p-value < 2.2e-16

shapiro.test(bcs\_data.sample$amount\_dispensed)

##   
## Shapiro-Wilk normality test  
##   
## data: bcs\_data.sample$amount\_dispensed  
## W = 0.67401, p-value < 2.2e-16

### Grubbs Test for Feature Amount Administered Outliers

Grubbs tests performed in this section of the study are for detecting existance of outliers on one tail of the distribution for feature Amount Administered.

The Grubbs statistical test outcome detects a high value at 1300 for feature Amount Administered. This highest value is an outlier which is supported by the evidence of the threshold p-value measure, e.g. 2.2e-16, as it is less than alpha signficanct confidence level at 0.05. The Null hypothesis is rejected and the statistical test concludes that the hightest value at 1300 is an outlier, therefore, accepting the alternative hypothesis.

With the outcome of this existing high value outlier, discovering at least one, as detected by the Grubbs statistical test, it is now evident that the feature Amount Administered data, with high confidence, would certainly cause computation discrepency when computing the balance with Amount Medication Dispensed. This high value outlier will also cause shewness in the normal distribution.

grubbs.test(bcs\_data.discrepency$amount\_administered, type=10)

##   
## Grubbs test for one outlier  
##   
## data: bcs\_data.discrepency$amount\_administered  
## G = 11.46529, U = 0.90327, p-value < 2.2e-16  
## alternative hypothesis: highest value 1300 is an outlier

### Grubbs Test for Feature Amount Wasted Outliers

Grubbs tests performed in this section of the study are for detecting existance of outliers on one tail of the distribution for feature Amount Wasted.

The Grubbs statistical test detects a high value at 450 for feature Amount Wasted. This highest value is an outlier which is supported by the evidence of the threshold p-value measure, e.g., 2.2e-16, as it is less than alpha signficanct confidence level at 0.05. The Null hypothesis is rejected and the statistical test concludes that the hightest value at 450 is an outlier, therefore, accepting the alternative hypothesis.

With the outcome of this existing high value outlier, discovering at least one, as detected by the Grubbs statistical test, it is now evident that the feature Amount Wasted data, with high confidence, would certainly cause computation discrepency when computing the balance with Amount Medication Dispensed. This high value outlier will also cause shewness in the normal distribution.

grubbs.test(bcs\_data.discrepency$amount\_wasted, type=10)

##   
## Grubbs test for one outlier  
##   
## data: bcs\_data.discrepency$amount\_wasted  
## G = 9.97737, U = 0.92675, p-value < 2.2e-16  
## alternative hypothesis: highest value 450 is an outlier

### Grubbs Test for Feature Amount Returned Outliers

Grubbs tests performed in this section of the study are for detecting outliers on one tail of the distribution for feature Amount Returned.

The Grubbs statistical test detects a high value at 400 for feature Amount Returned, and that this highest value is an outlier which is supported by the evidence of the threshold p-value measure, e.g., 1.017e-05, as it is less than alpha signficanct confidence level at 0.05. The Null hypothesis is rejected and the statistical test concludes that the highest value at 400 is an outlier, therefore, accepting the alternative hypothesis.

With the outcome of this existing high value outlier, discovering at least one, as detected by the Grubbs statistical test, it is now evident that the feature Amount Returned data, with high confidence, would certainly cause computation discrepency when computing the balance with Amount Medication Dispensed. This high value outlier will also cause shewness in the normal distribution.

grubbs.test(bcs\_data.discrepency$amount\_returned, type=10)

##   
## Grubbs test for one outlier  
##   
## data: bcs\_data.discrepency$amount\_returned  
## G = 5.62996, U = 0.97668, p-value = 1.017e-05  
## alternative hypothesis: highest value 400 is an outlier

### Rosner’ tests for outliers

The Rosner’s test is called a test for “generalized extreme on Studentized deviate tests for up to k potential outliers”. [Steven P. Millard ([EnvStats@ProbStatInfo.com](mailto:EnvStats@ProbStatInfo.com))]

The Rosner tests in this study is a battery of tests which is a critical link section of this investigation for its purpose to accuratly identify outliers. The outcome results, e.g., rosnerTest all.stats results, can be utilized as the base data in order to build a framework for implementing a data quality feedback loop for the purpose to integrate with doctor’s data practices.

#### Rosner’s test potential outliers in Amount Administered

With the former Grubbs statistical tests, it is now verified that at least one outlier exists with each feature data (Amount Administered, Amount Wasted, Amount Returned). Having established with this evidence, this study next needs to verify using the Rosner tests on whether we can detect multiple specific outliers within each data features. Secondarily, this Rosner test may also find outliers that exist undetected with masked values of similar magnitude to other outliers.

In each Rosner tests, the study pre-sets the estimated number of suspected outliers to detect, as per specification to set the argument (k) to be k=3, for 3 maximum outliers. For the benefit of the full analysis outcome, it is observed that boxplot.stats() detected for feature data Amount Administered, that the number of suspected outliers in truth was found to be 92. Boxplot.stats() further detected for feature data Amount Wasted that the number of suspected outliers was found to be 257. Furthermore, boxplot.stats() detected for feature data Amount Returned which indicated that the number of suspected outliers was found to be 281.

For the outcome observed Rosner test for data feature Amount Administered, there were confirmed 3 suspected outliers were detected within the distribution (limited by the argument k=3). These were identifed at values 1300 & (2) 1059.99; and are located in observation numbers 587, 886, and 886, repectedly in the distributution dataset [bcs\_data.discrepency]. Note that the outlier (2) 1059.99 are two masked values, easily detected by the Rosner test. The outcome of $all.status Outlier table columns specifically confirms outliers by marking each TRUE.

The Rosener sample size was 1,361; the statistical alpha is 0.05 (the default in the Rosner test)

rosner.test <- rosnerTest(bcs\_data.discrepency$amount\_administered, k=3)  
rosner.test$sample.size

## [1] 1361

rosner.test$alpha

## [1] 0.05

rosner.test$all.stats

## i Mean.i SD.i Value Obs.Num R.i+1 lambda.i+1 Outlier  
## 1 0 69.19367 107.35061 1300.00 587 11.465295 4.114948 TRUE  
## 2 1 68.28867 102.06423 1059.99 886 9.716443 4.114772 TRUE  
## 3 2 67.55894 98.48878 1059.99 887 10.076590 4.114595 TRUE

#### Rosner’s test potential outliers in Amount Wasted

For the outcome observed Rosner test for feature Amount Wasted, there were confirmed 3 suspected outliers were detected within the distribution (limited by the argument k=3). These were identifed at value (3) 450; and are located in observation number 1055, 1056, and 1059, repectedly in the distributution dataset [bcs\_data.discrepency]. Note that the outlier (3) 450 are three masked values, easily detected by the Rosner test. The outcome of $all.status Outlier table columns specifically confirms outliers by marking each TRUE.

The Rosener sample size was 1,361; the statistical alpha is 0.05 (the default in the Rosner test)

rosner.test <- rosnerTest(bcs\_data.discrepency$amount\_wasted, k=3)  
rosner.test$sample.size

## [1] 1361

rosner.test$alpha

## [1] 0.05

rosner.test$all.stats

## i Mean.i SD.i Value Obs.Num R.i+1 lambda.i+1 Outlier  
## 1 0 12.84928 43.81423 450 1055 9.977369 4.114948 TRUE  
## 2 1 12.52785 42.19451 450 1056 10.367987 4.114772 TRUE  
## 3 2 12.20594 40.50500 450 1059 10.808396 4.114595 TRUE

### Rosner’s test potential outliers in Amount Returned

For the outcome observed Rosner test for feature Amount Returned, there were confirmed 3 suspected outliers were detected within the distribution (limited by the argument k=3). These were identifed at value (3) 400; and are located in observation number 371, 662, and 663, repectedly in the distributution dataset [bcs\_data.discrepency]. Note that the outlier (3) 400 are three masked values, easily detected by the Rosner test. The outcome of $all.status Outlier table columns specifically confirms outliers by marking each TRUE.

The Rosener sample size was 1,361; the statistical alpha is 0.05 (the default in the Rosner test)

rosner.test <- rosnerTest(bcs\_data.discrepency$amount\_returned, k=3)  
rosner.test$sample.size

## [1] 1361

rosner.test$alpha

## [1] 0.05

rosner.test$all.stats

## i Mean.i SD.i Value Obs.Num R.i+1 lambda.i+1 Outlier  
## 1 0 30.06466 65.70835 400 371 5.629960 4.114948 TRUE  
## 2 1 29.79265 64.96145 400 662 5.698878 4.114772 TRUE  
## 3 2 29.52024 64.20358 400 663 5.770391 4.114595 TRUE