**Surveillance Research Program, NCI, Technical Report #2018-01**

# Detecting outliers with sample size adjustment in the SEER Box Plot Outlier Tool

Huann-Sheng Chen[1], Benmei Liu[1] and James Cucinelli[2]

1. Statistical Methodology and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute
2. Information Management Services Inc.

## Introduction

The SEER Box Plot Outlier Tool (BOT) is developed to detect outliers in the Surveillance, Epidemiology, and End Results (SEER) database for 118 variables (either at person level or tumor case level) across all of the SEER cancer registries and multiple diagnosis years from 1988 to the most recent year (2015). Table 1 provides the full list of variables that are currently implemented in the SEER BOT.

For each variable included in the SEER BOT, for each specific value or group of values of interest, three analyses are performed: 1) Compute percentage of a specific value (or group of values) out of all records in each registry with all years combined, and then compare the percentages across the SEER registries and detect outlier registries if any; 2) Compute percentage of a specific value (or group of values) out of all records in each registry using the most recent (2015) data only, and then compare the percentages across the SEER registries and detect outlier registries if any; 3) Compute percentage of a specific value (or group of values) out of all records in each diagnosis year within an individual cancer registry, and then compare the percentages across the diagnosis years and detect outlier years if any.

The SEER Box Plot Outlier Tool (BOT) is integrated in the SEER Quality Assurance Program. Each year, new data submissions are examined using the BOT and any outliers are investigated.

This brief technical report describes the statistical procedure used in SEER BOT to detect outlier registries or diagnosis years in the SEER database.

## Statistical Methods

For analysis 1, assume y is a variable of interest (e.g., Race 1), $y_{ikt}$ denotes the individual value of y for person or case $i$ $(i = 1, \dots, n_{kt})$ in cancer registry $k$ $(k = 1, \dots, K)$ at diagnosis year $t$ $(t = 1988, \dots, 2015)$, where $K$is the total number of cancer registries included in the SEER database. Let $\tilde{y}_{ikt}$ denote the binary recode for a specific value (or groups of values) of interest. Without loss of generosity, we use the specific value of interest as those “unknowns” for illustration below. That is, $\tilde{y}_{ikt} = \begin{cases} 1, & if\ y_{ijt} = \text{"}Unknown\text{"} \\ 0, & otherwise \end{cases}$.

Assume each $\tilde{y}_{ikt}$ follows a Bernoulli distribution with probability of $p$ to get value of 1 and $1-p$ to get value of 0. Let $\tilde{y}_k = \sum_{t=1988}^{2015} \sum_{i=1}^{n_{kt}} \tilde{y}_{ikt},\ n_k = \sum_{t=1988}^{2015} n_{kt}$.

Then, for each $k$,

$$\tilde{y}_k \sim Binomial(n_k, p).$$

Let $\hat{p}_k = \frac{\tilde{y}_k}{n_k}$, then $\hat{p}_k$'s are estimates of $p$, and

$$E(\hat{p}_k) = p,\ \ Var(\hat{p}_k) = \frac{p(1-p)}{n_k}.$$

The confidence interval for $\hat{p}_k$ is:

$$\hat{p}_k \pm 1.96\sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n_k}}.$$

It is noted that with the same $\hat{p}$, if the sample size $n$ is smaller, the standard deviation is larger. Therefore, there is more variation for a smaller registry, and it is more likely become outliers just by chance. To take this factor into account, we use box-plot of the point estimates and the confidence intervals associated with each point estimates together to identify real outliers.

We use the estimate $\hat{p}_k, k = 1, \dots, K$ , i.e., the registry level percentage of a specific value (or groups of values) for a specific variable of interest, to draw box-plot (Figure 1).

To determine whether the outside values are really outliers, the confidence interval of each $\hat{p}_k$ is compared to the box boundaries calculated from the inter-quartile range (IQRs), the width of the box and whisker plot. $IQR = Q_3 - Q_1$.

The lower boundary is defined as Lower=max(0, $Q_1 - 1.5 * IQR$), and the upper boundary is defined as $Q_3 + 1.5 * IQR$. So the box boundary range is: (max(0, $Q_1 - 1.5 * IQR$), $Q_3 + 1.5 * IQR$).

For any $\hat{p}_k$, if its confidence interval falls completely outside of the boundary range, then it is considered as an outlier. Values with confidence intervals overlap or partially overlap with the box boundary range cannot be considered as outliers. If the confidence interval falls completely outside of the boundary range formed using $(\max(0, Q_1$ −3*IQR), $Q_3$ +3*IQR), then the $\hat{p}_k$ is considered as an extreme outlier.

Same outlier detection procedure is used for all the three analyses, all the variables and all the specific values (or group of values) of interest built in the SEER BOT. The most recent years of data will be updated when new years of SEER data is available.

Table 1: SEER Variables Currently included in the SEER BOT

Age at Dx
Age Recode with < 1 Year Olds
Ambiguous Terminology DX
AYA Site Recode / WHO 2008
Behavior ICD-O-3
Behavior Recode for Analysis
Cause of Death
Census Tract 2000
Census Tract Certainty 2000
Computed Ethnicity
County at DX
CS Extension
CS Lymph Nodes
CS Lymph Nodes Eval
CS Mets at Dx
CS Mets Eval
CS Site-Specific Factor 1
CS Site-Specific Factor 2
CS Site-Specific Factor 3
CS Site-Specific Factor 4
CS Site-Specific Factor 5
CS Site-Specific Factor 6
CS Site-Specific Factor 7
CS Site-Specific Factor 8
CS Site-Specific Factor 9
CS Site-Specific Factor 10
CS Site-Specific Factor 11
CS Site-Specific Factor 12
CS Site-Specific Factor 13
CS Site-Specific Factor 14
CS Site-Specific Factor 15
CS Site-Specific Factor 16
CS Site-Specific Factor 17
CS Site-Specific Factor 18
CS Site-Specific Factor 19
CS Site-Specific Factor 20
CS Site-Specific Factor 21
CS Site-Specific Factor 22

Grade
Histologic Type ICD-O-3
ICCC Site Recode ICD-O-3 / WHO 2008
ICD Revision Number
ICD-O-3 Conversion Flag
IHS Link
Laterality
Lymphoma Subtype Recode / WHO 2008
Marital Status at DX
Month of Birth
Month of Conclusive DX
Month of Dx
Month of Multiple Tumors
Mult Tum Rpt as One Prim
Multiplicity Counter
NAPIIA Derived Pacific Islander Origin
NHIA Derived Hispanic Origin
Origin Recode NHIA (Hispanic, Non-Hisp)
Override Site / Type
Place of Birth
Primary Payer at DX
Primary Site
Race 1
Race 2
Race 3
Race 4
Race 5
Race Recode (W, B, AI, API)
Race Recode (White, Black, Other)
Reason For No Surgery
Regional Nodes Examined
Regional Nodes Positive
RX Summ -- BRM
RX Summ -- Chemo
RX Summ -- Hormone
RX Summ -- Other
RX Summ -- Radiation
RX Summ -- Scope Reg LN Sur

CS Site-Specific Factor 23
CS Site-Specific Factor 24
CS Site-Specific Factor 25
CS Size/Ext Eval
CS Tumor Size
CS Version 1st
CS Version Latest
Date of Initial RX -- SEER - Month
Date of Last Contact - Month
Date of Last Contact - Year
Day of Conclusive DX
Day of Multiple Tumors
Derived AJCC 6th Stage
Derived AJCC 7th Stage
Diagnostic Confirmation
DSS2000S
EOD10 Extension
EOD10 Extension Prostate Path
EOD10 Lymph Node Involvement
EOD10 Tumor Size
Ethnicity Source

RX Summ -- Surg / Rad Seq
RX Summ -- Surg Oth Reg/Dis
RX Summ -- Surg Prim Site
RX Summ -- Transplant / Endocr
RX summ--Systemic Surg Seq
SEER Record Number
SEER Summary Stage 1977
SEER Summary Stage 2000
SEER Type of Follow-up
Sequence Number - Central
Sex
Site Recode ICD-O-3 / WHO 2008
Spanish / Hispanic Origin
Tumor Marker 1
Tumor Marker 2
Tumor Marker 3
Type of Reporting Source
Vital Status
Vital Status Recode
Year of Birth
Year of Conclusive DX

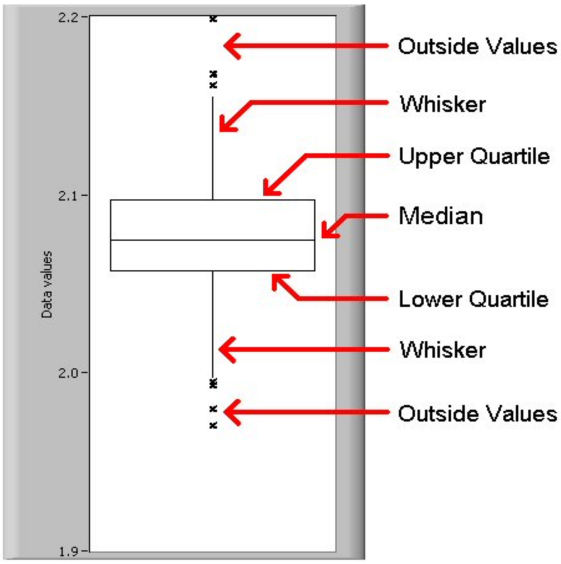

*Figure 1: A boxplot illustration. The ends of the vertical lines or "whiskers" indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range (IQR). Extreme outliers are the ones that falls 3 times IQR outside of the box boundary. Outliers are the ones that falls 1.5 times IQR outside of the box boundary.*