tttt

Iyad Alkrunz

2/23/2022

## Winsorization of outliers

There are two common methods for treating outliers in a data set. The first is to remove outliers as a means of trimming the data set. The second method involves replacing the values of outliers with suitable statistic such as mean, median, mode or quantile-based technique as follows:

In this technique the outliers are replaced with the arthematic mean of the remaining observations after removing outliers.

The median value that is the middle value in a ordered remaining observations

is used to replace the detected outliers.

The outliers are replaced with the mode value of the remaining observations, which is appears most often in a set of data values.

in this quantile-based technique, the maximum outliers are replaced with upper fence (capped), and the minimum outliers are replaced with lower fence (floored).

The following section investigates the effect of the four considered winsorization statistics on the performance of parameter estimates for different probability distributions via Monte carlo simulation.

# Simulation (Numerical Study)

An R code has been developed and implemented in Studio environment to generate random data sets from three different probability distributions namely, normal, negative binomial and exponential distribution.

## Settings of Data Generation

Data were generated with four different sample sizes, = 20,50,100 and 200, in such a way that () of data are generated from the original distribution () and the rest of data are generated from the contamination distribution (). Thus, the contaminated data structure can be formulated as , where is the contamination level and =0.05, 0.10 or 0.15.The following three probability distributions are considered:

### Normal distribution

For normal random variable, ; data were generated from the standard normal distribution with and . For contamination procedure; the contaminated data were generated from another normal distribution with and .

The maximum likelihood estimator () of the mean and standard deviation are obtained as the sample mean ,and , respectively.

### Negative binomial distribution

Random variable follows the negative binomial distribution with mean and variance if is the count of independent Bernoulli trials required to achieve the successful trials when the probability of success is a constant . The probability of trials is . The of is given by: .

For negative binomial random variable; data are generated with parameters and , while the contaminated data are generated from Poisson distribution with , where the probability of successes is .

### Exponential distribution

The Exponential distribution is the most commonly used model in reliability and life-testing analysis, (i.e for ). The of is given by .  
Data were generated with parameter , and the contaminated data were generated from exponential distribution with .

For each combination of the considered probability distributions,sample sizes, contamination levels and winsorization statistics; the generation procedures are repeated 1000 iterations to ensure the convergence.

## Performance indicators

The impact of the considered four outliers winsorization statistics on the parameter estimates are measured by three common indicators as follows:

1. , is the difference between the estimator’s expected value and the true value of the parameter being estimated.
2. , is a measure of the quality of an estimator. As it is derived from the square of Euclidean distance, it is always a positive value that decreases as the error approaches zero. where and are the true and estimated values of the considered parameters .
3. , are statistical tests aiming to determine whether a set of observed values match those expected under the applicable distribution. There are different goodness-of-fit tests, in this article the Shipiro-Wilk test is used in the case of normal and exponential distributions, while Kolmogorov-Smirnov test is used in the case of negative binomial distribution.

library(robust)

## Warning: package 'robust' was built under R version 4.1.2

## Loading required package: fit.models

## Warning: package 'fit.models' was built under R version 4.1.2

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning in as.POSIXlt.POSIXct(Sys.time()): unable to identify current timezone 'W':  
## please set environment variable 'TZ'

library(MLmetrics )

##   
## Attaching package: 'MLmetrics'

## The following object is masked from 'package:base':  
##   
## Recall

library(SimDesign)

## Warning: package 'SimDesign' was built under R version 4.1.2

##   
## Attaching package: 'SimDesign'

## The following objects are masked from 'package:MLmetrics':  
##   
## MAE, RMSE

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.1.2

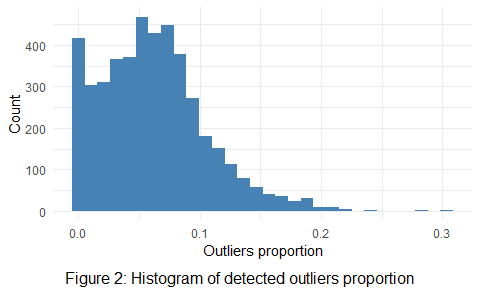
library(stringr)

# Application

Internet usage data set containing more than 2 million session records for 4500 random users are obtained for internet service provider company in Palestine through Ministry of Telecom and Information Technology of the State of Palestine. Each session has many features like start time, end time, traffic and duration.

In this study, we are interested only in sessions’ duration which are commonly hypothesized to be exponentially distributed (see Akmeroth and Ammaram, 1996, Sripanidkulchai, et al, 2004, Chetlapalli, et al, 2020). Consonance with that, we assume that session duration are exponentially distributed, therefore, the sessions rows are aggregated for each user. A total of 1416 (31.467%) of users sessions’ duration have been fitted by exponential distribution at 0.05 level of significance according to Shapiro-Wilk goodness-of-fit test.

The outlier of session duration for each user have been detected. Figure 2 presents the proportions of detected outliers for each user, it is ranged between 0% and 30% , with mean of 6% and positively skewed distribution.



Three winsorization methods are applied on users data with detected outliers, the summary of fitted users before and after winsorization is presented in Table 6. The results show that the proportion of fitted users data after winsorization are increase significantly, where the mean has the highest proportion, followed by median and then the qunatile-bases method which are consistent with the findings of the simulation study. The Chi-square test of independence shows that there are significant association between the status of users data (i.e fitted by exponential distribution) before and after winsorization at 0.05 level of significance,which reveal that insignificant number of the exponential fitted users data before winsorization has been alternated to be not fitted by exponential after winsorization has been conducted.