

Background

Numerous traditional statistical methods can be used to quantify parameters under certain assumptions. However, these methods are only true if the assumptions made in the statistical analysis are fulfilled. Classical statistics usually assumed a Gaussian distribution because it closely approximates a large number of distributions usually when the mean is much larger than the standard deviation, and so these techniques have found very wide applications. For the investigation of two-sample testing, where $H_0 : \mu_0 = \mu_1$, the most common methods include the Independent Sample t-tests which comprise of the Student-t and Welch-t.

However, real-world cases represent a paucity of obeying theoretical standards since data may demonstrate skewness or heavy-tail distribution. And as such, many would use transformations or robust estimates of the location to deal with these behaviors. More specifically, analysts would employ alternative methods such as Winsorized-t and α -trimmed-t (also called g-times trimmed-t or trimmed-t) because they are efficient and allows control over Type 1 error in nonnormal conditions.

A Gaussian distribution enables possible identification of outliers—ideally data points that are considered very low and very high within the data set. Observing these outliers can be advantageous to a study because understanding the fundamental reason as to why they were even recorded can, in turn, lead to more fitting statistical analysis and better inferences. Thus, asymmetrical trimming of special data sets can provide accurate statistical inferences than classical and common robust statistical methods because data sets are trimmed to target the ideal cases for analysis rather than symmetrically trimming or substituting a specific percentage of data.

Statistical Method

The following derivation of the method of α, β -trimmed means was proposed by Stephen M. Stigler from the University of Wisconsin, Madison (1973):

Let a sample X_1, X_2, \dots, X_n be n i.i.d. random variables.

By theory of order statistic, $X_{(1),n}, X_{(2),n}, \dots, X_{(n),n}$ are random variables in ascending order where $X_{(1),n}$ is $\min(X_{(i),n})$, $X_{(2),n}$ is the second smallest of X_i and so forth.

To trim, the α^{th} percentile are removed from the bottom portion of the ordered data while the β^{th} percentile are trimmed from the top.

- The α and β values are any desired integers assigned to be percentile cutoffs such that $0 < \alpha < \beta < 1$.

Arithmetic trimmed mean is then,

$$S_n = \frac{1}{([\beta n] - [\alpha n])} \sum_{i=[\alpha n]+1}^{[\beta n]} x_i$$

where $[\alpha n]$ and $[\beta n]$ are the floor functions to the number of data values that remain.

Variance of the trimmed mean is then,

$$s_{it\alpha\beta}^2 = \frac{\sigma^2}{(\beta - \alpha)} + \frac{1}{(\beta - \alpha)^2} [\beta(1 - \beta)(X_{[\beta n]} - \mu)^2 - 2\alpha(1 - \beta)(X_{[\alpha n]} - \mu)(X_{[\beta n]} - \mu) + \alpha(1 - \alpha)(X_{[\alpha n]} - \mu)^2]$$

Together, the t-test may be carried out with the usual testing hypothesis:

H_0 : There is no significant difference between the trimmed population scores, $\mu_1 = \mu_2$

H_1 : There is a significant difference between the trimmed population scores, $\mu_1 \neq \mu_2$

With $S_{(n,i)}$, $s_{it\alpha\beta}^2$, and h_i (trimmed sample size), the t-score is:

$$t_{(\alpha,\beta)} = \frac{(S_{n,1} - S_{n,2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_{1t\alpha\beta}^2}{h_1} + \frac{s_{2t\alpha\beta}^2}{h_2}}}$$

- It can be tested against the Student-t table with $[\beta n] - [\alpha n]$ degrees of freedom.

R code

```
ab.trim.Ttest = function(data1, data2, a, b){
  data1 = sort(data1)           # Order Statistics
  n = dim(as.data.frame(data1))
  data2 = sort(data2)           # Order Statistics
  n = rbind(n, dim(as.data.frame(data2)))

  an = c() ; bn = c();
  for (i in 1:2){
    an[i] = floor(a*n[i]+1)     # Locates minimum score at alpha-th percentile
    bn[i] = floor(b*n[i])       # Locates maximum score at beta-th percentile
  }

  Tdata1 = data1[an[1]:bn[1]]   # Trims dataset #1
  p = dim(as.data.frame(Tdata1))
  Tdata2 = data2[an[2]:bn[2]]   # Trims dataset #2
  p = rbind(p, dim(as.data.frame(Tdata2)))

  tm = c(); tv = c(); vtm = c();
  tm[1] = mean(Tdata1)
  tm[2] = mean(Tdata2) # Trimmed data
  tv[1] = var(Tdata1)
  tv[2] = var(Tdata2) # Variance of trimmed data
  # Trimmed Variance
  vtm[1] = (tv[1]/(b-a)) + (1/(b-a)^2) *
    ((b*(1-b) * ((data1[bn[1]] - tm[1])^2)) - ((2*a) * (1-b) * (data1[an[1]] - tm[1]) * (data1[bn[1]] - tm[1])) +
    (a*(1-a) * (data1[an[1]] - tm[1])^2))
  vtm[2] = (tv[2]/(b-a)) + (1/(b-a)^2) *
    ((b*(1-b) * ((data2[bn[2]] - tm[2])^2)) - ((2*a) * (1-b) * (data2[an[2]] - tm[2]) * (data2[bn[2]] - tm[2])) +
    (a*(1-a) * (data2[an[2]] - tm[2])^2))

  tt = (tm[1]-tm[2]) / sqrt((vtm[1]/p[1]) + (vtm[2]/p[2])) # Independent Sample t-test of trimmed data

  # ~~~~~Output~~~~~
  cat('
    Adaptive Trimmed Mean Two Sample t-test
  -----
  Data Sets | Trimmed Mean | Variance
  -----
  Data Set 1 | ',round(tm[1],2),'| ',round(vtm[1],2),'
  Data Set 2 | ',round(tm[2],2),'| ',round(vtm[2],2),'
  -----
  t-score based on trimmed mean = ',tt,'
  degrees of freedom = ',bn[1]-an[1],
  -----')
}
```

Example

The data set contain participants who were undergraduate students at a major university in the Northern Region who were enrolled in either mathematics or statistics course in the Summer of 2015. Some courses were strictly taught as a lecture-discussion course while others were hybrid mastery-based courses, which combine lectures and special computer software for teaching and active studying. Overall, except for the learning environment presented, there were no significant differences in variables between the instructional methods. This would include the demographics of the students (i.e. age, gender, race, prior knowledge as related to the course), availability of outside resources, and the representation of a professor-student relationship.

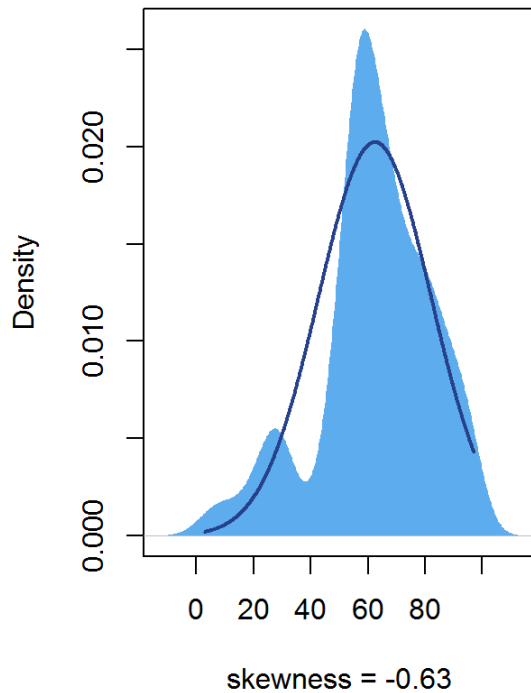
```
library(tidyverse)
df = read.csv("Trimmed means.csv", header = TRUE, sep = ",")
sets = split(df, df$trad_hyb)

data1 = sets[["trad"]][["trad_hyb_sco"]]
data2 = sets[["hyb"]][["trad_hyb_sco"]]
```

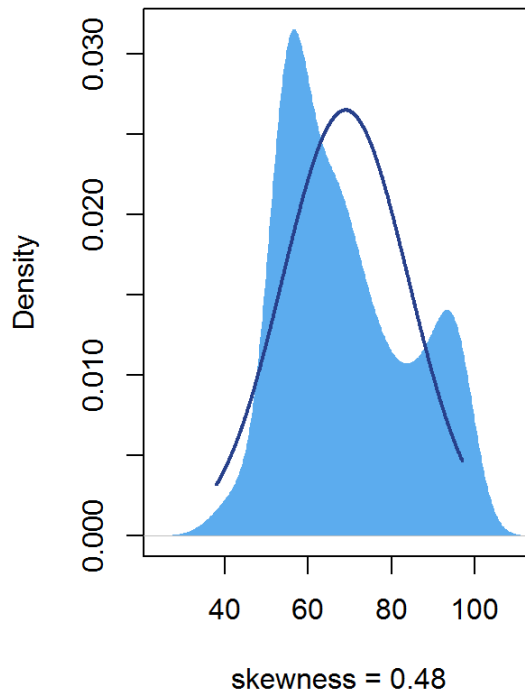
Upon statistical inspection of the students' performance scores from the traditional lecture and mastery-based courses, there is a need for asymmetrical trimming because of the real-world contamination: accelerated Summer courses and confidence possessed by students. Therefore, the assumption was based on that no matter which instruction mode a percentile of students were subjected to, they are likely to remain in the respected percentile of the other class because of their behaviors are assumed to be unique. This further suggests that the untrimmed portion of students is those where the variability matters and α, β -trimmed means would isolate these students for better statistical analysis.

```
par(mfrow = c(1,2))
rcompanion::plotNormalDensity(
  data1, main = "Traditional Courses",
  xlab = sprintf("skewness = %1.2f", psych::describe(data1)[11]),
  col2 = "steelblue2", col3 = "royalblue4")
rcompanion::plotNormalDensity(
  data2, main = "Mastery-based Courses",
  xlab = sprintf("skewness = %1.2f", psych::describe(data2)[11]),
  col2 = "steelblue2", col3 = "royalblue4")
```

Traditional Courses



Mastery-based Courses



Immediately, the traditional t-test statistic suggests that there is a difference between the result of students' final grades in a course. This indicates that neither mode of instructions is significant against a two-tail 95% confidence interval of 1.96. However, the trimmed means t-test, with a trimming of 20% from the bottom and 10% from the top, highlights that there are no differences between final grades. This further suggests that there is no difference in the ways students process information and challenge problems when taught in either a lecture-discussion or a mastery-based class.

```
t.test(data1, data2, var.equal = FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: data1 and data2
## t = -3.4016, df = 259.18, p-value = 0.0007758
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -10.36053 -2.76328
## sample estimates:
## mean of x mean of y
## 62.38095 68.94286
```

```
ab.trim.Ttest(data1, data2, 0.2, 0.9)
```

```
##
## Adaptive Trimmed Mean Two Sample t-test
## -----
## Data Sets | Trimmed Mean | Variance
## -----
## Data Set 1 | 66.37 | 304.15
## Data Set 2 | 69.95 | 398.3
## -----
## t-score based on trimmed mean = -1.505901
## degrees of freedom = 102
## -----
```

This was expected since the purpose of the college offering summer courses is to expedite the graduation date for students who are determined to put forth their best in such a rigorous course term. Thus, with performance being the dependent variables, a confident student in mathematics may remain the same no matter which mode of instruction they were exposed, in addition to, their likeliness of being successful in either class.

Works Cited

- Stigler, Stephen M. (1973) *The Asymptotic Distribution of the Trimmed Mean*. The Annals of Statistics 1.3: 472-477.