ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(4); 405-429

Drawing a Sample with Desired Properties from Population in R Package *"drawsample"*

Kübra ATALAY KABASAKAL * Tuba GÜNDÜZ **

Abstract
The aim of this study is to develop an R package called *drawsample*, which will be used to draw samples with the desired properties from a real data set. In accordance with the aim of the study, a sample with the desired properties can be drawn by purposive sampling with determining several conditions, such as deviation from normality (skewness and kurtosis) and sample size. Different applications of the package *drawsample* are illustrated using real data from the "Science and Technology(Score_1)" and "Social Studies (Score_2)" subtests of 6th Grade Public Boarding and Scholarship Examinations (PBSE). As the importance given to research with real data has increased in recent years, a good approach would be to draw a sample of the population. With this package, it is expected that researchers will draw samples as close as possible to the desired properties from the population or a large sample. It is thought that using the drawn samples obtained from real data with package *drawsample* will provide an alternative to simulation studies as well as a complement for these studies.

Key Words: R package "drawsample", distribution, real data, simulation.

INTRODUCTION

In the field of measurement and evaluation in education and psychology, the distribution of scores has an important role in the description of the groups. In addition to the description of groups, testing for normality to conduct many procedures of statistical inference, which are based on the assumption of normality, is crucial. However, as Erceg-Hurn and Mirosevich (2008) pointed out, the assumption of normality is rarely met when analyzing real data. Therefore, in applications, non-normal distributions are more common than normal distributions (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Geary, 1947; Micceri, 1989; Olivier & Norberg, 2010; Pearson, 1932). Due to the failure of the normality assumption, violation of normality, and distribution types have been the focus of many researchers working on important issues such as test equating, computer adaptive testing, differential item functioning, classification, and latent score estimation (Custer, Omar, &Pomplun, 2006; Finney & DiStefano, 2006; Gotzmann, 2011; Kieftenbeld & Natesan, 2012; Kirisci, Hsu, & Yu, 2001; Kolen, 1985; Kogar, 2018; Seong, 1990; Uysal, 2014; Yıldırım, 2015).

In the process of collecting data in a study, researchers may obtain different types of distributions. For example, most of the time, mathematics achievement scores differ from a normal distribution (skewed to the right) in selection exams (Ministry of National Education-MoNE, 2020; Student Selection and Placement Center- SSCP, 2019). If a researcher plans to conduct a study to investigate relations to antecedent and subsequent factors with mathematics scores obtained by a selection exam, and the statistical analysis intended to be used requires normality assumption, the researcher would not make use of the data because the results would be suspenseful. Since a sample selected from this data would also be skewed to the right, drawing a sample from this population will not solve the problem either. Otherwise, the scenario may be the opposite. For example, the aim of researchers may be to test the violations of the normality assumption in a psychometric analysis, and the data they collected may show normal distribution.

* Assist. Prof. Dr., Hacettepe University, Department of Educational Sciences, Ankara-Turkey, katalay@hacettepe.edu.tr, ORCID ID: 0000-0002-3580-5568
** Res. Assist., Gazi University, Gazi Faculty of Education, Ankara-Turkey, tubagunduz@gazi.edu.tr, ORCID ID: 0000-0002-0921-9290

To cite this article:
Atalay Kabasakal, K. & Gündüz, T. (2020). Drawing a Sample with Desired Properties from Population in R Package "drawsample". *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 405-429. doi: 10.21031/epod.790449

Received: 04.09.2020
Accepted: 22.12.2020

In order to conduct studies with different distribution types, generated data are used in simulation studies (Abdel-Fattah, 1994; Bıkmaz-Bilgen & Doğan, 2017; Dolma, 2009; Kaya, Leite, & Miller, 2015; Urry, 1974; Yıldırım Uysal-Saraç, & Büyüköztürk, 2018; Yoes, 1993). There are many software packages used to generate data with different distribution types such as normal, uniform, and skewed distributions. Bahry (2012), using a beta distribution, generated samples with three distribution types (extreme and

moderate skewness and a baseline condition) and seven sample sizes (from n = 100 to n = 3,000) by using WinGen 3.1 (Han, 2007). As an alternative to WinGen 3.1, SAS software (SAS Institute, 2009) can also be used to obtain different types of distribution. Gotzmann (2011) simulated normal and negatively skewed population distributions of ability parameters (N = 2,000,000). In his study, the population distributions of thetas were generated using the Normal Distribution function in SAS, and the negatively skewed distributions were created using the RAND Beta Distribution function in SAS (SAS Institute, 2009). Of these data, the ability parameter was determined to be appropriate for the purpose of his study, and random samples of different sample sizes (1,500 and 3,000) were selected. The use of beta distributions makes it easy to simulate skewed score distributions (Han & Hambleton, 2007). The components of beta distribution are parameters α and β. Some researchers draw a sample from the simulation data based on the desired properties. For this purpose, Fleishman's (1978) power method is suitable to draw a sample with skewed or platykurtic/flat distribution from the original data set (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2017; Kieftenbeld & Natesan, 2012; Sen, Cohen, & Kim, 2014; Stone, 1992).

Simulation methods are flexible and can be applied to a number of problems to obtain quantitative answers to questions that may not be possible to derive (Hallgren, 2013). Although simulation is a powerful technique, it has some limitations, which include difficulty in generalizing the results, organizing the results, and applying the results to real data (Wicklin, 2013). Simulation data provide a perfect fit that cannot be reached in real data. As Hallgren (2013) pointed out, real-world datasets are likely to be more "dirty" than the "clean" datasets that are generated in simulation studies, which are often generated under idealistic conditions which can be referred to as a perfect fit. Sireci (1991) stated that when real test data were not used, it was difficult to know whether the simulated data accurately reflected the characteristics of small sample data encountered in practice, and its validity could not be tested. Therefore, the use of real data has increased the importance of the studies conducted lately. In addition, some prestigious journals such as Educational Measurement: Issues and Practice (EM: IP) and the Journal of Educational and Behavioral Statistics (JEBS) have stated that simulation-based studies are from "examples of inappropriate manuscript topics" or considered to "have low priority," although most of the articles in these journals so far are simulation studies (American Educational Research Association, 2020; John Wiley & Sons Inc., 2019).

In empirical research, the process of data collection is challenging. The sample may not be representative of the population distribution; alternately, it may not be normally distributed, or it may be unsuitable for the desired distribution. To meet the assumption of normality in the literature, many studies in which the data set was manipulated have been found. For example, Gelbal (1994), in accordance with the purpose of his research, examined test scores, which included approximately two thousand fifth grade students who took both the Turkish language test and Math test. In order to get the desired distributions, approximately five hundred students from each test were removed. Doğan and Tezbaşaran (2003), in their study, selected participants with the required attributes to ensure the desired distribution. The researchers stated that random and purposive sampling techniques were used in the selection of the samples. For the purpose of their study, the students were drawn from a population consisting of students who had taken the Secondary Education Institutions Student Selection and Placement Examination in 2001. The samples were drawn randomly, right-skewed, left-skewed, flattened, and normal distribution, ranging in sample size from 2,353 to 29,244. In their study, in skewed samples, absolute values of skewness (±1.00) and kurtosis (1.37) were kept equal among samples to increase the accuracy for comparisons. Similar to the study of Doğan and Tezbaşaran (2003), Şahin ve Yıldırım (2018), obtaining the ability parameters, both right-skewed and left-skewed ability distributions were chosen from the real data. The real data were obtained from mathematics subtests of the Placement Test (SBS) applied in 2012. The selection of the right-skewed distributions was made randomly because it was originally a

right-skewed data set (skewness value=1.05). For the left-skewed data sets, the intended sample distribution was achieved through purposive sampling, and the groups whose skewness value is approximately -1.00 were chosen for all samples.

In addition to the above, in the literature, many researchers have chosen to draw samples from the real data set (population) in accordance with the purpose of their studies (Courville, 2004; Doğan & Kılıç, 2018; Fan, 1998; Nartgün, 2002; Reyhanlıoğlu Keçeoğlu, 2018). In the process of sampling from the population, it is important for future studies to have a function that makes the sample selection easier and brings it closer to the desired properties. In fact, it is suggested that the study of different abilities with non-normal distributions or samples with different levels of ability is the result of some research in the literature (Çelikten & Çakan, 2019). When the studies are examined, it was concluded that there is a need for a tool to enable researchers to draw samples with the desired properties from a large data set.

Purpose of the Study

In this study, the package *drawsample*, which aims to draw a sample based on the information of total score or ability parameter in accordance with the desired sample size and deviation from normality (skewness and kurtosis), was developed.With this package, it is expected that researchers will draw samples as close as possible to the desired properties from the population or a large sample, and it is thought that it will pave the way for the studies to be conducted on different topics based on the distribution in the literature. With this function, it is possible for researchers to draw samples with desired properties from large data in order to conduct statistical analysis under different conditions.

Fleishman's Power Method

In this section, Fleishman's (1978) power method, which is used to select the desired measures of deviation from normality (skewness and kurtosis), is explained briefly. Fleishman (1978) used a cubic transformation of a standard normal variable to create a distribution with pre-specified moments. Fleishman's (1978) power method, $Y = a + bz + cz^2 + dz^3$, was used to generate a non-normal distribution, where *Y* is a non-normal deviate with specified skewness and kurtosis. The value of z is a standard normal deviate, and a, b, c, and d are constants for transforming the standard normal variable to a variable with known skewness and kurtosis. (Kirisci, 2001). These constants for the normal distribution are 0.0, 1.0, 0.0, and 0.0 ($a = c$) respectively.

Fleishman (1978) tabulated these coefficient values for the selected skewness and kurtosis values. Writing the function in R, the values in this table were used to get the non-normal distributions. The values in this table can also be accessed using the *find_constants()* function in the "SimMultiCorrData" (Fialkowski, 2018) package in R. The *find_constants()* function is a function that calculates Fleishman's third or Headrick's (2002) fifth-order constants, converting a standard normal random variable into a continuous variable with a certain skewness and standardized kurtosis value. When the skewness value of the function is 0 and the standardized kurtosis value is 0, the usage example is given in Table 1.

Table 1. R Code to Find Fleishman's Third Order Constants

```
library(SimMultiCorrData)
find_constants(method = c("Fleishman"), skews = 0, skurts = 0)
## $constants
## c0 c1 c2 c3
##  0  1  0  0
##
## $valid
## [1] "TRUE"
```

Since the use of the function used in the example given in Table 1 extends the operation process, an R object named "constants_table" was created with the values obtained using this function.

Skewness and Kurtosis Statistics

The first four moments of the distribution are mean, variance, skewness, and kurtosis, respectively, which are the most important characteristic of frequency distributions (D'agostino, Belanger, & D'Agostino, 1990).

The following equations are for the third and fourth moments, skewness and kurtosis statistics, in Equation 1 and 2. These equations are used routinely; for example, SAS and SPSS give skewness and kurtosis statistics using them in their descriptive statistics output (D'agostino, Belanger, & D'Agostino 1990).

$$\text{skewness} = \frac{n\sum(X-\bar{X})^3}{(n-1)(n-2)S^3} \quad (1)$$

$$\text{kurtosis} = \frac{n(n+1)\sum(X-\bar{X})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (2)$$

There are many R packages to calculate the skewness and kurtosis values. In this study, the *describe()* function in the *psych* package was used to calculate skewness and kurtosis values. Table 2 shows the example of calculating descriptive statistics of the vectors of "normal_dis" and "skew_dis" generated by *rnorm()* and *rbeta()* functions, respectively.

Table 2. R Code to Calculate Descriptive Statistics of a Vector

```
library(psych)
set.seed(41)
normal_dis <- rnorm(1000)
describe(normal_dis)
##    vars    n mean sd median trimmed  mad   min  max range  skew kurtosis   se
## X1    1 1000    0  1   0.03    0.01 1.01 -3.26 3.33  6.58 -0.01    -0.03 0.03
skew_dis <- rbeta(1000,2,5)
describe(skew_dis)
##    vars    n mean   sd median trimmed  mad min  max range skew kurtosis se
## X1    1 1000 0.28 0.16   0.26    0.27 0.17   0 0.83  0.83 0.56    -0.19  0
```

As shown in Table 2, the *describe()* function has 13 different outputs. From the output of this function, the skewness and kurtosis values can be extracted, as shown in Table 3.

Table 3. R Code to Extract Skewness and Kurtosis Values

```
describe(normal_dis)$skew
## [1] -0.006120114
describe(normal_dis)$kurtosis
## [1] -0.03008443
```

Drawing Samples

The most commonly used function for selecting samples in R is the *sample()* function in the base package. This function takes a sample of the specified size from a determined vector using either with or without replacement

In this study, *sample_n()* function which is a function of *dplyr* package (Wickham, François, Henry, & Müller; 2019) is used to select samples. The *sample_n()* function has similar arguments with the

sample() function in the base package. The *sample()* function works with vectors, while the *sample_n()* function works with data sets. The *sample_n()* function has the "weight" argument instead of the "prob" argument in the *sample()* function. The value of the "weight" argument can be any column in the data set or data frame. In order to demonstrate the use of the *sample_n()* function, "example1" data set consisting of four variables with 100 observations was created. The variables in the data set "example1" are "id," "gender," "math_score" and "science_score." In order to create a new data frame with students who have higher science scores, the "weight" argument was used with the value of this variable (science_score). Table 4 shows the example of using *sample_n()* function.

Table 4. An Example of Using *sample_n()* Function

```
library(dplyr)
set.seed(41)
example1 <- data.frame( id=paste("id",101:200,sep=""),
gender = sample(c("F","M"),replace=TRUE,100),
math_score = sample(0:100,100,replace=TRUE),
science_score =sample(0:100,100,replace=TRUE))
summary(example1)
##       id                gender            math_score    science_score
##  Length:100         Length:100         Min.   : 1.00   Min.   : 0.00
##  Class :character   Class :character   1st Qu.:27.75   1st Qu.:26.75
##  Mode  :character   Mode  :character   Median :50.50   Median :57.00
##                                        Mean   :49.87   Mean   :52.88
##                                        3rd Qu.:71.25   3rd Qu.:76.00
##                                        Max.   :99.00   Max.   :98.00
example2 <- sample_n(example1, 10, weight = science_score)
summary(example2)
##       id                gender            math_score    science_score
##  Length:10          Length:10          Min.   :15.00   Min.   :32.00
##  Class :character   Class :character   1st Qu.:22.75   1st Qu.:43.00
##  Mode  :character   Mode  :character   Median :52.50   Median :72.00
##                                        Mean   :51.40   Mean   :63.30
##                                        3rd Qu.:76.75   3rd Qu.:79.75
##                                        Max.   :91.00   Max.   :89.00
```

In Table 4, "example1" data set was created, and summary information about the data set was printed. While creating the "example2" data set, the students were weighted according to the "science score" variable, and the sampling was selected. When the summary information about "example2" data set is examined, it is seen that the minimum, quartiles, median, and median values of "science_score" are higher than "example1".

In the *drawsample* package, the *draw_sample()* function has been improved to get a sample with the desired distribution properties and sample size in accordance with skewness and kurtosis. The code belonging to this function is explained below.

R CODE FOR *draw_sample()* FUNCTION

draw_sample() function with 6 arguments was written to draw a sample with the desired properties. The arguments of the function are given in Table 5.

Table 5. The arguments of *draw_sample()* function

Argument	*Value*
dist	data frame: consists of id and scores with no missing
n	numeric: desired sample size
skew	numeric: the skewness value
kurts	numeric: the kurtosis value
replacement	logical: sample with or without replacement? (default is FALSE).
output_name	character: a vector of two components. The first component is the name of the output file, user can change the second component.

When determining "skew" and "kurts" from the arguments in Table 5, the Fleishman Power Method Weights table must be consulted. Fleishman coefficients corresponding to some combinations, such as skewness value 1 and kurtosis value 0, are absent. The minimum and maximum values of the kurtosis coefficient corresponding to a determined skewness coefficient are presented in this table created by using the Flesihman's (1978) Power Method Weights Table. For example, if the skewness coefficient is selected as 2, the kurtosis coefficient must be entered between 5 and 20. In other words, the minimum and maximum value of kurtosis values corresponding to each skewness coefficient that can be used are presented in Table 6.

Table 6. Minimum and Maximum Kurtosis Coefficient Corresponding to the Skewness Coefficient

Skewness	*Kurtosis (min)*	*Kurtosis (max)*	*Skewness*	*Kurtosis (min)*	*Kurtosis (max)*
0	-1.2	20	1.9	4.4	20
0.1	-1.2	20	2	5	20
0.2	-1.1	20	2.1	5.6	20
0.3	-1.1	20	2.2	6.3	20
0.4	-0.9	20	2.3	7.1	20
0.5	-0.8	20	2.4	7.8	20
0.6	-0.6	20	2.5	8.6	20
0.7	-0.4	20	2.6	9.5	20
0.8	-0.2	20	2.7	10.4	20
0.9	0.1	20	2.8	11.4	20
1	0.4	20	2.9	12.4	20
1.1	0.7	20	3	13.4	20
1.2	1	20	3.1	14.4	20
1.3	1.4	20	3.2	15.5	20
1.4	1.8	20	3.3	16.5	20
1.5	2.3	20	3.4	17.6	20
1.6	2.7	20	3.5	18.8	20
1.7	3.2	20	3.6	19.9	20
1.8	3.8	20			

R commands for *draw_sample()* function are given in Table 7. In this function, the value of the "dist" argument must be a data frame that has two columns. Note that the data includes student IDs in the first column and student total test scores or abilities (thetas) in the second column. For that purpose, with the command of names(dist), the columns of the imported object columns in the R environment are named "id" and "x" (Table 6, Line 8). Then, the x is extracted as the variable x" in Line 10, so "x" becomes a vector that can provide convenience. If "n" from the arguments of the function, the desired sample size, is larger than the length of the data, it gives the following error: "Cannot take a sample larger than the length of the data". For example, although the sample size of the imported data is 1,000 and users desire to take sample size 2,000, the function gives the error and stops running (Lines 13 to 16).

The values in Fleishman's (1978) table are used to get a sample with the desired distribution properties. These values are found in the object called "constants_table" in the package. *SimMultiCorrData* (Fialkowski, 2018) package was used in the preparation of the table including these constants. In this object, there are b, c, and d constants belonging to a set of 5,292 lines consisting of kurtosis values corresponding to each skewness value and consisting of skewness values increasing by 0.1 units from 0

to 3.6. This table includes only 0 and positive skewness values and corresponding kurtosis and constants. Between lines 18-22, if the user enters the skew as a negative value, "Skew" and "c" columns of the table by multiplying -1 will be rearranged with *if* statement. If the skewness value given by the user is not included in the table or if the kurtosis value corresponding to that value is not found in the table even if the skewness value is found in the table, the function stops working with an error "No valid power method constants could be found for the specified values. Change the values" (Line 25 to 32). It is suggested to use Fleishman's (1978) table when making choices with regard to skewness and kurtosis values. Within the *repeat* loop, the reference distribution with the skewness and kurtosis values entered by the user between Line 38 and Line 53 is formed. According to the minimum and maximum values of the distribution formed in this loop and then included in the user's data set (Line 65), the rescaled "reference_v4" distribution forms the basis for the function's work. Before the *repeat* loop, an empty vector was created to form a distribution with the skewness and kurtosis values entered by the user. Firstly, an object with a normal distribution called "reference" with a mean of 0 and a standard deviation of 1 is formed in the loop (Line 41). Within the *repeat* loop, the "reference_v2" object is formed by multiplying the "reference" object by the b, c, and d coefficients in the table, respectively. When the skewness and kurtosis values of the "reference_v2" object are equal to the skewness and kurtosis values entered by the user, the loop is stopped, and the "reference_v2" object is assigned to "reference_v3" (line 50). If the calculated values are not equal to the values defined by the user, the "reference_v3" object is left empty, and the loop is repeated. With the *draw_sample()* function, it is aimed to form a similar distribution from the values in the user's data set based upon the "reference_v4" object formed in accordance with the values entered by the user. On lines 67-69, the outputs of the *hist (reference_v4)* function are used for this purpose. The starting and ending points of each bar of the histogram are assigned to "x_break" objects, the number of bars in the histogram to "n_break" objects, and the number of elements in each bar to "x_counts" objects.

The vector "x" is categorized by "x_break" and identified as "x_v1". The categorized object is added as a new column to the user's data set. The information about how many individuals are in each category is assigned to the "x_n" object. The specified operations are defined between 71-73. The information on how many individuals there are in each category is crucial in terms of determining whether the function will select the sample of the user's desired properties without resampling. When the number of individuals in each category in the data set is higher than in each category of the reference distribution, the function can be performed without resampling, with the default value of the *"*replacement" argument. This situation is checked between lines 73 and 79. If the number of the individuals in at least one category in the data set is less than the number of the individuals in the relevant category of the reference distribution, the function gives an error: "Cannot take a sample form that data without replacement. Please change replacement = TRUE." In this situation, the function can be used by changing the value of the *"*replacement" *argument*. The codes working up to line 83 have been written in order to prepare for drawing sample. The drawing sample process is carried out through the *for* loop between 89-105. For data manipulation in the loop, *filter()* and *sample_n()* functions in the package of *dplyr* (Wickham, François, Henry, & Müller; 2019) are used. The scores belonging to the individuals to be formed in the *for* loop were created in the "new_sample" and the empty matrices named "ID_list" for the identity information of the individuals on lines 83 and line 84. In both matrices formed, the number of lines was determined as the number of categories ("n_break") and the number of columns as the maximum number of individuals in these categories.

Table 7. R Commands for *draw_sample()* Function

```
draw_sample <-  function(dist,n,skew,kurts,
                         replacement =FALSE,
                         output_name = c("sample","default")){

  # rename the data
  skew <- round(skew,1)
  kurts <- round(kurts,1)
  names(dist) <- c("id","x")
  # extract x column
  x <- dist$x


  N <- length(x)
  if(n >= length(x)){
    stop("Cannot take a sample larger than the length of the data")
  }

  # arrange table for negative skewness
  if(skew<0){
    constants_table$c <- -1*constants_table$c
    constants_table$Skew <- -1*constants_table$Skew
  }


  if(skew %in% constants_table$Skew == FALSE){
    stop("No valid power method constants could be found for
          the specified values. Change the values")
  }else if (skew %in% constants_table$Skew == TRUE &
  kurts %in%  constants_table[constants_table$Skew==skew,]$Kurtosis == FALSE){
    stop("No valid power method constants could be found for
           the specified values.Change the values")
  }

  reference_v3 <- NULL

  # conduct Fleishman's power method for the specified
  # skewness and standardized kurtosis
  repeat{
    for( i in 1:dim(constants_table)[1]){
      #  random generation for the normal distribution
      reference <- stats::rnorm(n,0,1)
      constants <- constants_table[i,3:5]
      b <- constants$b
      c <- constants$c
      d <- constants$d
      reference_v2 <- -c + b*reference + c*(reference^2) + d*(reference^3)
      skew_value <- round(psych::describe(reference_v2)$skew,1)
      kurt_value <- round(psych::describe(reference_v2)$kurtosis,1)
      if(skew_value == skew & kurt_value == kurts){
        reference_v3 <- reference_v2
        break
      }
    }
    if(is.null(reference_v3) == FALSE)
      break
  }

  # Rescale the reference vector to have specified minimum and maximum
  scale_ref <- function(x, from, to) {
    x <- x - min(x)
    x <- x / max(x)
    x <- x * (to - from)
    x + from
  }
  reference_v4 <- scale_ref(reference_v3, from=min(x),to=max(x))

  x_counts <-  graphics::hist(reference_v4)$counts
  n_break <-   length(graphics::hist(reference_v4)$breaks) -1
  x_break <-   graphics::hist(reference_v4)$breaks

  x_v1 <- as.numeric(cut(x,x_break,include.lowest = TRUE))
  dist2 <- data.frame(dist,x_v1)
  x_n <- unname(table(x_v1))
```

```
  control <- sum(x_n>= x_counts)
  if(control!=length(x_counts)){
    if(replacement==FALSE){
      stop("Cannot take a sample form that data without replacement.
           Please change replacement=TRUE")
    }
  }

  new_sample <-  matrix(NA, nrow = n_break, ncol = max(x_counts))
  ID_list <-  matrix(NA, nrow = n_break, ncol = max(x_counts))

  new_sample_2 <- list()
  ID_list_2 <- list()

  for(i in 1:n_break){
    new_count <- 0
    j <- 0
    while(new_count < x_counts[i]){
      j <- j + 1
      IDx <- dplyr::filter(dist2,x_v1==i)
      IDx <- dplyr::sample_n(IDx,1)
      if(replacement==FALSE){
        dist2 <- dplyr::filter(dist2,id!=IDx$id)
      } else{ dist2 <- dist2}
      new_count <- new_count + 1
      new_sample[i,j] <- IDx$x
      ID_list[i,j]<-   IDx$id
    }
    new_sample_2[[i]] <-  stats::na.omit(new_sample[i,])
    ID_list_2[[i]] <-  stats::na.omit(ID_list[i,])
  }

  new_sample_3 <- unlist(new_sample_2)
  ID_list_3 <- unlist(ID_list_2)

  S1 <- data.frame(id=ID_list_3,x=new_sample_3)

  # Save the output
  if (output_name[2] == "default") {
    wd <- paste(getwd(), "/", sep = "")
  }else {wd <- output_name[2]}
  fileName <- paste( output_name[1], wd,".dat", sep = "")
  utils::capture.output(data.frame(S1), file = fileName)


  # Organize the output
  dist3 <-  dplyr::select(dist2,id,x)
  dist3 <-  dplyr::mutate(dist3,type="population")
  S2   <-   dplyr::mutate(S1,type="sample")
  result <-  rbind(dist3,S2)

  # to capture the graph
  graph <-      lattice::histogram(~x|type,data=result,xlab="Score",
                 nint = n_break,
                scales = list(x = list(tick.number = 5,relation = "free")))

  lattice::trellis.device(device="png",
           filename=paste( output_name[1], wd,".png", sep = ""))
  print(graph)
  grDevices::dev.off()

  desc <-  rbind(psych::describe(x),
                 psych::describe(reference_v4),
                 psych::describe(S1$x))[,c(2:4,8:9,11:12)]
  rownames(desc) <- c("population","reference","sample")
  # output with three components
  output <- list(desc =desc ,
                 sample = tibble::as_tibble(data.frame(S1)),
                 graph = lattice::trellis.last.object()
  )

  return(output)
}
```

With the *while* loop in the *for* loop that repeats as many as the number of categories, the number of individuals required to be included in each category of the reference vector divided into categories continues until the required number of people is reached by selecting one by one from the relevant category of the data set with the *sample_n()* function. The code in the *while* loop works in two different ways depending on the value of the *"*replacement" *argument*. If the value of the argument is FALSE, the person selected for sampling once is not included in the resampling and is removed from the data (line 96 to 98); if the value of the argument is TRUE, this part of the code will not work. Each line of the "new_sample" and "ID_list" objects formed within the *while* loop contains the "NA" value (max (x_counts) - x_counts [i] values), except for the amount of data that should be in each category. In the *for* loop, the missing data of the objects in the matrices formed in the *while* loop are removed and assigned to the empty list objects formed in 86-87 lines. After the loops were completed, lists containing information about each individual were selected for the sampling, and that person was assigned to the vector objects (Lines 107-108). After that, these vectors are combined in the S1 data set to form the desired sample.

Between lines 141 and 144, the output of the function is formed. The output, which is a three-component list, consists of descriptive statistics of the data and sample, the sample formed and the histogram graphs of the data, and the distribution of the sample. Descriptive statistics, which are the first component of the list, were formed between lines 136-139 by using the *describe()* function in the *psych* package (Revelle, 2018), the *graph*, which is the third component, was formed by using the *histogram()* function in the "lattice" package (Sarkar, 2008) between 120-134 lines. The *desc* component consisting of descriptive statistics information is a matrix. This matrix includes the mean, standard deviation, skewness, and kurtosis of the population, sample and the reference distribution. The second component is called *sample*, and it is from the *tibble* package (Wickham, Francois, and Müller; 2016). It is situated between 112-117 lines required to extract this data. It includes ids and x scores which are sampled. The third component is called "graph," and it includes two histogram graphs one is for "population" (imported data), and one is for the "sample" (extracted data). The third component of the output is also extracted.

EXAMPLES WITH REAL DATA

In the examples, related functions and outputs are presented based on the "Science and Technology" and "Social Studies" subtests data of the 6th Grade Public Boarding and Scholarship Examinations (PBSE) in 2013. At the secondary school level, the PBSE test consists of 100 multiple-choice test items, which include 25 items in each subtest (Turkish, Mathematics, Science and Technology, and Social Studies). It was administered in two booklet types, A and B (MoNE, 2013).

In 2013, 242,598 students participated in PBSE at the 6th-grade level, and 121,523 (50.09%) received booklet A. Of the students, 133,866 (55.18%) were female and 108,732 (44.82%) were male students. Within the scope of the study, randomly selected 5,000 students taking booklet A were considered as the "population." Of this group, 2,745 (54.90%) are female students. The data were obtained by the Directorate General for Measurement, Assessment and Examination Services of the Ministry of National Education in accordance with written permission. The total score distributions for each test were examined. Then, two datasets were used for the demonstration. The Science and Technology subtest was chosen as an example of left-skewed distribution. The Social Studies subtest was used as an example of platykurtic distribution. In each example, a sample of 500 students was drawn from the population for the related subtests. That the samples have the desired properties in terms of distribution type was taken into consideration. The functions and outputs for this process were given in Tables 8-14 and Figures 1-3.

In the first two examples, particular importance has been given to draw samples with a normal distribution and both negatively skewed and leptokurtic distribution from the data of the science and technology subtest, respectively. The command for the first example is shown in Table 8.

Table 8. Draw the Sample with a Normal Distribution of 500 Students from the Total Score Distribution of Science and Technology

```
install.packages("drawsample")
library(drawsample)
data(example_data)
Score1_normal <- drawsample::draw_sample(dist= example_data [,c(1,2)],
 n=500, skew = 0, kurts = 0,output_name = c("sample","1"))
```

First, the package *drawsample* is installed and then loaded. After then the object "example_data" which is automatically provided by the package is loaded. It has three columns including the total scores of the PBSE subtests of 5,000 students and IDs. The first column contains IDs (1: 5000), the second column contains the total scores of "Score_1 (Science and Tecnology subtest) ", and the third column contains "Score_2 (Social Studies subtest)" respectively.

In the function *draw_sample()* given in Table 8, the value of the "dist*"* argument contains the first and the second columns of "example_data [,c(1,2)]". In the output produced by the function in this model, the IDs and total scores of 500 students are recorded in the working directory of "Sample1.dat" extracted by "Sample1.png", which includes histogram graphs of the "population" of 5,000 students and the of 500 students with a distribution close to normal distribution. Table 9 shows the descriptive statistics of the distribution of 500 students drawn from the total score distribution of Science and Technology and the output of some of the students in the sample extracted.

Table 9. Outputs of the Distribution of 500 Students Drawn from the Total Score Distribution of Science and Technology

```
> Score1_normal
$desc
             n  mean   sd min max  skew kurtosis
population 5000 14.61 4.90   0  25 -0.40    -0.35
reference   500 13.21 4.58   0  25  0.04    -0.03
sample      500 13.73 4.59   0  25  0.01    -0.13

$sample
# A tibble: 500 x 2
      id     x
   <dbl> <dbl>
 1   416     2
 2   456     2
 3  3169     0
 4  4918     2
 5  4411     3
 6  4847     4
 7  4752     3
 8  1159     3
 9  4018     4
10  2963     4
# ... with 490 more rows

$graph
```

When the descriptive statistics in Table 9 were examined, the distribution of the total score of the Science and Technology subtest of 5,000 students was left-skewed, and the drawn sample was very close to the normal distribution. Figure 1 shows the "Sample1.png" which includes histogram graphs.

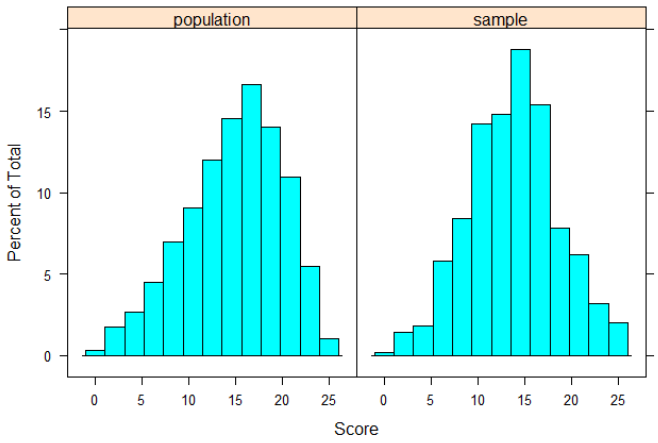

Figure 1. Histograms for the "population" and the "sample" Desired to Have Normal Distribution (Sample1.png)

As seen in Figure 1, the drawn sample distribution given according to the total scores of Science and Technology subtest was very close to the normal distribution. The command for this second example is shown in Table 10. In the second example, different from example 1, the value of skew and kurts are changed to -1 and 2, respectively.

Table 10. A sampling of 500 Students with Left Skewed and Leptokurtic Distribution from the Total Score Distribution of Science and Technology

```
Score1_nskew_lepto<- draw_sample(dist= example_data [,c(1,2)], n=500,skew = -1, kurts =
5,
 output_name = c("sample","2"))
```

Table 11 shows the descriptive statistics of the distribution of 500 students from the total score distribution of Science and Technology and the output of some of the students in the sample drawn.

Table 11. Outputs of the Distribution of 500 Students Drawn From the Total Score Distribution of Science and Technology.

```
Score1_nskew_lepto
## $desc
##                n  mean   sd min max  skew kurtosis
## population 5000 14.61 4.90   0  25 -0.40    -0.35
## reference   500 15.71 2.40   0  25 -0.95     5.03
## sample      500 16.24 2.43   2  25 -0.75     3.48
##
## $sample
## # A tibble: 500 x 2
##       id     x
##    <dbl> <dbl>
##  1  1124     2
##  2   768     8
##  3    82     8
##  4  2748     7
##  5  3196    10
##  6  3049     9
##  7  3456    10
##  8  3319    10
##  9  3942     9
## 10  2558    10
## # ... with 490 more rows
##
## $graph
```

When Table 11 is examined, although the skewness of "sample 2" is almost the same as the skewness of the "reference", the "sample 2" is flattened than the "reference". Figure 2 shows "Sample2.png" which includes histogram graphs for this model.

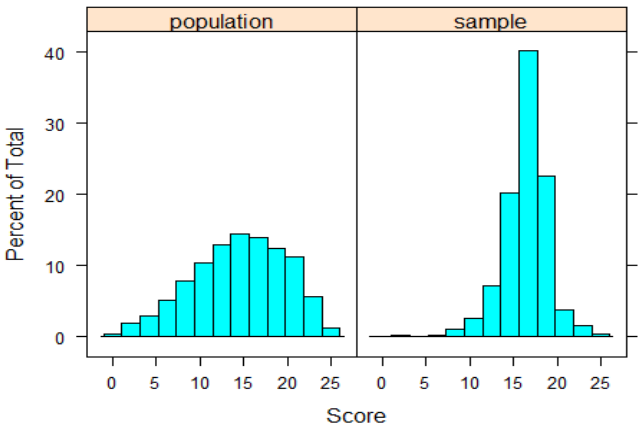

Figure 2. Histograms for the "Population" and the "Sample" Desired with Left Skewed and Leptokurtic Distribution (Sample2.Png)

The next two examples for real data were to draw sample with right-skewed and leptokurtic distribution (skewness value is =1.5 and kurtosis value=3) drawn from the distribution given according to the total scores of Social Sciences subtest. The command required for this situation is presented in Table 12.

Table 12. A sampling of 500 Students with Right Skewed and Leptokurtic Distribution Drawn from the Sum Score Distribution of Social Sciences

```
Score2_pskew_lepto<- draw_sample(dist= example_data [,c(1,3)], n=500,skew = 1.5, kurts =
3, output_name = c("sample","3"))
```

When the code in Table 12 is set to work, since the function cannot draw the data with the desired properties from the provided data without resampling, it gives an error and suggests allowing resampling. The argument "replacement," which is FALSE by default, has been replaced to meet the distribution conditions set out in Table 13.

Table 13. Sampling with Replacement of 500 Students with Right Skewed and Leptokurtic Distribution Drawn from the Sum Score Distribution of Social Sciences

```
Score2_pskew_lepto<- drawsample::draw_sample(dist= example_data [,c(1,3)], n=500,skew =
1.5, kurts = 3,replacement = TRUE,
 output_name = c("sample","3"))
```

Resampling is allowed when "TRUE" is entered in the "replacement" argument. In other words, an individual selected from "population" to "sample" is allowed to be repeatedly selected to provide the desired distribution.

Table 14 shows the descriptive statistics of the distribution of 500 students drawn from the total score distribution of Social Sciences and the output of some of the students in the sample extracted. In this case, the dist data frame contains the columns "ID" and "Score_2", which are used for defining the student identity and total score of the Social Studies subtest.

Table 14. Outputs of the Distribution of 500 Students Drawn from the Total Score Distribution of Social Sciences

```
> Score2_pskew_lepto
$desc
             n  mean   sd min max  skew kurtosis
population 5000 12.78 5.22   0  25 -0.17    -0.88
reference   500  6.79 3.72   0  25  1.49     2.96
sample      500  7.32 3.78   0  25  1.39     2.75

$sample
# A tibble: 500 x 2
      id     x
   <dbl> <dbl>
 1  3756     1
 2   390     1
 3  1483     1
 4  2855     2
 5  4792     1
 6  3660     2
 7  3937     1
 8  4280     2
 9   930     0
10  4324     2
# ... with 490 more rows

$graph
```

When the descriptive statistics in Table 14 were examined, the total score distribution of the Social Sciences subtest of 5,000 students (population) is slightly left-skewed; the desired sample is right-

skewed. On the other hand, the "population" s distribution shape is flatty, but the "sample" distribution shape is leptokurtic. Figure 3 shows the "Sample3.png" which includes histogram graphs.

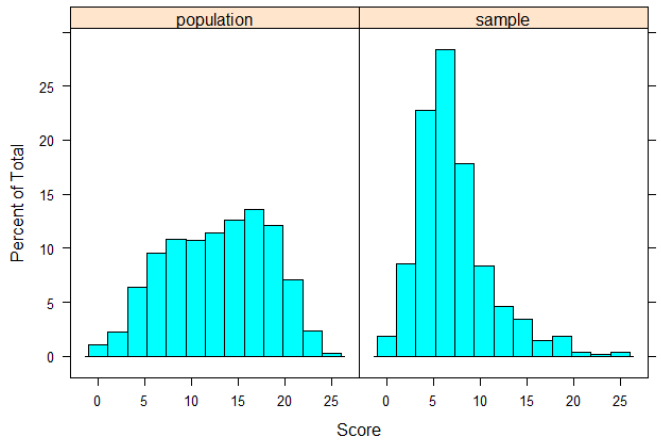

Figure 3. Histograms for the "population" and the "sample" Desired to have Skewed Distribution with Replacement (Sample3.png)

Evaluating the Function's Stability

Measures of kurtosis and skewness are used to determine if indicators met normality assumptions (Kline, 2005). The extent to which a frequency distribution diverges from symmetry is described as skewness. Kurtosis is a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. A perfect symmetrical distribution will have a skewness of 0 and a kurtosis of -3 ('excess' kurtosis of 0). The original kurtosis value is sometimes called kurtosis (proper), and West, Finch, & Curran (1995) proposed a reference of substantial departure from normality as an absolute kurtosis (proper) value > 7. Most statistical packages such as SPSS provide 'excess' kurtosis obtained by subtracting 3 from the kurtosis (proper). In this study, 'excess' kurtosis is used for practical reasons. Distributions that are more flat-topped than normal distributions are called platykurtic, and their kurtosis values are less than 3. Distributions that are less flat-topped than normal distributions are called leptokurtic, and their kurtosis values are more than 3 (Flott, 1995; Wuensch, 2005).

There is no consensus about the skewness and kurtosis values which indicate normality in the literature. It is widely accepted that absolute skew and kurtosis values up to one provide normality. (Büyüköztürk, Çokluk, & Köklü, 2014; Huck, 2012; Ramos et al., 2018). Furthermore, there are some suggestions that much larger values of skewness and kurtosis indicate normality (Brown, 2006; Kim, 2013; West et al., 1995). Furthermore, kurtosis is generally interesting only when dealing with approximately symmetrical distributions. Skewed distributions are always leptokurtic. Besides, kurtosis can be thought of as a measurement which adjusts to remove the effect of skewness (Blest, 2003). Moreover, social science researchers are concerned with the deviation of the distribution from symmetry rather than its flatness. In addition, high kurtosis should be considered for the researcher to look for outliers in one or both tails of the distribution (Wuensch, 2005). For this reason, although the possible skewness and kurtosis values can be selected in the *draw_sample()* function, the data provided by the function provides very close results in the skewness values, but not in the kurtosis values. We recommend that users should choose kurtosis values closest to 0 for normal distributions and higher than 3 for leptokurtic distributions, and lower than 3 for platykurtic distributions. If the aim of the researcher is to obtain data with outliers, the value of kurtosis can be increased up to 20 according to the number of outliers.

In order to determine how close the drawn sample to the reference distribution, a function called *draw_sampleRMSE()* is written. This function can take samples from the data with different set.seed values as much as the specified number of replications. The functions' output is the skewness and

kurtossis values for each replication of *draw_sample() function*. R commands for *draw_sampleRMSE()* function are given in Table 15.

Table 15. R Commands for *draw_sampleRMSE()* Function

```
## A function for running the draw_sample() function
draw_sampleRMSE <- function(df,rep=rep,n=n, skew=skew,kurts=kurts){
  skew.rep <- c();kurts.rep <-c()
  i=1
  while(i < (rep+1)) {
    skip_to_next <- FALSE
    tryCatch({
      set.seed(sample(1:10000,1))
      result <- drawsample::draw_sample(dist =df, n = n,
                                        skew = skew,kurts = kurts, output_name = c("samp
le",paste(i)))$desc },
      error = function(e) { skip_to_next <<- TRUE})
    if(skip_to_next) { next }
      if(is.na(result[3,6])==FALSE){
      skew.rep[i] <- result[3,6]
      kurts.rep[i] <- result[3,7]
      i= i +1
    }else{i=i}
  }
  return(data.frame(skew.rep,kurts.rep))
}
```

To illustrate the stability of *draw_samples()*, two simulated datasets are used. First, negatively skewed and platykurtic data was generated with a sample size of 10000 by using *rbeta()* function, called "datfra".

Then, 100 different samples were drawn from "datfra" with a different set.seed values with *draw_sampleRMSE()* function. After calculating the skewness and kurtosis values for each sample, the RMSE values and descriptive statistics were presented in Table 16 for skewness values, and only descriptive statistics were presented for kurtosis

Table 16. Drawing Samples from Negatively Skewed and Platykurtic Distribution

```
# Drawing samples from negatively skewed and platykurtic distribution
datfra <- data.frame(id=1:10000, x = rbeta(10000,1,0.2))
# a. draw normal distribution with skew=0 & kurts=0
sim_1 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=0,kurts=0)
attach(sim_1)
skew=0
psych::describe(sim_1$skew.rep,skew=FALSE)
##    vars   n mean   sd   min  max range   se
## X1    1 100 0.04 0.07 -0.09 0.19  0.29 0.01
psych::describe(sim_1$kurts.rep,skew=FALSE)
##    vars   n  mean   sd   min  max range   se
## X1    1 100 -0.12 0.17 -0.56 0.16  0.71 0.02
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_1$skew.rep)
## [1] 0.07759446
# b. draw positively skewed and leptokurtic skew=1 & kurts=5
sim_2 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=1,kurts=5)
skew=1
psych::describe(sim_2$skew.rep,skew=FALSE)
##    vars   n mean   sd min  max range   se
## X1    1 100 0.87 0.12 0.6 1.19  0.59 0.01
psych::describe(sim_2$kurts.rep,skew=FALSE)
##    vars   n mean   sd  min max range   se
## X1    1 100 3.63 0.48 2.03 4.8  2.77 0.05
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_2$skew.rep)
## [1] 0.1719165
# c. draw negatively skewed and platykurtic skew=-0.5 & kurts=1.5
sim_3 <- draw_sampleRMSE(df=datfra,rep=100,n=300, skew=-0.5,kurts=1.5)
skew=-0.5
psych::describe(sim_3$skew.rep,skew=FALSE)
##    vars   n  mean   sd   min   max range   se
## X1    1 100 -0.38 0.09 -0.56 -0.19  0.37 0.01
psych::describe(sim_3$kurts.rep,skew=FALSE)
##    vars   n mean  sd  min  max range   se
## X1    1 100 0.99 0.3 0.25 1.66  1.41 0.03
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_3$skew.rep)
## [1] 0.1525628
```

In the first example in Table 16, normal distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean of skewness and kurtosis values of the distributions produced in this example are quite close to the determined value, 0. The skewness values vary between -0.09 and 0.19, and kurtosis varies between -0.56 and 0.16. RMSE calculated for the skewness value was determined as 0.078.

In the second example in Table 16, positively skewed and leptokurtic distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean skewness value of the distributions produced in this example is quite close to the determined value, 1. However, the mean kurtosis value of the distributions produced in this example is larger than 3, as expected for leptokurtic distributions. The skewness values vary between 0.6 and 1.19, and RMSE calculated for the skewness value was determined as 0.172.

In the third example in Table 16, negatively skewed and platykurtic distributions are drawn from the negatively skewed and leptokurtic distribution. It is seen that the mean skewness value of the distributions produced in this example is quite close to the determined value, -0.5. However, the mean kurtosis value of the distributions produced in this example is smaller than 3, as expected for platykurtic

distributions. The skewness values vary between -0.56 and -0.19, and RMSE calculated for the skewness value was determined as 0.153.

Second, positively skewed and platykurtic data was generated with a sample size of 10000 by using *rbeta()* function, called "datfra2".Then, 100 different samples were drawn from the "datfra2" with a different set.seed values with *draw_sampleRMSE()* function. After calculating the skewness and kurtosis values for each sample, the RMSE values and descriptive statistics were presented in Table 17 for skewness values, and only descriptive statistics were presented for kurtosis.

Table 17. Drawing Samples from Negatively Skewed and Platykurtic Distribution

```
# Drawing samples from positively skewed and platykurtic distribution
datfra2 <- data.frame(id=1:10000, x = rbeta(10000,1,6))
# d. draw positively skewed and leptokurtic skew=2 & kurts=5
sim_4 <- draw_sampleRMSE(df=datfra2,rep=100,n=300, skew=2,kurts=5)
skew=2
psych::describe(sim_4$skew.rep,skew=FALSE)
##    vars   n mean   sd  min  max range   se
## X1    1 100    2 0.17 1.51 2.27  0.76 0.02
psych::describe(sim_4$kurts.rep,skew=FALSE)
##    vars   n mean   sd  min  max range   se
## X1    1 100 5.21 0.77 3.01 6.66  3.66 0.08
# RMSE for skewness # sqrt(sum((skew.rep- skew)^2)/rep)
Metrics::rmse(skew, sim_4$skew.rep)
 ## [1] 0.1740071
```

In Table 17, positively skewed and leptokurtic distributions are drawn from the positively skewed and leptokurtic distribution. It is seen that the mean of skewness values of the distributions produced in this example are quite close to the determined value, 2. The skewness values vary between 1.51 and 2.27, and kurtosis values are higher than 3. RMSE calculated for the skewness value was determined as 0.174. As a result, it was found that the function gives more consistent results at more common skewness values (between -1 + 1).

INSTALLING THE drawsample PACKAGE

The R package *drawsample* can be installed from CRAN with *install.packages("drawsample")* command. The package *drawsample* automatically provides the example data set "example_data". Additionally, package's files are available from the GitHub repository https://github.com/atalay-k/drawsample.

FINAL REMARKS

In this study, an R package *drawsample* has been developed to draw samples with desired properties from a given distribution. Contrary to simulation studies, the importance given to studies with real data has increased in recent years. It is thought that using the drawn samples obtained from the real data with *drawsample* package will provide an alternative to simulation studies as well as a complement for these studies. In addition, since the real data is used instead of the simulation studies, the descriptive characteristics of the study groups can be examined. Thus, it may be possible to examine the demographic characteristics of the individuals making up the sample.

In this study, four examples with real data are presented. It can be inferred from the examples in the study; the sample drawn from the real data is very close to the desired properties. However, it should be noted that it is not so easy to draw samples that perfectly match the desired properties in real data sets to draw sample from simulation data sets. Apart from the examples discussed in the study, two simulation data were genareted to evaluate the stability of the of *draw_sample()*. Then samples were drawn from these data sets under four cases. For each case in the the *draw_sample()*, 100 replications were performed and RMSE values are reported. As a limitation, *draw_sample()* yields more inconsistent

results at less common skew values. In addition, this inconsistency is directly related to the characteristics of the data from which the sample is taken. For example, the size of the population and its similarity (distribution shape) with the desired data are directly related to the amount of inconsistency. Due to the nature of random assignment, the function will get different samples in each time, even for the same values. Users are advised to run the function several times if they cannot obtain samples with the desired properties the first time.

Researchers can access the web-wide data sets provided by the "https://toolbox.google.com/datasetsearch" search engine, as well as they can access large public data such as TIMSS (Trends in International Mathematics and Science Study), PIRLS (The Progress in International Reading Literacy Study), and PISA (The Program for International Student Assessment). Various studies can be done by drawing samples using the data sets mentioned above based on distribution properties. In situations like this, a good approach would be to draw a sample of the population. As authors, we are open to all kinds of suggestions in the development of the *drawsample* package.

REFERENCES

Abdel-fattah, A.-F. A. (1994, April). *Comparing BILOG and LOGIST estimates for normal, truncated normal and beta ability distributions.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

American Educational Research Association. (2020). *Journal of Educational and Behavioral Statistics*. Retrieved from https://journals.sagepub.com/description/jeb

Bahry L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size* (Unpublished Master's disertation). University of Alberta, Edmonton, Canada.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması [Comparison of Polytomous Parametric and Nonparametric Item Response Theory Models]. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 354-372. DOI: 10.21031/epod.346650

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology,9*(2), 78–84. DOI: 10.1027/1614-2241/a000057

Blanca, M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option?. *Psicothema, 29*(4), 552-557. DOI: 10.7334/psicothema2016.383

Blest, D. C. (2003). A new measure of kurtosis adjusted for skewness. *Australian & New Zealand Journal of Statistics, 45*(2), 175-179.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Büyüköztürk, Ş., Çokluk, Ö., & Köklü, N. (2014). *Sosyal Bilimler için istatistik.* Ankara: Pegem Akademi.

Çelikten, S., & Çakan, M. (2019). Bayesian ve nonbayesian kestirim yöntemlerine dayali olarak siniflama indekslerinin TIMSS 2015 matematik testi üzerinde incelenmesi. [Investigation of classification indices on TIMSS 2015 mathematic-subtest through Bayesian and nonbayesian estimation methods]. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *13*(1), 105-124.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (Doctoral dissertation, Texas A&M University).

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19*(2), 133-149. DOI: 10.1207/s15324818ame1902_4

D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(4), 316-321. DOI: 10.2307/2684359

Doğan, N. & Tezbaşaran, A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. [Comparison of classical test theory and latent traits theory by samples]. *Hacettepe University Journal of Education, 25*, 58–67. DOI: 10.17860/efd.86348

Doğan, N., & Kılıç, A. F. (2018). The Effects of Sample Size, Correlation Technique, and Factor Extraction Method on Reliability Coefficients. *Kastamonu Eğitim Dergisi*, *26*(3), 697-706. DOI: 10.24106/kefdergi.413303

Dolma, S. (2009). *Çok ihtimalli rasch modeli ile derecelendirilmiş yanıt modelinin örtük özellikleri tahminleme performansı açısından simülasyon yöntemiyle karşılaştırılması* [*A simulation study for the comparison of*

the polytomous Rasch model and graded response model according to their performance on recovering the latent traits]. (Unpublished Doctoral dissertation). İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, Turkey.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. doi: 10.1037/0003-066X.63.7.591

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381. doi: 10.1177/0013164498058003001

Fialkowski, A. C. (2018). SimMultiCorrData: Simulation of Correlated Data with Multiple. Retrieved from: https://cran.r-project.org/web/packages/SimMultiCorrData/index.html

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In Hancock, G.R. & Mueller R. O. (Eds.), *Structural equation modeling: A second course*, (pp. 269-314). Information Age Publishing, U.S.A.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532. doi: 10.1007/BF02293811

Flott, L. W. (1995). Quality control: Measurement error. *Metal Finishing*, *93*(9), 72-75.

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*(3/4), 209-242. DOI: 10.1093/biomet/34.3-4.209

Gelbal, S. (1994). *P madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma* [*A comparison of item difficulty index P and Rasch model b parameters and their ability measures based on them*].Doctoral disertation, Hacettepe University, Ankara. Retrieved from

Gotzmann, A. J. (2011). *Comparison of vertical scaling methods in the context of NCLB.* (Doctoral dissertation, University of Alberta, Alberta). Retrieved from https://era.library.ualberta.ca/items/04a8d59c-791d-435b-bde6-7a6de3012169

Hallgren, K. A. (2013). Conducting simulation studies in the R programming environment. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 43–60. DOI:10.20982/tqmp.09.2.p043

Han, K. T. (2007). WinGen: Windows software thatgenerates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi: 10.1177/0146621607299271

Han, K. T., & Hambleton, R. K. (2007). *User's Manual: WinGen (Center for Educational Assessment Report No. 642*). Amherst, MA: University of Massachusetts, School of Education.

Headrick, T. C. 2002. Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics & Data Analysis. 40*(1),685–711. doi: 10.1016/S0167-9473(02)00072-5

Huck, S. W. (2012). *Reading statistics and research* (6th ed). Boston: Pearson.

John Wiley & Sons, Inc.-a. (2019). *Educational Measurement: Issues and Practice.* Retrieved from https://onlinelibrary.wiley.com/page/journal/17453992/homepage/productinformation.html

Kaya, Y., Leite, W. L., & Miller, M. D. (2015). A comparison of logistic regression models for DIF detection in polytomous items: The effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education*, *2*(1), 22-39. doi: 10.21449/ijate.239563

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399-419. DOI: 10.1177/0146621612446170

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52-54.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. doi: 10.1177/01466210122031975

Kline, R. B. (2005). *Principles and practice of structural equations modeling.* New York: Guilford.

Kogar, H . (2018). Effects of Various Simulation Conditions on Latent-Trait Estimates: A Simulation Study. International Journal of Assessment Tools in Education , 5 (2) , 263-273. DOI: 10.21449/ijate.377138

Kolen, M. J. (1985). Standard errors of Tucker Equating. *Applied Psychological Measurement, 9*(2), 209-223, doi: 10.1177/014662168500900209.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. DOI: 10.1037/0033-2909.105.1.156

Ministry of National Education. (2013). *Parasız Yatılılık Ve Bursluluk Sınavı (PYBS) Sınav Kılavuzu* [*Guide of Public Boarding And Scholarship Examination (PBSE)*]. Retrieved from http://www.meb.gov.tr/sinavlar/dokumanlar/2013/kilavuz/2013_PYBS_2.pdf

Ministry of National Education. (2020). *Ortaöğretim Kurumlarına İlişkin Merkezi Sınav Kılavuzu [Guide for Central Examination Secondary Education Institution]*. Retrieved from: http://www.meb.gov.tr/meb_iys_dosyalar/2020_07/17104126_2020_Ortaogretim_Kurumlarina_Iliskin_Merkezi_Sinav.pdf

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi [Examination of item and scale properties of likert type scale and metric scale to measure the same attitude according to classical test theory and item response theory]*. (Unpublished doctoral dissertation). Hacettepe University Social Sciences Institute, Ankara.

Olivier, J., & Norberg, M. M. (2010). Positively skewed data: revisiting the box-cox power transformation. *International Journal of Psychological Research*, *3*(1), 68-95. DOI: 10.21500/20112084.846

Pearson, E.S. (1932). The analysis of variance in cases of non-normal variation. *Biometrika. 23*, 114-133.

Pomplun, M., Omar, M. H., & Custer, M. (2004). A comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. *Educational and Psychological Measurement, 64*(4), 600-616. doi: 10.1177/0013164403261761

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved September 10, 2019, from http://www.R-project.org/

Ramos, C., Costa, P. A., Rudnicki, T., Marôco, A. L., Leal, I., Guimarães, R., ... & Tedeschi, R. G. (2018). The effectiveness of a group intervention to facilitate posttraumatic growth among women with breast cancer. *Psycho-oncology*, *27*(1), 258-264. DOI:10.1002/pon.4501

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12.

Reyhanlıoğlu Keceoğlu, Ç. (2018). *Parametrik ve Parametrik Olmayan Madde Tepki Kuramında Farklı Örneklem Büyüklüklerine ve Boyutluluklarına Göre Parametre Değişmezliğinin İncelenmesi.* Unpublished doctoral dissertation). Hacettepe University Social Sciences Institute, Ankara.

Şahin, M. G., & Yıldırım, Y. (2018). The examination of item difficulty distribution, test length and sample size in different ability distribution. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *9(3)*, 277-294. DOI: 10.21031/epod.385000

Sarkar, D. (2008). *lattice: Multivariate Data Visualization with R.* Springer-Verlag, New York.

SAS Institute Inc. (2009). *SAS/Stat User's Guide, version 9.2,* (Version 9.2). Cary, NC.

Sen, S., Cohen, A. S., & Kim, S. H. (2014, November). Robustness of mixture IRT models to violations of latent normality. In *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society* (Vol. 89, p. 27). Springer.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311. DOI: 10.1177/014662169001400307

Sireci, S. G. (1991). *"Sample-Independent" Item Parameters? An Investigation of the Stability of IRT Item Parameters Estimated from Small Data Sets.* Paper presented at the annual Conference of Northeastern Educational Research Association, New York, NY.

SSCP (2019). *2019 YKS Değerlendirme Raporu [2019 Examinations of the Council of Higher Education Assessment Report]*. Retrieved from https://dokuman.osym.gov.tr/pdfdokuman/2019/GENEL/yksDegRaporweb03092019.pdf

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16. doi: 10.1177/014662169201600101

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, *34*, 253-269. doi: 10.1177/001316447403400206

Uysal, İ. (2014). *Comparison of irt test equating methods for mixed format tests.* [*Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması*]. (Master disertation, Bolu Abant Izzet Baysal University, Bolu). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Variable Types. R package version 0.2.2.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (p. 56–75). Newbery Park, CA: Sage

Wickham, H., François, R., Henry, L., & Müller, K. (2016) *tibble: Simple data frames.* Retrieved from https://CRAN.Rproject.org/package=tibble. R package version 3.0.3

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of data manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr

Wicklin, R. (2013). *Simulating data with SAS*. SAS Institute.

Wuensch, K. L. (2005). *Kurtosis. Encyclopedia of Statistics in Behavioral Science*. doi:10.1002/0470013192.bsa334

Yıldırım, H., Uysal-Saraç, M., & Büyüköztürk, Ş. (2018). Farklı örneklem büyüklüğü ve dağılımı Koşullarında WLS ve Robust WLS yöntemlerinin karşılaştırılması. *Ilkogretim Online, 17*(1), 431-439. doi: 10.17051/ilkonline.2018.413794

Yıldırım, Y. (2015). *Derecelendirilmiş tepki modeli temelli parametre kestiriminde normallik sayıltısı ihlalinin ölçme kesinliğine etkisi* [*The effect of normality violation in the process of parameter estimation based upon Graded Response Model on measurement precision*]. (Master disertation, Gazi University, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/

Yoes, M. E. (1993*). A comparison of the effectiveness of item parameter estimation techniques used with the three-parameter logistic item response theory model. (Volumes I and II).* Unpublished Ph.D., University of Minnesota, Minneapolis/St. Paul, MN.

R'da "*drawsample"* Paketi ile Evrenden İstenilen Özelliklere Sahip Örneklem Çekme

Giriş

Eğitimde ve psikolojide ölçme ve değerlendirme alanında, puanların dağılımı grupların betimlenmesinde önemli bir role sahiptir. Grupların betimlenmesine ek olarak, normallik varsayımına dayanan birçok anlam çıkarıcı istatistiksel teknikleri kullanmak için normallik varsayımını test etmek çok önemlidir. Ancak Erceg-Hurn ve Mirosevich'in (2008) belirttiği gibi, gerçek veriler analiz edilirken normallik varsayımı nadiren karşılanmaktadır. Yapılan birçok araştırmada belirtildiği gibi, normal olmayan dağılımlar normal dağılıma göre daha yaygındır (Blanca, Arnau, López-Montiel, Bono ve Bendayan, 2013; Geary, 1947; Micceri, 1989; Olivier ve Norberg, 2010; Pearson, 1932).

Normallik varsayımının karşılanamaması, normalliğin ihlali ve dağılım türleri; test eşitleme, bilgisayarda bireyselleştirilmiş test uygulamaları, değişen madde fonksiyonu, sınıflandırma ve gizil puan kestirimleri gibi önemli konularda birçok araştırmacının odak noktası olmuştur (Custer, Omar, &Pomplun, 2006; Finney & DiStefano, 2006; Gotzmann, 2011; Kieftenbeld & Natesan, 2012; Kirisci, Hsu, & Yu, 2001; Kolen, 1985; Kogar, 2018; Seong, 1990; Uysal, 2014; Yıldırım, 2015). Normal dağılım ile normal olmayan dağılım türlerini elde etmede simülasyon ile veri üretmeye çok sık başvurulmaktadır (Abdel-fattah, 1994; Bıkmaz-Bilgen & Doğan, 2017; Dolma, 2009; Kaya, Leite, & Miller, 2015; Urry, 1974; Yıldırım, Uysal-Saraç, & Büyüköztürk, 2018; Yoes, 1993). Araştırmacılar farklı dağılım türlerinde veri üretilmesinde ise çeşitli yazılımlardan yararlanmıştır. Bahry (2012), WinGen 3.1 (Han, 2007) ile beta dağılımı ile çarpıklık katsayısı 0; 0,5 ve 1,00 olan örneklem büyüklüğü 100 ile 3000 arasında olan örneklemler üretmiştir. WinGen'e benzer şekilde SAS yazılımı (SAS Enstitüsü, 2009) ve R'da (R Core Team, 2014) farklı dağılım türleri elde etmede kullanılabilir. Örneğin Gotzmann (2011), normal dağılım gösteren dağılımı üretmede SAS'daki "Normal Distribution Function" dan ve çarpık dağılım gösteren dağılımı üretmede "RAND Beta Distribution Function" dan yararlanarak iki durum için 2000000 birey parametresi üretmiş ve bu verilerden yetenek parametresi ortalamaları araştırmanın amacına uygun olacak şekilde belirlenmiş ve farklı örneklem büyüklüklerinde (1500, 3000) rastgele veri setleri seçilmiştir. Veri üretmede, beta dağılımlarının kullanımı, çarpık puan dağılımlarını üretmeyi kolaylaştırmaktadır (Han ve Hambleton; 2007). Beta dağılımının bileşenleri α ve β parametreleridir. Bazı araştırmacılar ise simülasyon verisinden istendik özelliklere sahip örneklem çekmektedirler. Bu amaç doğrultusunda, orijinal veri setinden çarpık dağılıma sahip örneklem çekmede, Fleishman'ın (1978) güç yöntemi uygundur (Blanca, Alarcón, Arnau, Bono ve Bendayan, 2017; Stone, 1992; Kieftenbeld ve Natesan, 2012; Sen, Cohen ve Kim; 2014).

Simülasyon yöntemleri esnektir ve elde edilmesi mümkün olmayan problemlere nicel yanıtların sağlanması için uygulanabilir (Hallgren, 2013). Simülasyon güçlü bir teknik olmasına rağmen, sonuçları

genelleme, düzenleme ve gerçek verilere uygulama güçlüğü gibi bazı sınırlıkları vardır (Wicklin, 2013). Simülasyon verileri, gerçek verilerde ulaşılamayan mükemmel bir uyum sağlar. Hallgren'in (2013) belirttiği gibi, gerçek dünya veri setleri, simülasyon çalışmalarında oluşturulan ve genellikle mükemmel uyum olarak adlandırılabilecek idealist koşullar altında üretilen "temiz" veri setlerinden daha "kirli" olacaktır. Sireci (1991) gerçek test verileri kullanılmadığında, üretilen veriler pratikte karşılaşılan ilgili durumun özelliklerini doğru olarak yansıtıp yansıtmadığı bilinemeyeceğini ve geçerliği test edilemeyeceğini ifade etmiştir. Bu yüzden çalışmalarda gerçek veri kullanımı çalışmaların önemini artırmaktır. Ayrıca, Educational Measurement: Questions and Practice (EM: IP) ve Journal of Educational and Behavioral Statistics (JEBS) gibi bazı prestijli dergiler, uzun bir süre zarfında simülasyon çalışmalarını kabul etseler de, günümüzde simülasyon temelli çalışmaların "uygun olmayan makale konu örneklerinden" veya "düşük önceliğe sahip" olarak kabul edildiğini belirtmiştir (John Wiley & Sons Inc., 2019; American Educational Research Association, 2020).

Gerçek uygulamalarda veri toplama süreci zorluklarla doludur. Elde edilen örneklemler evren dağılımını temsil etmiyor, normal dağılmıyor veya istenen dağılıma uygun olmayan bir halde olabilir. Araştırma problemlerine dayalı olarak normalliği sağlamada, normal dağılımdan uzaklaştıran verilerin çıkarıldığı ya da ayrıştırıldığı bazı araştırmalara rastlanılmıştır. Gelbal (1994) yaptığı araştırmada, araştırmasının amacına uygun olarak, 2072 beşinci sınıf öğrencisine uygulanan Türkçe testi ve 2077 beşinci sınıf öğrencisine uygulanan Matematik testine ait verileri incelemiş test puanları normal dağılım göstermediği için Türkçe testinden 506 öğrenciyi, Matematik testinden 521 öğrenciyi çıkarmış ve normal dağılım gösteren iki yeni veri seti elde ederek test puanlarının hem normal dağıldığı hem de normal dağılmadığı durumlar üzerinde çalışmıştır. Doğan ve Tezbaşaran (2003) yaptıkları çalışmada amaçları doğrultusunda istenen dağılımı sağlayan verilerin örnekleme alınmasını sağlamışlardır. Örneklemlerin seçiminde random ve kasıtlı örnekleme tekniklerinin kullanıldığını ifade etmişlerdir. Araştırmacılar; amaçları doğrultusunda Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavı (ÖSYS) 2001 giren bireylerin oluşturduğu evrenden örneklem büyüklükleri 2353 ile 29244 arasında değişen random, sağa çarpık, sola çarpık, basık ve normal dağılım gösteren beş örneklem seçmiştir. Çarpık örneklemlerde, yapılacak karşılaştırmaların isabetliliğini artırmak için çarpıklık mutlak değerleri (1,00) ve basıklık değerleri (1,37) eşit tutulmuş ve basık dağılımda çarpıklık katsayısının 0 olması sağlanmıştır. Doğan ve Tezbaşaran'ın (2003) çalışmasına benzer şekilde, Şahin ve Yıldırım (2018), başlangıçta sağa çarpık bir veri seti olan Seviye Belirleme Sınavı (SBS) verilerinden (çarpıklık katsayısı = 1,05) hem sağa çarpık hem de sola çarpık yetenek dağılımları elde etmiştir. Sola çarpık veri setleri için, amaçlanan örneklemeyle istenen örneklem dağılımı sağlanmış ve araştırma kapsamında ele alınan tüm örneklem büyüklüklerinde örneklemler için çarpıklık katsayısı ≈-1,00 olan gruplar seçilmiştir. Yukarıdakilere ek olarak, literatürde birçok araştırmacı, yaptıkları çalışmaların amacına uygun olacak şekilde gerçek veri setinden (evren) örneklem almayı seçmiştir (Courville, 2004; Doğan ve Kılıç, 2018; Fan, 1998; Nartgün, 2002; Reyhanlıoğlu Keçeoğlu, 2018). Evrenden örnekleme sürecinde, gelecekteki çalışmalar için örneklem seçimini kolaylaştıran ve istenen özelliklere yaklaştıran bir araca sahip olmanın önemli olacağı düşünülmektedir. Öyle ki normal dağılımdan farklı yetenek dağılımlarına sahip örneklemler üzerinde çalışılması literatürde bazı araştırmaların sonucunda önerilmiştir (Çelikten ve Çakan, 2019). Araştırmalar incelendiğinde, araştırmacıların geniş bir veri setinden istenen özelliklere sahip örneklem seçimini sağlayacak bir araca ihtiyaç olduğu sonucuna varılmıştır.

Bu çalışmanın amacı gerçek bir veri setinden istendik özelliklere sahip örneklemlerin seçilmesinde yararlanılacak *drawsample* adlı bir R paketinin geliştirilmesidir. Bu amaç doğrultusunda seçilecek örneklemin sahip olması beklenen özelliklerden normallikten sapma ölçüleri (çarpıklık ve basıklık) ve örneklem büyüklüğü gibi koşulların bir veya birden fazlasının belirlenmesi ile kasıtlı örnekleme yapılabilmektedir. Çalışmanın amacı için Fleishman'ın (1978) güç yönteminden yararlanılmıştır.

Geliştirilen *drawsample* paketinde yer alan *draw_sample()* fonksiyonu ile evrenden ya da büyük bir örneklemden, istendik özelliklere sahip örneklem çekilmesinin olabildiğince kullanılışlı olması beklenmekte ve gerçek veriler ile yapılan araştırmaların önem kazandığı bu dönemde simülasyon çalışmalarına alternatif oluşturarak, dağılıma ilişkin gerçek verilere dayalı farklı konularda yapılacak olan çalışmalara katkı sağlayacağı düşünülmektedir.

R'da draw_sample() Fonksiyonu

İstenilen özelliklerde örneklem seçmek amacıyla 6 argümanlı bir *draw_sample()* fonksiyonu yazılmıştır. Fonksiyonun argümanlarına Tablo 1'de yer verilmiştir.

Tablo 1: *draw_sample()* Fonksiyonun Argümanları

Argümanlar	
dist	Very seti: ID ve puanları içeren iki sütunlu bir veri seti
n	Sayısal: Örneklem büyüklüğü
skew	Sayısal: Çarpıklık değeri
kurt	Sayısıal: Basıklık değeri
replacement	Mantıksal: Yeniden örnekleme yapılsın mı? (varsayılan FALSE'dur)
output_name	Karakter: İki bileşenli çıktı dosyasının adı

Bu fonksiyonda, "dist" argümanı örneklemlerin çekileceği veri setidir. Veri seti ilk sütununu öğrenci kimlik numaraları (ID) ve ikinci sütununu öğrencinin toplam test puanını veya yetenek puanı (theta) içerecek şekilde iki sütundan oluşmalıdır. Fonksiyonun argümanlarından istenen örneklem büyüklüğü olan "n" örneklemlerin çekileceği veri setinin uzunluğundan büyükse şu hatayı verir: "Cannot take a sample larger than the length of the data". Örneğin, içe aktarılan verilerin örneklem büyüklüğü 1000 olmasına rağmen ve kullanıcılar örneklem büyüklüğü 2000 olan bir örneklem almak isterlerse fonksiyon hata verir ve çalışmayı durdurur.

Tablo 1'deki argümanlardan "skew " ve "kurts" belirlenirken, Fleishman Güç Yöntemi Ağırlıkları tablosuna başvurulmalıdır. Örneğin çarpıklık değeri 1 ve basıklık değeri 0 gibi bazı kombinasyonlara karşılık gelen Fleishman katsayıları yoktur. Eğer kişinin verdiği çarpıklık değeri tabloda yoksa veya çarpıklık değeri olsa bile o değere karşılık gelen basıklık değeri tabloda bulunmuyorsa fonksiyon "No valid power method constants could be found for the specified values. Change the values" hatası vererek çalışmayı durdurur.

Fonksiyon ile kullanıcının girdiği çarpıklık ve basıklık değerlerine sahip olan bir referans dağılım oluşturulmakta ve daha sonra evrenden referans dağılım baz alınarak bir örnekleme yapılmaktadır. Bu kısımda kullanıcının örneklemi çekmek istediği veri setinde yer alan dağılımın minumum ve maksimum değerlerine göre yeniden ölçeklenmektedir. Veri setinde oluşturulan her bir kategoride bulunan birey sayısı, oluşturulan referans dağılımının her bir kategorisinde bulunundan daha fazla olduğunda fonksiyon, *replacement* argümanının varsayılan değeri olarak yeniden örnekleme yapılmadan yürütülebilir. Ancak *draw_sample()* veri setindeki en az bir kategoride bulunan birey sayısı, referans dağılımının ilgili kategorisinde bulunan birey sayısından az ise "Cannot take a sample form that data without replacement. Please change replacement=TRUE" hatasını verir. Bu durumda, fonksiyon *replacement* argümanının değeri değiştirilerek (FALSE) kullanılabilir.

draw_sample() fonksiyonu"output" olarak üç farklı liste içermektedir. Bunlardan biri "desc" olarak adlandırılan bir çıktıdır. Bu çıktıda "population (tüm veri, içe aktarılan veriler)", "reference (referans alınan dağılım)" ve "sample (çekilen örneklem)" dağılımlarının ortalamasını, standart sapmasını, çarpıklığını ve basıklığını içerir. Diğeri "graph" olarak adlandırılırmaktadır ve sol tarafta "popülasyon (içe aktarılan veriler)" ve sağ tarafta "sample" (çıkarılan veriler) olmak üzere iki histogram grafiği içermektedir. Bir diğeri ise "sample" olarak adlandırılan ve ilk sütununu seçilen örneklemdeki bireylerin kimlik numaraları (ID) ve diğer sütununu bireylerin puanını içerecen iki sütunlu bir veri setidir.

Bu çalışmada geliştirilen fonksiyonun kullanımını göstermek amacıyla 2013 yılı 6. Sınıf Parasız Yatılılık ve Bursluluk Sınavı'na (PYBS) ait Fen ve Teknoloji ile Sosyal Bilimler alt testleri verilerine dayalı olarak örnek dört uygulama sunulmuştur. Yapılan bu örnek uygulamalarda 5000 kişilik veri setlerinden 500 kişilik örneklemler çekilmiştir. İlk iki örnek uygulamada normal dağılımdan daha sola çarpık bir dağılım şekli gösteren Fen ve Teknoloji alt testi (Score_1) verilerinden sırasıyla normal

dağılıma sahip ve hem sola çarpık hem de sivri bir dağılıma sahip örneklemler çekilmiştir. Üçüncü örnek uygulamada ise normal dağılımdan daha basık bir dağılım şekli gösteren Sosyal Bilimler alt testi (Score_2) verisinden sağa çarpık ve sivri dağılıma sahip bir örneklem çekilmesi için komutlar verilmiştir. Ancak fonksiyon burada çalışmamış ve yeniden örneklemeye olanak sağlacak şekilde terar komut verilerek istenen özellikte bir örneklem elde edilmiştir. Elde edilen örneklemlerin dağılım türü açısından istenilen özelliklere sahip olduğu görülmüşür. Fonksiyonun tutarlılığının incelenmesi amacıyla *rbeta()* fonksiyonu ile üretilen iki veri setinden çekilen örneklemlere ilişkin yapılan replikasyonlar sonucu çarpıklık değerlerine ilişkin ortalama RMSE değerleri raporlanmıştır. Fonksiyonun daha sık rastlanan çarpıklık değerlerinde (-1 ve +1 arasında) daha tutarlı sonuçlar verdiği bulgusuna ulaşılmıştır.

Sonuç ve Tartışma

Bu çalışmada, belirli bir dağılımdan istenen özelliklere sahip örneklemler elde etmek için R'da *draw_sample()* adlı bir fonksiyon ve örnek veri seti sunan *drawsample* paketi geliştirilmiştir. Simülasyon çalışmalarının aksine gerçek verilerle yapılan çalışmalara verilen önem son yıllarda artmıştır. *draw_sample()* fonksiyonu ile gerçek verilerden çekilen örneklerin kullanılmasının simülasyon çalışmalarına alternatif oluşturacağı gibi bu çalışmaları tamamlayacağı düşünülmektedir. Ayrıca örneklemler gerçek kişilerden oluşacağı için çalışma gruplarının betimleyici özellikleri incelenebilir. Böylece örneklemi oluşturan bireylerin demografik özelliklerini incelemek mümkün olabilir.

Bu çalışmada hem gerçek veriden, hem de simülasyon verisine dayalı olarak örnek uygulamalar sunulmuştur. Çalışmadaki örneklerden anlaşılacağı üzere simülasyon verilerinden alınan örneklem istenilen özelliklere çok yakındır. Bununla birlikte, gerçek veri setinden istenen özelliklere mükemmel bir şekilde uyan örneklemler seçmenin, özellikle yeniden örnekleme yapılmadığında, kolay olmadığı unutulmamalıdır. Sunulan bu örnek uygulamaların dışında fonksiyonun tutarlılığının değerlendirilmesi amacıyla üretilen simülasyon verisinden tekrarlı çekilen örneklemlere dayalı olarak *draw_sample()* fonksiyonunun, daha az rastlanan büyük çarpıklık değerlerinde daha tutarsız sonuçlar verdiği görülmüştür. Ayrıca bu tutarsızlık, örneklemin çekildiği evrenin özellikleriyle de doğrudan ilişkilidir. Örneğin evren ile çekilecek örneklemin büyüklükleri ve istenen özelliklere sahip örneklemin evren ile dağılım türlerinin benzerliği tutarsızlık miktarı ile doğrudan ilişkilidir. Rastgele atamanın doğası gereği, fonksiyondaki argümanların aynı değerleri için bile her seferinde farklı örneklem çekebilir. Kullanıcılar istenilen özelliklere sahip örneklemi ilk seferde çekemezlerse, kullanıcılara fonksiyonu birkaç kez çalıştırmaları önerilir.

Araştırmacılar, "https://toolbox.google.com/datasetsearch" arama motoru tarafından sağlanan web genelindeki veri kümeleri ile TIMSS (Trends in International Mathematics and Science Study), PIRLS (The Progress in International Reading Literacy Study) ve PISA (The Program for International Student Assessment) gibi açık büyük verilere erişebilirler. Dağılım özelliklerine göre yukarıda belirtilen veri setlerinden örneklemler alınarak çeşitli çalışmalar yapılabilir. Bu gibi durumlarda, *drawsample* paketi ile evrenden örneklem çekmek iyi bir yaklaşım olacaktır. Yazarlar olarak *drawsample* paketinin geliştirilmesinde her türlü öneriye açık olduğumuzu bildiririz.