ANALYSIS OF WORLD UNIVERSITY RANKING DATASET

CS544 Foundations of Analytics

Archana Balachandran

August 2016

Data Preparation

1. **Dataset Download**
   The dataset World University Ranking was downloaded from Kaggle. The data selected is based on Shanghai Ranking, which is a powerful ranking method considering the important university measures for rank determination. The dataset was downloaded and viewed in Excel.

   From the dataset containing a total of 4897 entries for rankings from 2005-2015, the top 100 universities in the year 2012 were selected.

   Some of the relevant columns in the dataset is described below[1]:

   - world_rank - world rank for university. Contains rank ranges and equal ranks (eg. 101-152).
   - university_name - name of university.
   - national_rank - rank of university within its country.
   - score - total score, used to determine rank.
   - alumni - Alumni Score, based on the number of alumni of an institution winning nobel prizes and fields medals.
   - award - Award Score, based on the number of staff of an institution winning Nobel Prizes in Physics, Chemistry, Medicine, and Economics and Fields Medals in Mathematics.
   - year - year of ranking (2012).

2. **Conversion to CSV format**
   File -> Save As -> Select CSV in Type

3. **Dataset Import**
   RStudio -> Tools -> Import Dataset -> From Local file
   To view the imported dataset in R.
   View(UniRanking)

world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	patents	score	year
1	Harvard University	USA	1	7	9	1	1	1	1	5	100.00	2012
2	Massachusetts Institute of Technology	USA	2	9	17	3	12	4	4	1	91.67	2012
3	Stanford University	USA	3	17	11	5	4	2	2	15	89.50	2012
4	University of Cambridge	United Kingdom	1	10	24	4	16	16	11	50	86.17	2012
5	California Institute of Technology	USA	4	2	29	7	37	22	22	18	85.21	2012
6	Princeton University	USA	5	8	14	2	53	33	26	101	82.50	2012
7	University of Oxford	United Kingdom	2	13	28	9	15	13	19	26	82.34	2012
8	Yale University	USA	6	14	31	12	14	6	15	66	79.14	2012
9	Columbia University	USA	7	23	21	10	13	12	14	5	78.86	2012
10	University of California, Berkeley	USA	8	16	52	6	6	5	3	16	78.55	2012
11	University of Chicago	USA	9	15	26	8	34	20	28	101	73.82	2012
12	Cornell University	USA	10	21	42	14	22	21	16	10	73.69	2012
13	University of Pennsylvania	USA	11	31	16	24	9	10	8	9	73.64	2012
14	University of Tokyo	Japan	1	32	19	31	8	19	23	3	69.49	2012
15	Johns Hopkins University	USA	12	34	77	20	11	9	9	7	66.94	2012

[1] Source: https://www.kaggle.com/mylesoneill/world-university-rankings

Analysis of Categorical Data

All columns in the dataset:

world_rank
institution
country
national_rank
quality_of_education
alumni_employment
quality_of_faculty
publications
influence
citations
patents
score
Year

The column 'country' can be selected for categorical analysis, where the number of colleges in each country(which is the frequency) can be identified.

```
country.data<-UniRanking$country
table(UniRanking$country)
```

```
> table(UniRanking$country) #majority in the US

    Australia        Canada       Denmark       Finland        France       Germany
            2             3             1             1             5             3
       Israel         Italy         Japan   Netherlands        Norway   South Korea
            4             1             5             2             1             1
       Sweden   Switzerland United Kingdom           USA
            1             4             8            58
>
```

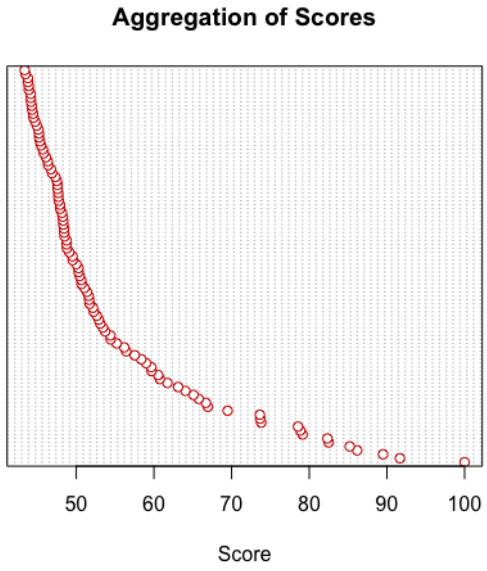
The countries having atleast 1 top university are USA, United Kingdon, Switzerland, Sweden, South Korea, Norway, Netherlands, Japan, Italy, Israel, Germany, France, Finland, Denmark, Canada and Australia.

As one can observe, USA has the maximum number of top colleges, followed by United Kingdom. Japan and France both share the next position with 5 top universities in each country.

Analysis of Numerical Data

The score for each university can be used as a measure for numerical data analysis.

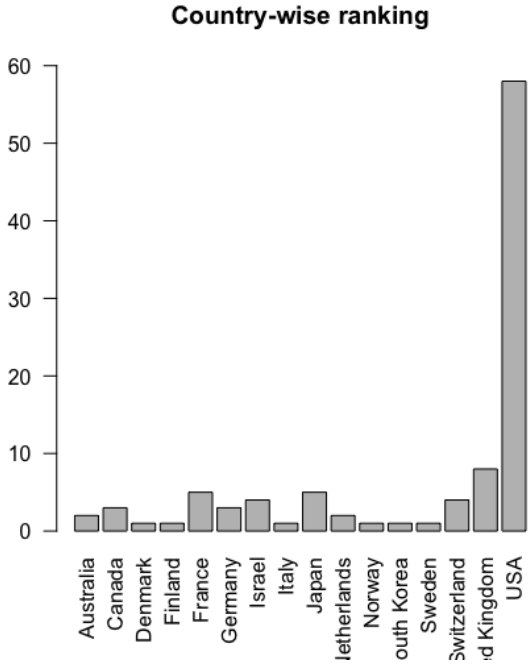
dotchart(UniRanking$score, col="red3", xlab="Score", main="Aggregation of Scores")



It can be observed that the scores primarily cluster around the rand 40 – 50, and is more spread out after 68. It may indicate that very few universities fall within the top 30%.

As observed earlier, a barplot of the countries rankings would indicate that USA has the largest number of top-ranked universities in the world.

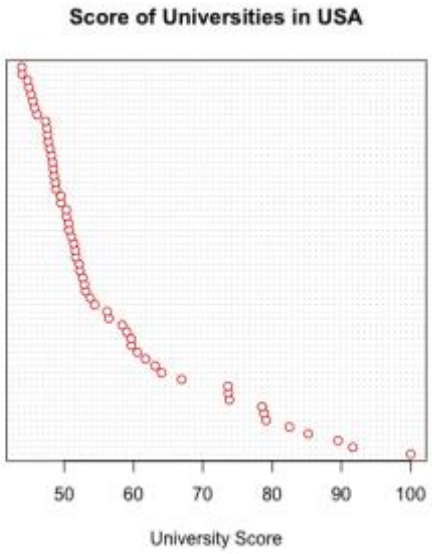
barplot(table(UniRanking$country),ylim = c(0,60), las=2,main="Country-wise ranking")



We will now take a good look at USA, since it has the majority of top-ranked universities –We will examine score in all top-ranked US universities.

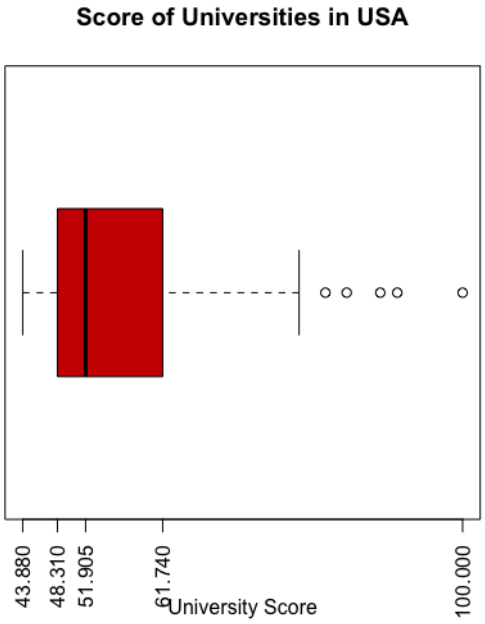
usa.score <- subset(UniRanking, country=="USA", select=(c("score")))

```
> dotchart(usa.score$score, col="red3", xlab="University Score", main="Score of Universities in USA")
```



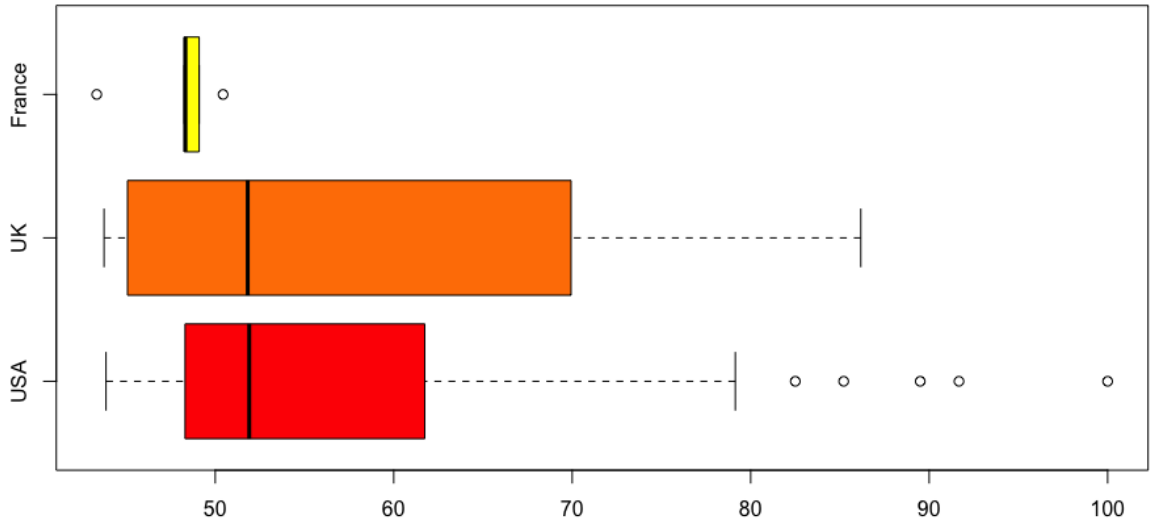
A boxplot can be considered more accurate in displaying the center, spread and skew of scores data. We can also note the outliers in the scores data.

```
> boxplot(usa.score$score, xaxt="n", horizontal=TRUE, col="red3",
+           xlab="University Score", main="Score of Universities in USA")
> axis(side=1, at=fivenum(usa.score$score), labels=TRUE, las=2)
```



Comparing Scores in Top three countries (countries having highest number of top-ranked universities):
USA(58 top universities) UK(8 top universities) and France(5 top universities) are taken for this comparative study.

```
> usa.score <- subset(UniRanking, country=="USA", select=c("score"))
> uk.score <- subset(UniRanking, country=="United Kingdom", select=c("score"))
> france.score <-subset(UniRanking,country=="France",select=c("score"))
> boxplot(usa.score$score, uk.score$score, france.score$score, names=c("USA", "UK","France"), col=heat.colors(3), horizontal=TRUE)
```

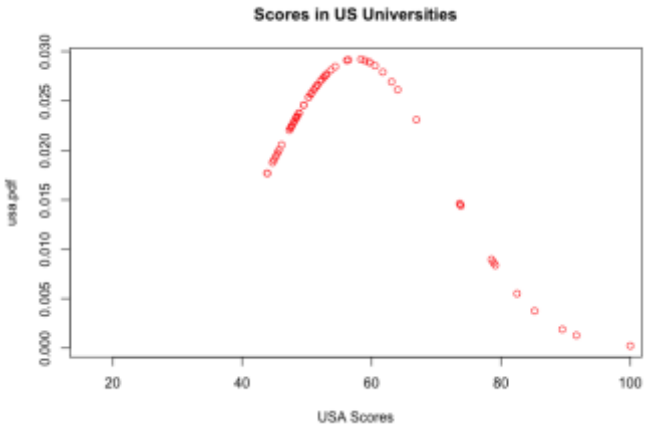


UK and USA both have a very similar median (around 52), which indicates that 50% of the data is greater than this value. No outliers are observed in the case of UK, whereas, outliers are observed for USA and France. 5 outliers are observed in USA, which indicates the top 5 colleges in USA with 5 highest scores.

In the case of France which has 5 records, University of Helsinki(having a score of 44) is the outlier in the lower section of data, while University of Paris-Sud is another outlier in the upper end, having a score of 50.44. The minimum, median and lower quartile appears to be the same.

Distribution of Scores

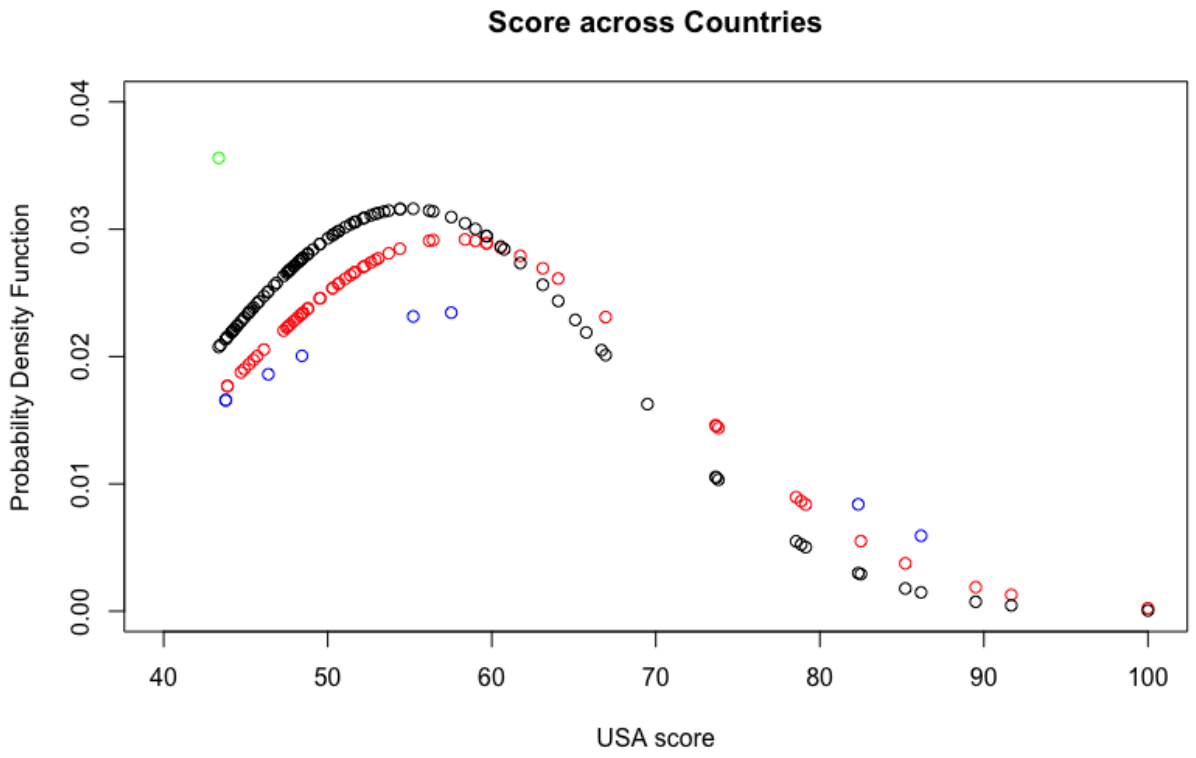
```
> usa.score<- subset(UniRanking, country=="USA", select=c("score"))
> usa.mean <- mean(usa.score$score)
> usa.sd <- sd(usa.score$score)
> usa.pdf <- dnorm(usa.score$score, mean=usa.mean, sd=usa.sd)
> plot(usa.score$score, usa.pdf, type="p", col="red", xlab="USA Scores", main="Scores in US Universities", xlim=c(usa.mean-3*usa.sd, usa.mean+3*usa.sd))
>
```



Comparing Scores across countries

The impact of standard deviation on the shape of the curve can be observed in the distributions shown below(ignore outliers). As the standard deviation decreases, shape of curve becomes teeper.

```
> all.mean <- mean(UniRanking$score)
> all.sd <- sd(UniRanking$score)
> all.pdf <- dnorm(UniRanking$score, mean=all.mean, sd=all.sd)
> all.sd
[1] 12.61927
> usa.score <- subset(UniRanking, country=="USA", select=c("score"))
> uk.score <- subset(UniRanking, country=="United Kingdom", select=c("score"))
> france.score <- subset(UniRanking, country=="France", select=c("score"))
>
> usa.mean <- mean(usa.score$score)
> usa.sd <- sd(usa.score$score)
> usa.pdf <- dnorm(usa.score$score, mean=usa.mean, sd=usa.sd)
> usa.sd
[1] 13.64813
> uk.mean <- mean(uk.score$score)
> uk.sd <- sd(uk.score$score)
> uk.pdf <- dnorm(uk.score$score, mean=uk.mean, sd=uk.sd)
> uk.sd
[1] 17.00839
> france.mean <- mean(france.score$score)
> france.sd <- sd(france.score$score)
> france.pdf <- dnorm(france.score$score, mean=france.mean, sd=france.sd)
> france.sd
[1] 2.684105
>
> plot(usa.score$score, usa.pdf, type="p", col="red", xlim=c(40,100), ylim=c(0.0, 0.04), xlab="USA score", ylab="Probability Density Function", main="Score across Countries")
> lines(uk.score$score, uk.pdf, type="p", col="blue")
> lines(france.score$score, france.pdf, type="p", col="green")
> lines(UniRanking$score, all.pdf, type="p", col="black")
>
```



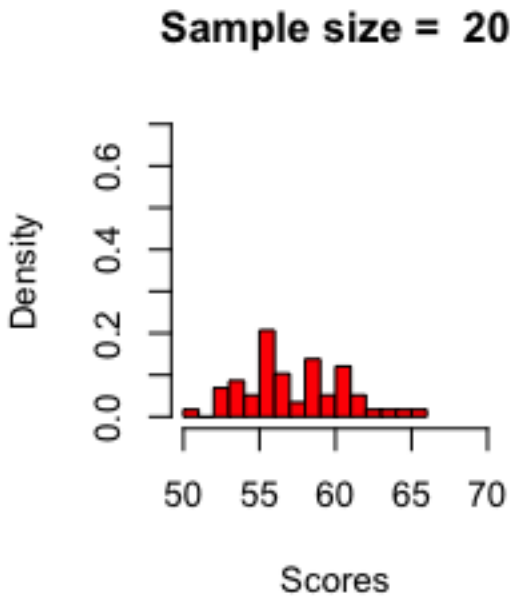
Standard Deviation in decreasing order: UK(blue)>US(red)>All(black)>France (green)
It is observed that, as the SD decreases, the shape of the curve becomes steeper.

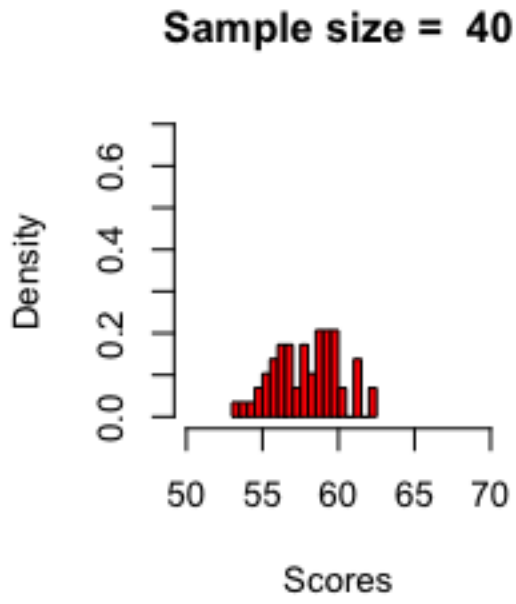
Central Limit Theorem: Application

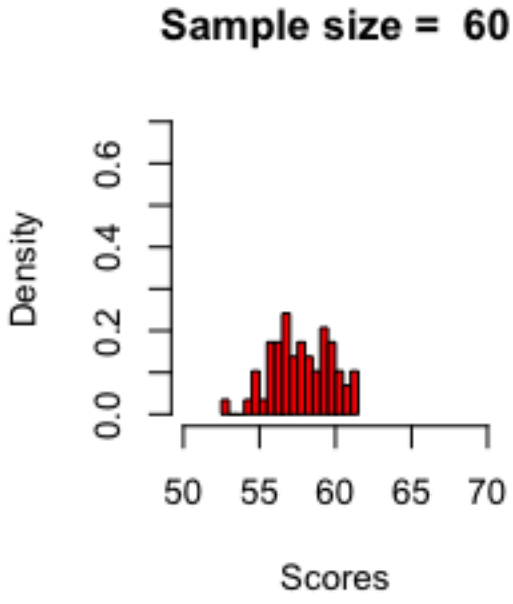
Sampling with Normal Distribution Data

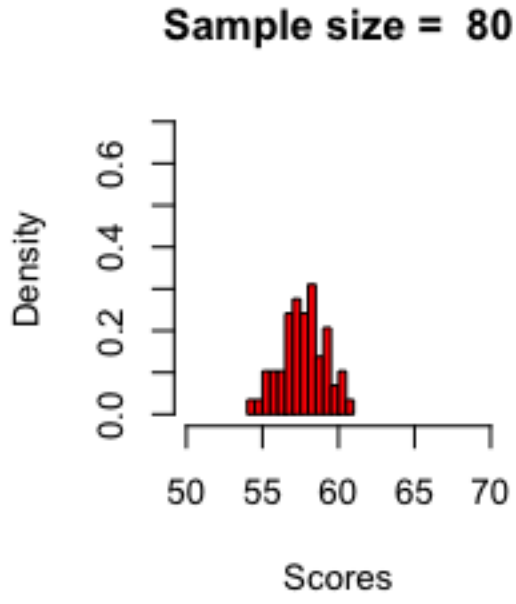
From the plots below, it can be observed that, as the sample size increases, the standard deviation of the scores of US universities decreases while the mean remains a constant.

```
> par(mfrow = c(2,2))
> set.seed(150)
> samples <- length(usa.score$score)
> sample.data <- numeric(samples)
> sample.size = 20
> for (i in 1:samples) {
+   sample.data[i] <- mean(rnorm(sample.size, mean=usa.mean, sd=usa.sd))
+ }
> hist(sample.data, prob = TRUE, breaks=15, xlim=c(50,70), ylim=c(0.0,0.7),  xlab="Scores", main=paste("Sample size = ", sample.size)
, col="red")
> mean(sample.data)
[1] 57.20023
> sd(sample.data)
[1] 3.281101
>
>
> sample.size = 40
> for (i in 1:samples) {
+   sample.data[i] <- mean(rnorm(sample.size, mean=usa.mean, sd=usa.sd))
+ }
> hist(sample.data, prob = TRUE, breaks=15, xlim=c(50,70), ylim=c(0.0,0.7),  xlab="Scores", main=paste("Sample size = ", sample.size)
, col="red")
> mean(sample.data)
[1] 57.89757
> sd(sample.data)
[1] 2.162876
>
> sample.size = 60
> for (i in 1:samples) {
+   sample.data[i] <- mean(rnorm(sample.size, mean=usa.mean, sd=usa.sd))
+ }
> hist(sample.data, prob = TRUE, breaks=15, xlim=c(50,70), ylim=c(0.0,0.7),  xlab="Scores", main=paste("Sample size = ", sample.size)
, col="red")
> mean(sample.data)
[1] 57.75036
> sd(sample.data)
[1] 1.99973
>
> sample.size = 80
> for (i in 1:samples) {
+   sample.data[i] <- mean(rnorm(sample.size, mean=usa.mean, sd=usa.sd))
+ }
> hist(sample.data, prob = TRUE, breaks=15, xlim=c(50,70), ylim=c(0.0,0.7),  xlab="Scores", main=paste("Sample size = ", sample.size)
, col="red")
> mean(sample.data)
[1] 57.7055
> sd(sample.data)
[1] 1.461513
>
```


Sample size = 20
Density
0.0
0.2
0.4
0.6
50
55
60
65
70
Scores


Sample size = 40
Density
0.0
0.2
0.4
0.6
50
55
60
65
70
Scores


Sample size = 60
Density
0.0
0.2
0.4
0.6
50
55
60
65
70
Scores


Sample size = 80
Density
0.0
0.2
0.4
0.6
50
55
60
65
70
Scores

Sampling Methods

The following sampling methods are explored in the sections below:

1. Simple Random Sampling with Replacement
2. Simple Random Sampling without Replacement
3. Systematic Sampling

1. Simple Random Sampling With Replacement

All items in the data frame has the same probability for selection. For the UniRanking dataset, a sample of size 40 was selected using Simple Random Sampling with Replacement.

```
> #Simple Random Sampling with Replacement
> #All items within the frame have the same probability for selection
>
> set.seed(153)
> s <- srswr(40, nrow(UniRanking))
> s[s != 0]
 [1] 3 1 2 1 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1
> rows <- (1:nrow(UniRanking))[s!=0]
> rows <- rep(rows, s[s != 0])
> rows
 [1]  1  1  1  5 14 14 22 23 27 28 32 34 34 35 40 43 46 48 48 50 52 53 53 54 56 60 60 61 68 72 73 74 76 77 79 84 84 91 96 98
> sample.1 <- UniRanking[rows, ]
> #Display the dataframe containing the list of all institutions, along with their corresponding frequency after sampling.
> setNames(data.frame(table(sample.1$institution)), c( "Institution", "Freq"))
                                    Institution Freq
1                \xcc\xe4cole normale suprieure - Paris    1
2                          \xcc\xe4cole Polytechnique    1
3                            Arizona State University    0
4                                  Boston University    0
5                                  Brown University    0
6                      California Institute of Technology    1
7                            Carnegie Mellon University    1
8                       Case Western Reserve University    0
9                                 Columbia University    0
10                                Cornell University    0
11                               Dartmouth College    0
12                                 Duke University    1
13                                Emory University    0
14                     Georgia Institute of Technology    0
15                                Harvard University    3
16                      Hebrew University of Jerusalem    1
17                             Imperial College London    1
18                            Johns Hopkins University    0
19                               Karolinska Institute    0
20                                  Kyoto University    0
21                                 Leiden University    0
22               Ludwig Maximilian University of Munich    0
23                Massachusetts Institute of Technology    0
24                                McGill University    0
25                                  Mines ParisTech    0
26                                Nagoya University    1
27                              New York University    1
28                           Northwestern University    0
29                      Ohio State University, Columbus    1
```

	Institution	Freq
1	\xcc\xe4cole normale suprieure - Paris	1
2	\xcc\xe4cole Polytechnique	1
3	Arizona State University	0
4	Boston University	0
5	Brown University	0
6	California Institute of Technology	1
7	Carnegie Mellon University	1
8	Case Western Reserve University	0

9	Columbia University	0
10	Cornell University	0
11	Dartmouth College	0
12	Duke University	1
13	Emory University	0
14	Georgia Institute of Technology	0
15	Harvard University	3
16	Hebrew University of Jerusalem	1
17	Imperial College London	1
18	Johns Hopkins University	0
19	Karolinska Institute	0
20	Kyoto University	0
21	Leiden University	0
22	Ludwig Maximilian University of Munich	0
23	Massachusetts Institute of Technology	0
24	McGill University	0
25	Mines ParisTech	0
26	Nagoya University	1
27	New York University	1
28	Northwestern University	0
29	Ohio State University, Columbus	1
30	Osaka University	1
31	Pennsylvania State University, University Park	0
32	Pierre-and-Marie-Curie University	0
33	Princeton University	0
34	Purdue University, West Lafayette	0
35	Rice University	0
36	Rockefeller University	0
37	Ruprecht Karl University of Heidelberg	0
38	Rutgers University-New Brunswick	1
39	Sapienza University of Rome	1
40	Seoul National University	0
41	Stanford University	0
42	Swiss Federal Institute of Technology in Lausanne	0
43	Swiss Federal Institute of Technology in Zurich	0
44	Technical University of Munich	0
45	Technion \x89\xdb\xd2 Israel Institute of Technology	0
46	Tel Aviv University	1
47	Texas A&M University, College Station	1
48	Tohoku University	1
49	Tufts University	0
50	University College London	0
51	University of Arizona	1
52	University of Bristol	1
53	University of British Columbia	0
54	University of California, Berkeley	0
55	University of California, Davis	0
56	University of California, Irvine	0
57	University of California, Los Angeles	0
58	University of California, San Diego	0
59	University of California, San Francisco	0
60	University of California, Santa Barbara	1
61	University of Cambridge	0

62	University of Chicago	0
63	University of Colorado Boulder	1
64	University of Copenhagen	0
65	University of Edinburgh	2
66	University of Florida	0
67	University of Geneva	0
68	University of Helsinki	0
69	University of Illinois at Urbana\x89\xdb\xd2Champaign	0
70	University of Manchester	1
71	University of Maryland, College Park	1
72	University of Michigan, Ann Arbor	2
73	University of Minnesota, Twin Cities	0
74	University of North Carolina at Chapel Hill	0
75	University of Nottingham	0
76	University of Oslo	1
77	University of Oxford	0
78	University of Paris-Sud	2
79	University of Pennsylvania	0
80	University of Pittsburgh - Pittsburgh Campus	0
81	University of Queensland	0
82	University of Rochester	2
83	University of Southern California	0
84	University of Sydney	0
85	University of Texas at Austin	0
86	University of Texas MD Anderson Cancer Center	1
87	University of Texas Southwestern Medical Center	0
88	University of Tokyo	2
89	University of Toronto	1
90	University of Utah	0
91	University of Virginia	2
92	University of Washington - Seattle	0
93	University of Wisconsin\x89\xdb\xd2Madison	0
94	University of Zurich	0
95	Utrecht University	0
96	Vanderbilt University	0
97	Washington University in St. Louis	0
98	Weizmann Institute of Science	0
99	Williams College	0
100	Yale University	0

2. Simple Random sampling without Replacement

A simple random sample of size 40 was drawn from the UniRanking dataset (which has 100 rows) without replacement.

```
> set.seed(153)
>
> s <- srswor(40, nrow(UniRanking))
>
> sample.2 <- UniRanking[s != 0, ]
> head(sample.2[c("institution","world_rank")])
                        institution world_rank
1                 Harvard University          1
4             University of Cambridge          4
6                Princeton University          6
8                    Yale University          8
9                 Columbia University          9
10 University of California, Berkeley         10
> setNames(data.frame(table(sample.2$institution)), c( "Institution", "Freq"))
                                          Institution Freq
1                 \xcc\xe4cole normale suprieure - Paris    0
2                             \xcc\xe4cole Polytechnique    1
3                               Arizona State University    1
4                                     Boston University    1
5                                      Brown University    1
6                         California Institute of Technology    0
7                               Carnegie Mellon University    1
8                          Case Western Reserve University    0
9                                    Columbia University    1
10                                    Cornell University    1
11                                   Dartmouth College    0
12                                     Duke University    1
13                                    Emory University    1
14                         Georgia Institute of Technology    1
15                                    Harvard University    1
16                          Hebrew University of Jerusalem    0
```

	Institution	Freq
1	\xcc\xe4cole normale suprieure - Paris	0
2	\xcc\xe4cole Polytechnique	1
3	Arizona State University	1
4	Boston University	1
5	Brown University	1
6	California Institute of Technology	0
7	Carnegie Mellon University	1
8	Case Western Reserve University	0
9	Columbia University	1
10	Cornell University	1
11	Dartmouth College	0
12	Duke University	1
13	Emory University	1
14	Georgia Institute of Technology	1
15	Harvard University	1
16	Hebrew University of Jerusalem	0
17	Imperial College London	0
18	Johns Hopkins University	0
19	Karolinska Institute	0
20	Kyoto University	0
21	Leiden University	1
22	Ludwig Maximilian University of Munich	0
23	Massachusetts Institute of Technology	0

24	McGill University	0
25	Mines ParisTech	1
26	Nagoya University	0
27	New York University	0
28	Northwestern University	0
29	Ohio State University, Columbus	0
30	Osaka University	0
31	Pennsylvania State University, University Park	1
32	Pierre-and-Marie-Curie University	0
33	Princeton University	1
34	Purdue University, West Lafayette	1
35	Rice University	0
36	Rockefeller University	0
37	Ruprecht Karl University of Heidelberg	0
38	Rutgers University-New Brunswick	1
39	Sapienza University of Rome	0
40	Seoul National University	0
41	Stanford University	0
42	Swiss Federal Institute of Technology in Lausanne	0
43	Swiss Federal Institute of Technology in Zurich	0
44	Technical University of Munich	0
45	Technion \x89\xdb\xd2 Israel Institute of Technology	1
46	Tel Aviv University	0
47	Texas A&M University, College Station	0
48	Tohoku University	1
49	Tufts University	0
50	University College London	1
51	University of Arizona	1
52	University of Bristol	0
53	University of British Columbia	0
54	University of California, Berkeley	1
55	University of California, Davis	0
56	University of California, Irvine	1
57	University of California, Los Angeles	1
58	University of California, San Diego	0
59	University of California, San Francisco	0
60	University of California, Santa Barbara	0
61	University of Cambridge	1
62	University of Chicago	1
63	University of Colorado Boulder	0
64	University of Copenhagen	0
65	University of Edinburgh	0
66	University of Florida	0
67	University of Geneva	0
68	University of Helsinki	0
69	University of Illinois at Urbana\x89\xdb\xd2Champaign	0
70	University of Manchester	0
71	University of Maryland, College Park	0
72	University of Michigan, Ann Arbor	1
73	University of Minnesota, Twin Cities	0
74	University of North Carolina at Chapel Hill	0
75	University of Nottingham	0
76	University of Oslo	0

77	University of Oxford	0
78	University of Paris-Sud	0
79	University of Pennsylvania	0
80	University of Pittsburgh - Pittsburgh Campus	1
81	University of Queensland	0
82	University of Rochester	0
83	University of Southern California	1
84	University of Sydney	1
85	University of Texas at Austin	1
86	University of Texas MD Anderson Cancer Center	1
87	University of Texas Southwestern Medical Center	0
88	University of Tokyo	1
89	University of Toronto	0
90	University of Utah	0
91	University of Virginia	1
92	University of Washington - Seattle	0
93	University of Wisconsin\x89\xdb\xd2Madison	1
94	University of Zurich	0
95	Utrecht University	1
96	Vanderbilt University	0
97	Washington University in St. Louis	1
98	Weizmann Institute of Science	1
99	Williams College	1
100	Yale University	1

3. Systematic Sampling

For a sample size n, N items from the frame are partitioned into n groups. Each group has k items (K=N/n). First item for the sample is randomly selected from the first set of k items in the frame. After the first selection, remaining n-1 items are selected by taking every kth item from the frame. [2]

```
> #Systematic Sampling
> #Step 1 calculation of number of items in each group for frame size N and sample size n
> N <- nrow(UniRanking)
> n <- 40
> k <- ceiling(N / n)
> k
[1] 3
>
>
> #Step 2 : selecting an item at random from the first group of k items. n samples are then drawn from every kth item in subsequent k
-item groups
> r <- sample(k, 1)
> r
[1] 2
> s <- seq(r, by = k, length = n)
> sample.3 <- UniRanking[s, ]
> head(sample.3[c("institution","world_rank")])
                           institution world_rank
2  Massachusetts Institute of Technology          2
5     California Institute of Technology          5
8                     Yale University          8
11               University of Chicago         11
14                 University of Tokyo         14
17                    Kyoto University         17
> setNames(data.frame(table(sample.3$institution)), c( "Institution", "Freq"))
                                      Institution Freq
1                \xcc\xe4cole normale suprieure - Paris    0
2                           \xcc\xe4cole Polytechnique    0
3                             Arizona State University    1
4                                    Boston University    1
5                                     Brown University    0
6                 California Institute of Technology    1
7                           Carnegie Mellon University    0
8                    Case Western Reserve University    0
9                                 Columbia University    0
10                                 Cornell University    0
11                                  Dartmouth College    0
12                                    Duke University    0
13                                   Emory University    0
14                    Georgia Institute of Technology    0
15                                 Harvard University    0
16                  Hebrew University of Jerusalem    0
17                         Imperial College London    0
```

Confidence Level – 80%

We will now show the confidence intervals of the mean of numeric variable for various samples and compare against the population mean. The variable selected for this section is scores, and the number of samples drawn is 50(for both ci=80% and 90%).

[2] source: https://learn.bu.edu/bbcswebdav/pid-4255946-dt-content-rid-14691730_1/courses/16sum1metcs544sc1/CS544_Module5.pdf

```
set.seed(150)
conf.value <- 80
alpha.value <- 1 - conf.value / 100

#calculating z score of the upper tail

zscore <- qnorm(1- alpha.value / 2)
OUT <- 0
for (i in 1:draw.samples) {
  rows.to.sample <- srswr(sample.size, data.frame.size)
  uni.sample.data <- UniRanking$score[rows.to.sample != 0]
  mean.of.50.samples[i] <- mean(uni.sample.data, na.rm=TRUE)
  # calcualtion of confidence intervals (ci)
  ci.lower[i] <- mean.of.50.samples[i] - zscore * sample.means.sd
  ci.upper[i] <- mean.of.50.samples[i] + zscore * sample.means.sd
  # displaying the 80% confidence level for each of the sample means

  print.samples <- sprintf("%2d. Sample Mean=%.2f, CI= %.2f - %.2f, %s",
                           i, mean.of.50.samples[i], ci.lower[i], ci.upper[i],
                           ifelse(total.mean >= ci.lower[i] && total.mean <= ci.upper[i], "IN", "OUT"))
  if (total.mean < ci.lower[i] || total.mean > ci.upper[i]) inc(OUT) <-  1
  cat(print.samples,"\n")
```

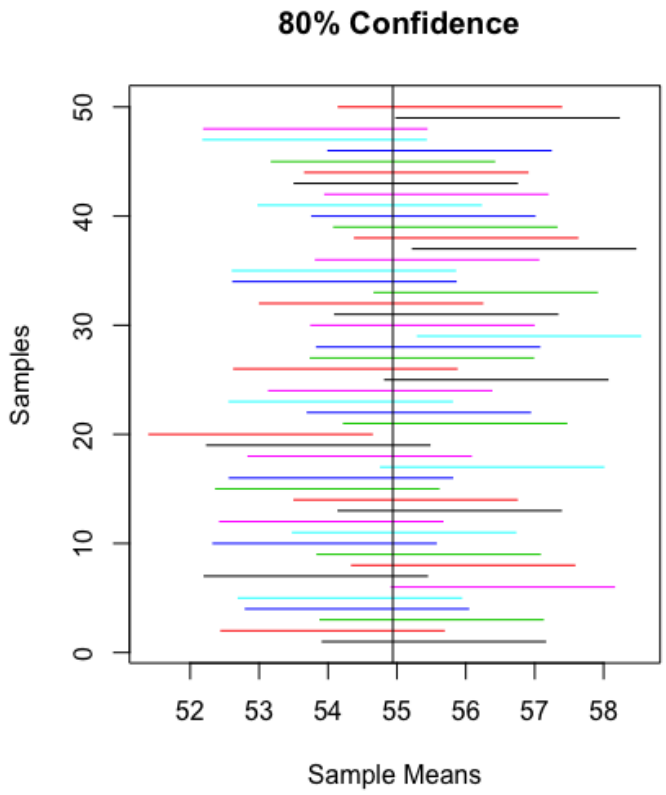
Mean of population:

```
> total.mean
[1] 54.94
>
```

1. Sample Mean=55.54, CI= 53.92 - 57.15, IN
2. Sample Mean=54.07, CI= 52.45 - 55.68, IN
3. Sample Mean=55.50, CI= 53.89 - 57.12, IN
4. Sample Mean=54.42, CI= 52.80 - 56.04, IN
5. Sample Mean=54.32, CI= 52.70 - 55.94, IN
6. Sample Mean=56.54, CI= 54.92 - 58.15, IN
7. Sample Mean=53.82, CI= 52.21 - 55.44, IN
8. Sample Mean=55.96, CI= 54.34 - 57.58, IN
9. Sample Mean=55.46, CI= 53.84 - 57.08, IN
10. Sample Mean=53.95, CI= 52.33 - 55.57, IN
11. Sample Mean=55.10, CI= 53.49 - 56.72, IN
12. Sample Mean=54.05, CI= 52.43 - 55.66, IN
13. Sample Mean=55.77, CI= 54.15 - 57.38, IN
14. Sample Mean=55.13, CI= 53.51 - 56.74, IN
15. Sample Mean=53.99, CI= 52.37 - 55.61, IN
16. Sample Mean=54.19, CI= 52.57 - 55.81, IN
17. Sample Mean=56.38, CI= 54.77 - 58.00, IN
18. Sample Mean=54.46, CI= 52.84 - 56.08, IN
19. Sample Mean=53.86, CI= 52.24 - 55.48, IN
20. Sample Mean=53.02, CI= 51.40 - 54.64, OUT
21. Sample Mean=55.84, CI= 54.23 - 57.46, IN
22. Sample Mean=55.32, CI= 53.70 - 56.94, IN
23. Sample Mean=54.18, CI= 52.57 - 55.80, IN
24. Sample Mean=54.76, CI= 53.14 - 56.37, IN
25. Sample Mean=56.44, CI= 54.83 - 58.06, IN
26. Sample Mean=54.25, CI= 52.63 - 55.87, IN
27. Sample Mean=55.36, CI= 53.75 - 56.98, IN

28. Sample Mean=55.45, CI= 53.84 - 57.07, IN
29. Sample Mean=56.92, CI= 55.30 - 58.53, OUT
30. Sample Mean=55.37, CI= 53.75 - 56.99, IN
31. Sample Mean=55.72, CI= 54.10 - 57.33, IN
32. Sample Mean=54.63, CI= 53.01 - 56.24, IN
33. Sample Mean=56.29, CI= 54.67 - 57.91, IN
34. Sample Mean=54.24, CI= 52.62 - 55.86, IN
35. Sample Mean=54.23, CI= 52.61 - 55.85, IN
36. Sample Mean=55.44, CI= 53.82 - 57.06, IN
37. Sample Mean=56.85, CI= 55.23 - 58.46, OUT
38. Sample Mean=56.01, CI= 54.39 - 57.62, IN
39. Sample Mean=55.70, CI= 54.09 - 57.32, IN
40. Sample Mean=55.38, CI= 53.77 - 57.00, IN
41. Sample Mean=54.61, CI= 52.99 - 56.22, IN
42. Sample Mean=55.57, CI= 53.95 - 57.19, IN
43. Sample Mean=55.13, CI= 53.51 - 56.75, IN
44. Sample Mean=55.28, CI= 53.66 - 56.90, IN
45. Sample Mean=54.80, CI= 53.18 - 56.41, IN
46. Sample Mean=55.62, CI= 54.00 - 57.24, IN
47. Sample Mean=53.80, CI= 52.19 - 55.42, IN
48. Sample Mean=53.82, CI= 52.20 - 55.43, IN
49. Sample Mean=56.61, CI= 54.99 - 58.22, OUT
50. Sample Mean=55.77, CI= 54.15 - 57.39, IN

```
> sprintf("Samples OUT the confidence interval = %d", OUT)
[1] "Samples OUT the confidence interval = 4"
>
> #each sample plotted against the population mean which is denoted by the vertical line
>
> matplot(rbind(ci.lower, ci.upper),
+         rbind(1:draw.samples, 1:draw.samples), type="l", lty=1,
+         ylab="Samples", xlab="Sample Means", main=" 80% Confidence Interval")
> abline(v = total.mean, lty="solid")
>
```



Confidence Level – 90%

```
> set.seed(150)
>
>    #storing values for confidence level and calculating its alpha value
>
> conf.value <- 90
> alpha.value <- 1 - conf.value / 100
>
>    #calculating z score of the upper tail
>
> zscore <- qnorm(1- alpha.value / 2)
>
> OUT <- 0
> for (i in 1:draw.samples) {
+   # using simple random sampling without replacement to select the rows to sample
+   rows.to.sample <- srswr(sample.size, data.frame.size)
+   uni.sample.data <- UniRanking$score[rows.to.sample != 0]
+   mean.of.50.samples[i] <- mean(uni.sample.data, na.rm=TRUE)
+   # calcualtion of confidence intervals
+   ci.lower[i] <- mean.of.50.samples[i] - zscore * sample.means.sd
+   ci.upper[i] <- mean.of.50.samples[i] + zscore * sample.means.sd
+
+   # displaying the 90% confidence level for each of the sample means
+   print.samples <- sprintf("%2d. Sample Mean=%.2f, CI= %.2f - %.2f, %s",
+                            i, mean.of.50.samples[i], ci.lower[i], ci.upper[i],
+                            ifelse(total.mean >= ci.lower[i] && total.mean <= ci.upper[i], "IN", "OUT"))
+   if (total.mean < ci.lower[i] || total.mean > ci.upper[i]) inc(OUT) <-  1
+   cat(print.samples,"\n")
```

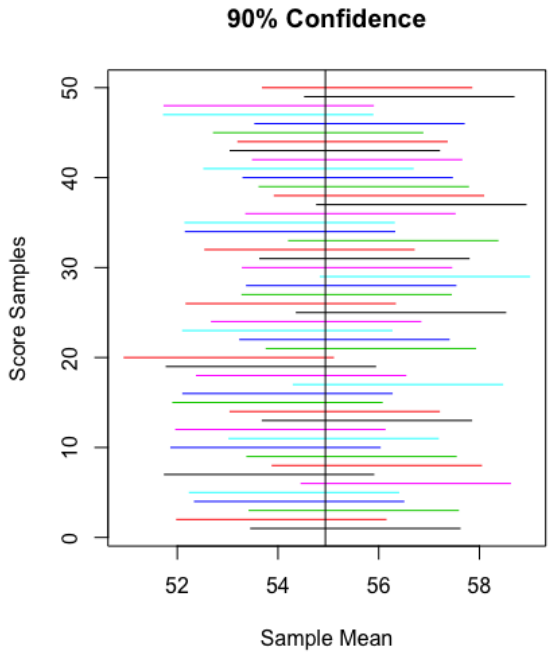
Mean of population:

```
> total.mean
[1] 54.94
>
```

1. Sample Mean=55.54, CI= 53.46 - 57.61, IN
2. Sample Mean=54.07, CI= 51.99 - 56.14, IN
3. Sample Mean=55.50, CI= 53.43 - 57.58, IN
4. Sample Mean=54.42, CI= 52.34 - 56.50, IN
5. Sample Mean=54.32, CI= 52.24 - 56.39, IN
6. Sample Mean=56.54, CI= 54.46 - 58.61, IN
7. Sample Mean=53.82, CI= 51.75 - 55.90, IN
8. Sample Mean=55.96, CI= 53.89 - 58.04, IN
9. Sample Mean=55.46, CI= 53.38 - 57.54, IN
10. Sample Mean=53.95, CI= 51.87 - 56.03, IN
11. Sample Mean=55.10, CI= 53.03 - 57.18, IN
12. Sample Mean=54.05, CI= 51.97 - 56.12, IN
13. Sample Mean=55.77, CI= 53.69 - 57.84, IN
14. Sample Mean=55.13, CI= 53.05 - 57.20, IN
15. Sample Mean=53.99, CI= 51.91 - 56.07, IN
16. Sample Mean=54.19, CI= 52.11 - 56.26, IN
17. Sample Mean=56.38, CI= 54.31 - 58.46, IN
18. Sample Mean=54.46, CI= 52.38 - 56.54, IN
19. Sample Mean=53.86, CI= 51.78 - 55.93, IN
20. Sample Mean=53.02, CI= 50.95 - 55.10, IN
21. Sample Mean=55.84, CI= 53.77 - 57.92, IN
22. Sample Mean=55.32, CI= 53.24 - 57.39, IN
23. Sample Mean=54.18, CI= 52.11 - 56.26, IN
24. Sample Mean=54.76, CI= 52.68 - 56.83, IN
25. Sample Mean=56.44, CI= 54.37 - 58.52, IN
26. Sample Mean=54.25, CI= 52.18 - 56.33, IN
27. Sample Mean=55.36, CI= 53.29 - 57.44, IN
28. Sample Mean=55.45, CI= 53.38 - 57.53, IN
29. Sample Mean=56.92, CI= 54.84 - 58.99, IN
30. Sample Mean=55.37, CI= 53.30 - 57.45, IN
31. Sample Mean=55.72, CI= 53.64 - 57.79, IN
32. Sample Mean=54.63, CI= 52.55 - 56.70, IN
33. Sample Mean=56.29, CI= 54.21 - 58.36, IN
34. Sample Mean=54.24, CI= 52.16 - 56.31, IN
35. Sample Mean=54.23, CI= 52.16 - 56.31, IN
36. Sample Mean=55.44, CI= 53.36 - 57.51, IN
37. Sample Mean=56.85, CI= 54.77 - 58.92, IN
38. Sample Mean=56.01, CI= 53.93 - 58.08, IN
39. Sample Mean=55.70, CI= 53.63 - 57.78, IN
40. Sample Mean=55.38, CI= 53.31 - 57.46, IN
41. Sample Mean=54.61, CI= 52.53 - 56.68, IN
42. Sample Mean=55.57, CI= 53.50 - 57.65, IN
43. Sample Mean=55.13, CI= 53.05 - 57.20, IN
44. Sample Mean=55.28, CI= 53.21 - 57.36, IN
45. Sample Mean=54.80, CI= 52.72 - 56.87, IN

46. Sample Mean=55.62, CI= 53.54 - 57.69, IN
47. Sample Mean=53.80, CI= 51.73 - 55.88, IN
48. Sample Mean=53.82, CI= 51.74 - 55.89, IN
49. Sample Mean=56.61, CI= 54.53 - 58.68, IN
50. Sample Mean=55.77, CI= 53.69 - 57.84, IN

```
> sprintf("Total OUT ci = %d", OUT)
[1] "Total OUT ci = 0"
> #each sample plotted against the population mean which is denoted by the vertical line
>
> matplot(rbind(ci.lower, ci.upper),
+         rbind(1:draw.samples, 1:draw.samples), type="l", lty=1,
+         ylab="Score Samples", xlab="Sample Mean", main="90% Confidence Interval")
> abline(v = total.mean, lty="solid")
> total.mean
[1] 54.94
>
```



Observations:

- When confidence level is 80%, the sample confidence intervals that contain the population mean is 96%. 4 samples have confidence intervals which do not contain the population mean.

- When confidence level is 90%, the sample confidence intervals that contain the population mean is 100%. The confidence intervals of all samples contain the population mean.

Reference Links:

a. https://www.kaggle.com/mylesoneill/world-university-rankings

b. https://learn.bu.edu/bbcswebdav/pid-4255941-dt-content-rid-14691726_1/courses/16sum1metcs544sc1/CS544_Module3.pdf

c. https://learn.bu.edu/bbcswebdav/pid-4255944-dt-content-rid-14691728_1/courses/16sum1metcs544sc1/CS544_Module4.pdf

d. https://learn.bu.edu/bbcswebdav/pid-4255946-dt-content-rid-14691730_1/courses/16sum1metcs544sc1/CS544_Module5.pdf 3.

e. https://learn.bu.edu/bbcswebdav/pid-4255953-dt-content-rid-14691711_1/courses/16sum1metcs544sc1/CS544_Module6.pdf

f. http://www.stat.osu.edu/~calder/stat528/Lectures/lecture21_2slides.PDF