CS544 Final Project
Data Analysis

Yixiong Zhang

Dataset selection

- Select the dataset of Students Performance in Exams from https://www.kaggle.com/datasets
- Dataset variables include:
  - gender
  - race/ethnicity
  - parental level of education
  - lunch
  - test preparation course
  - math score
  - reading score
  - writing score

```
Observations: 1,000
Variables: 8
$ gender                     <fct> female, female, female, male, male, female, female, male, male, f…
$ race.ethnicity             <fct> group B, group C, group B, group A, group C, group B, group B, gr…
$ parental.level.of.education <fct> bachelor's degree, some college, master's degree, associate's deg…
$ lunch                      <fct> standard, standard, standard, free/reduced, standard, standard, s…
$ test.preparation.course    <fct> none, completed, none, none, none, none, completed, none, complet…
$ math.score                 <int> 72, 69, 90, 47, 76, 71, 88, 40, 64, 38, 58, 40, 65, 78, 50, 69, 8…
$ reading.score              <int> 72, 90, 95, 57, 78, 83, 95, 43, 64, 60, 54, 52, 81, 72, 53, 75, 8…
$ writing.score              <int> 74, 88, 93, 44, 75, 78, 92, 39, 67, 50, 52, 43, 73, 70, 58, 78, 8…
```

Data preparing and preprocessing

- Import the dataset into R

- See if there are missing values in the dataset

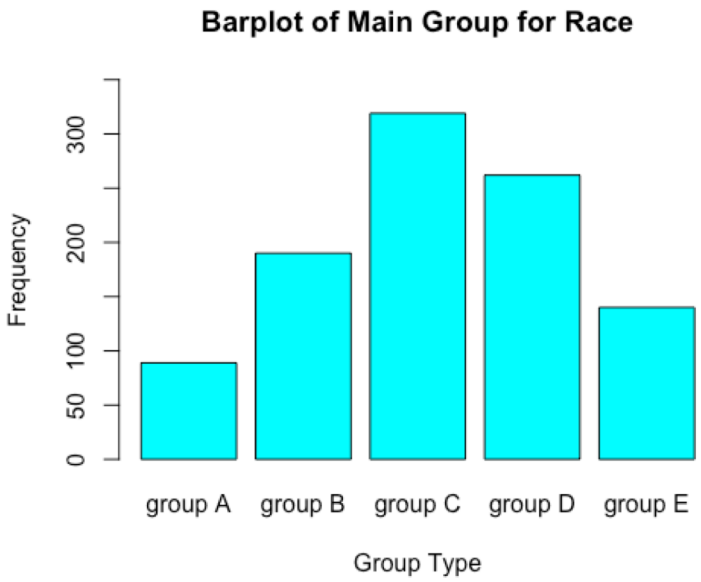
```
> # Preparing the data
> dataset <- read.csv("~/Downloads/StudentsPerformance.csv")
> dataset
    gender race.ethnicity parental.level.of.education       lunch test.preparation.course math.score
1   female        group B           bachelor's degree    standard                    none         72
2   female        group C                some college    standard               completed         69
3   female        group B             master's degree    standard                    none         90
4     male        group A          associate's degree free/reduced                    none         47
5     male        group C                some college    standard                    none         76
6   female        group B          associate's degree    standard                    none         71
7   female        group B                some college    standard               completed         88
8     male        group B                some college free/reduced                    none         40
9     male        group D                 high school free/reduced               completed         64
10  female        group B                 high school free/reduced                    none         38
11    male        group C          associate's degree    standard                    none         58
12    male        group D          associate's degree    standard                    none         40
13  female        group B                 high school    standard                    none         65
14    male        group A                some college    standard               completed         78
15  female        group A             master's degree    standard                    none         50
16  female        group C           some high school    standard                    none         69
17    male        group C                 high school    standard                    none         88
18  female        group B           some high school free/reduced                    none         18
19    male        group C             master's degree free/reduced               completed         46
20  female        group C          associate's degree free/reduced                    none         54
21    male        group D                 high school    standard                    none         66
22  female        group B                some college free/reduced               completed         65
23    male        group D                some college    standard                    none         44
24  female        group C           some high school    standard                    none         69
25    male        group D           bachelor's degree free/reduced               completed         74
26    male        group A             master's degree free/reduced                    none         73
27    male        group B                some college    standard                    none         69
> # See for missing values in the dataset
> is.na(dataset)
      gender race.ethnicity parental.level.of.education lunch test.preparation.course math.score
 [1,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [2,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [3,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [4,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [5,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [6,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [7,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [8,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
 [9,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[10,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[11,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[12,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[13,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[14,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[15,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[16,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[17,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[18,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[19,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[20,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[21,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[22,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[23,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[24,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[25,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[26,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[27,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[28,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[29,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[30,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[31,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[32,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[33,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[34,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[35,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
[36,]  FALSE         FALSE                        FALSE FALSE                    FALSE      FALSE
```

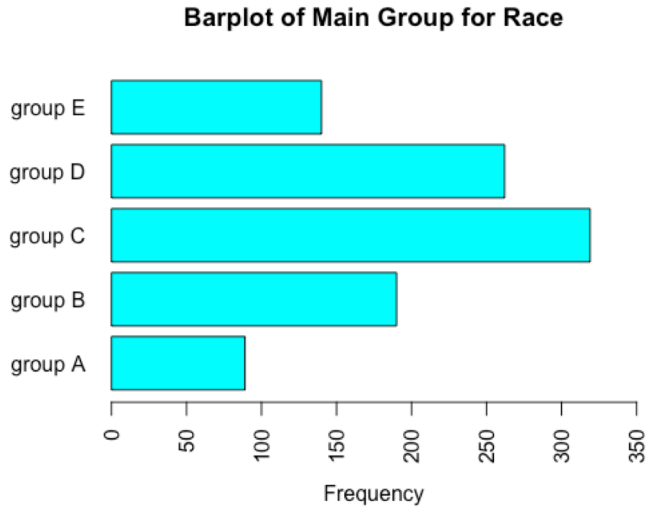
Analysis of categorical variable

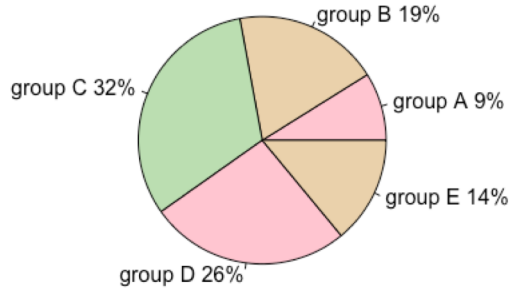
- Choose the categorical variable of race. ethnicity to analyze
- The table below shows the frequency of this categorical variable

Race. ethnicity	Frequency
Group A	89
Group B	190
Group C	319
Group D	262
Group E	140

Graphical representations of categorical data





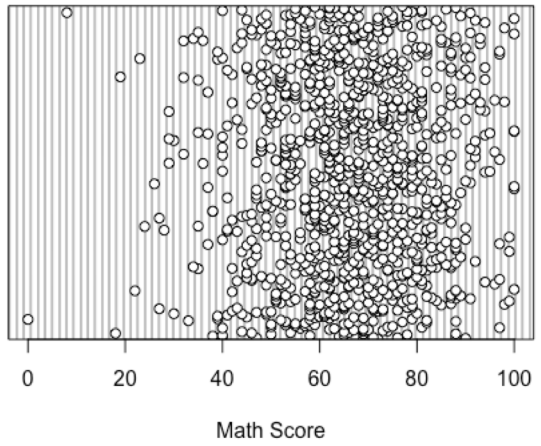


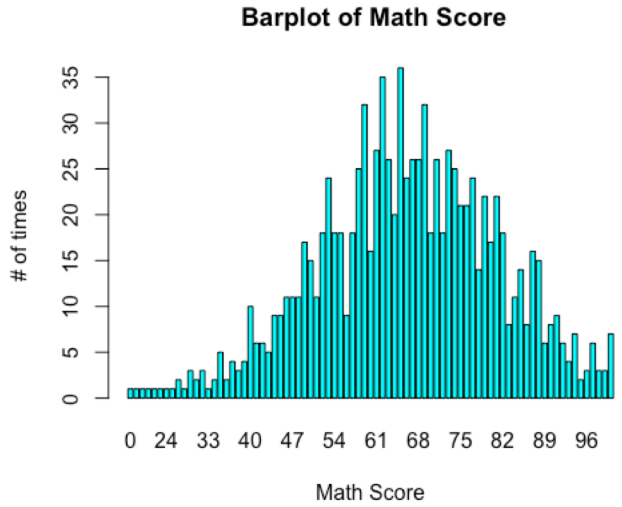
From those different kinds of graph representation, the group C is the top main group for race and ethnicity

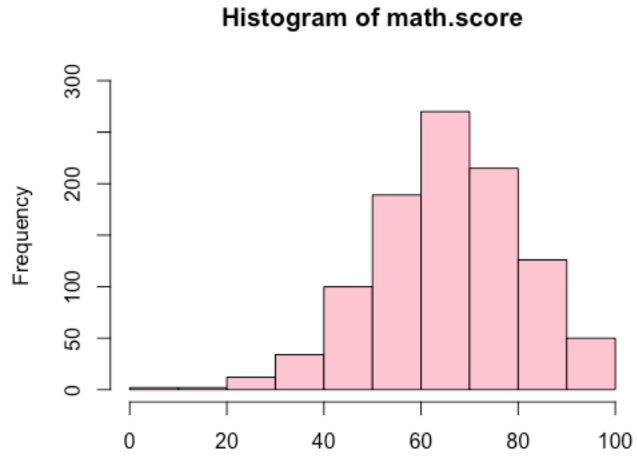
Analysis of numerical variable

- Choose the student's math score as a numerical variable to analyze
- The mean of math score variable is 66.09
- The median of math score variable is 66
- The variance and the standard deviation of math score are 229.9 and 15.16 respectively.

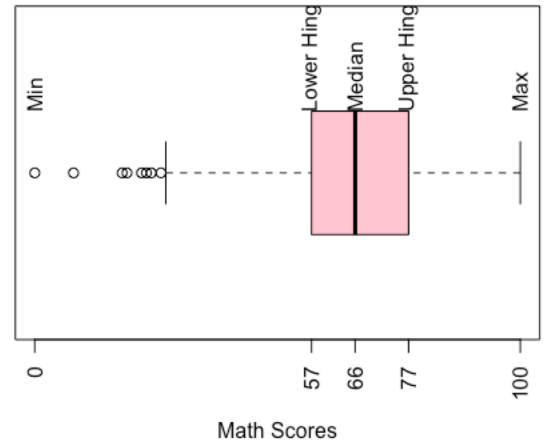
Graphical Representation of numerical variable







- As these graphs shown, student's maximum math score is 100 and minimum math score is 0
- In addition, most of student's math score are around from 60 to 70



Analysis of one set of two variables

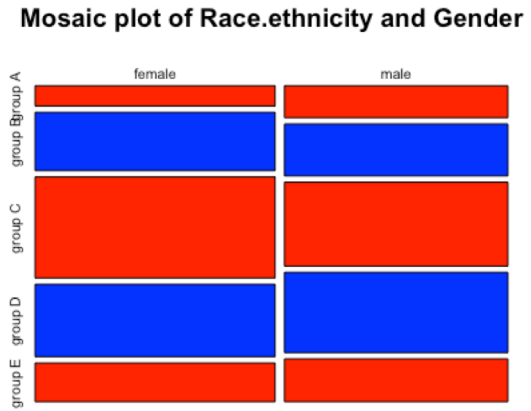
- The two variables of gender and race are shown as two-way table by adding marginal distribution

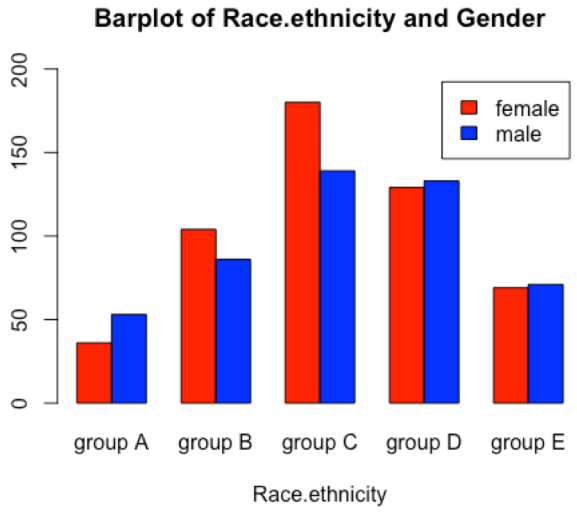
- The proportions for the gender and race dataset along the columns are shown below

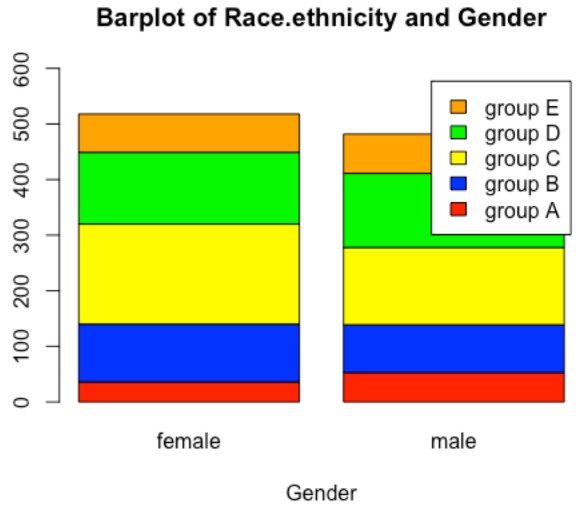
	group A	group B	group C	group D	group E	Sum
female	36	104	180	129	69	518
male	53	86	139	133	71	482
Sum	89	190	319	262	140	1000

	group A	group B	group C	group D	group E
female	0.036	0.104	0.180	0.129	0.069
male	0.053	0.086	0.139	0.133	0.071

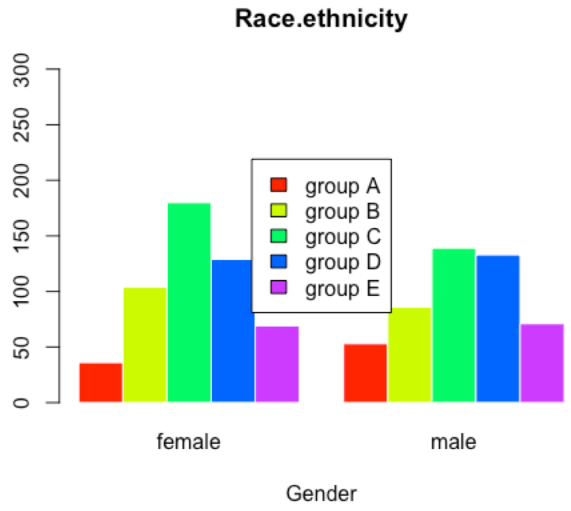
Graphical summarization of two-way tables







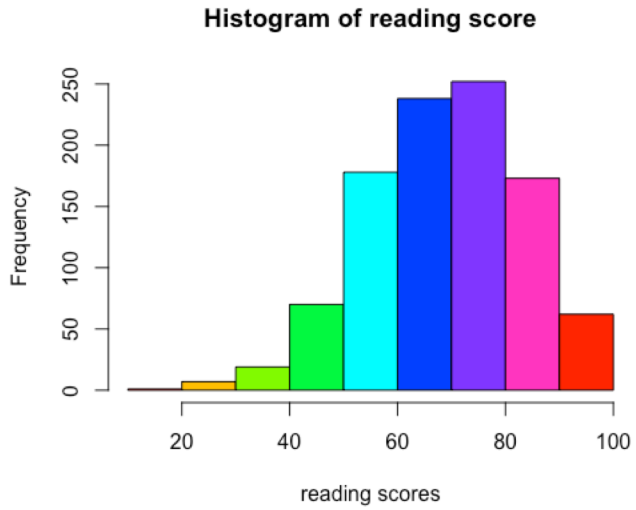
- From these plots, there are most female and male in the race of group C
- Conversely, among the group A of race, male and female are the least amount



Analysis of numerical data & examine distribution

- Pick the student's reading score as another numerical data and examine its distribution

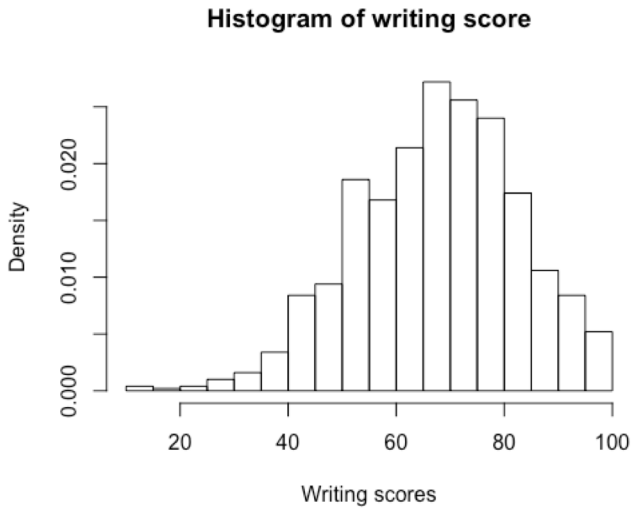
- From the distribution of the graph below, the distribution is left skewed

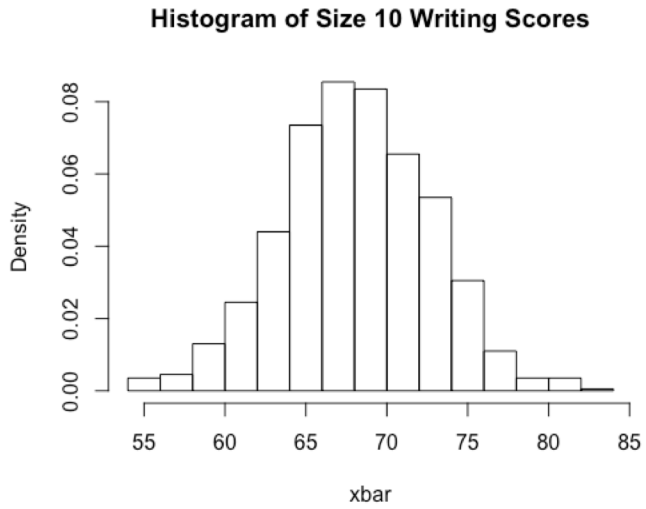


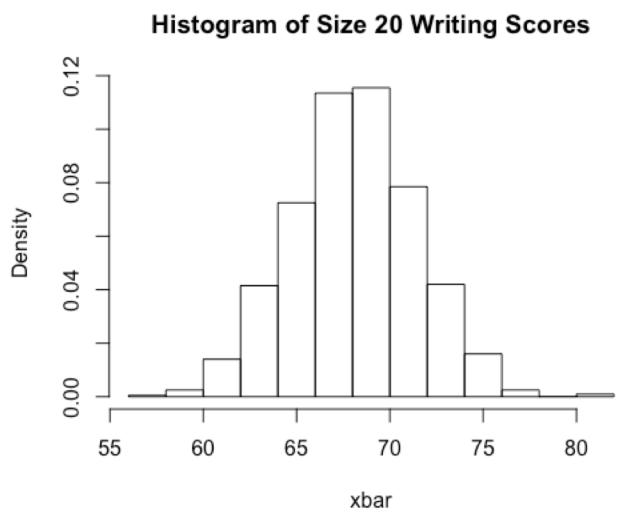
Various random samples of data & applicability of Central Limit Theorem

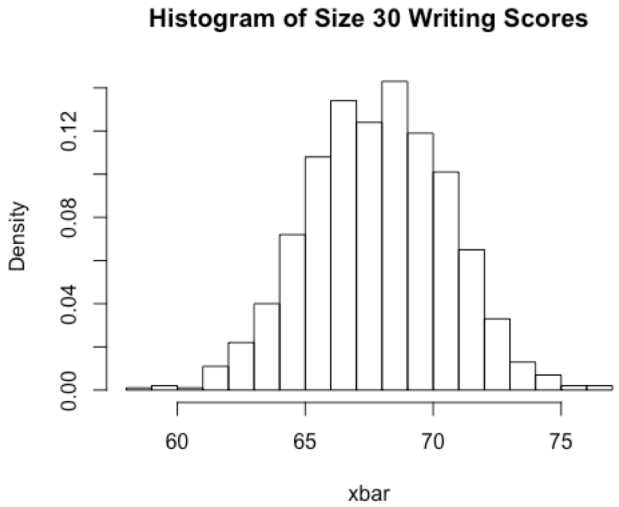
- Draw 1000 samples of the data of size 10, size 20, size 30, size 40 to show the histogram of densities sample means

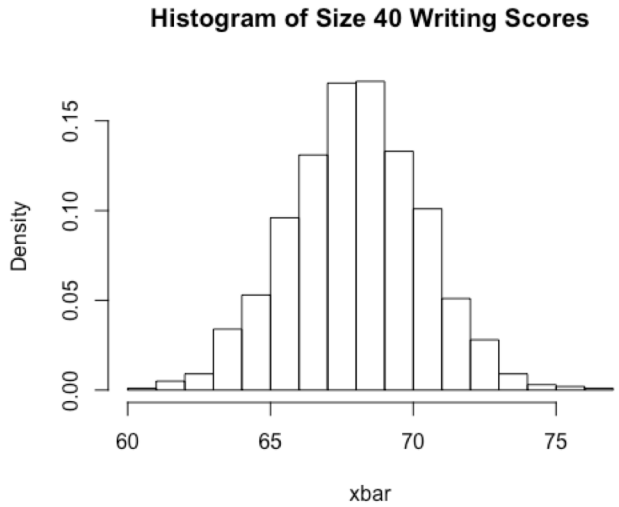
- Compute the means and standard deviations of the above four distribution












```
> data.info
   mean     sd
1 68.05 15.196
2 68.11  4.631
3 68.10  3.323
4 67.85  2.706
5 68.00  2.346
```

- As the sample size increase, the spread of the distribution becomes narrower and approach to the shape of normal distribution

- In the meantime, the mean of sample means gets closer to the mean of the data and the standard deviation will decrease as the sample size increases

Various sampling methods

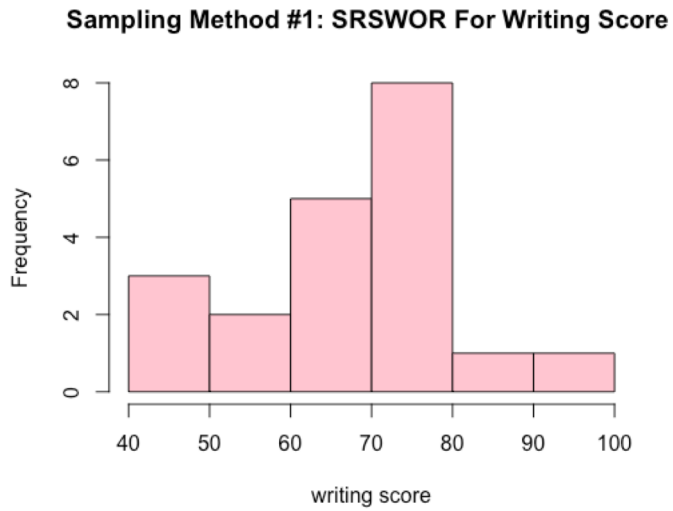
➢ Sampling method 1 — Simple Random Sampling

- Use sample size of 20 and show the frequencies for each student's writing scores in the dataset

```
> table(sample.1$writing.score)

40 45 47 54 55 62 64 65 70 72 73 76 77 78 79 89 95
 1  1  1  1  1  2  1  1  1  3  1  1  1  1  1  1  1
```

	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score	writing.score
26	male	group A	master's degree	free/reduced	none	73	74	72
143	female	group E	some college	free/reduced	completed	42	55	54
211	male	group D	some high school	free/reduced	completed	80	79	79
348	female	group C	bachelor's degree	standard	completed	77	94	95
355	female	group C	some college	standard	none	59	71	70
373	male	group D	some high school	standard	none	74	74	72



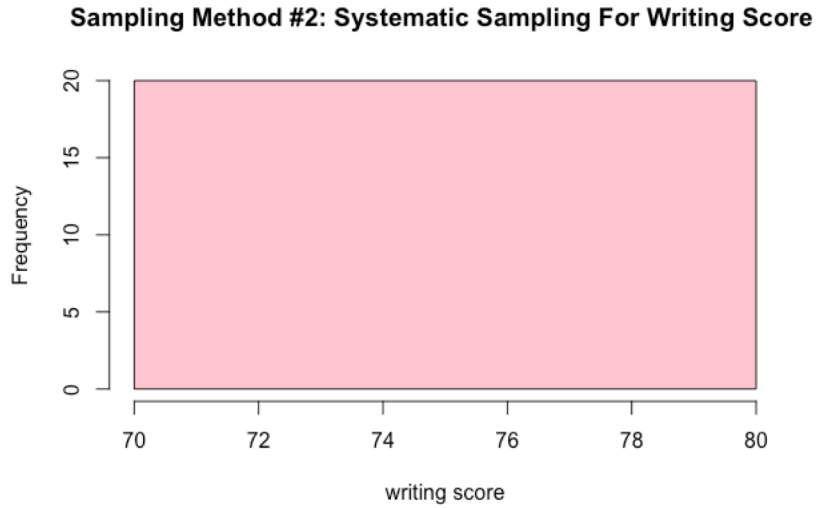
➢ Sampling method 2 — Systematic Sampling

- Use sample size of 20 and show the frequencies for each student's writing scores in the dataset

```
    gender race.ethnicity parental.level.of.education    lunch test.preparation.course math.score
1   female        group B           bachelor's degree standard                    none         72
1.1 female        group B           bachelor's degree standard                    none         72
1.2 female        group B           bachelor's degree standard                    none         72
1.3 female        group B           bachelor's degree standard                    none         72
1.4 female        group B           bachelor's degree standard                    none         72
1.5 female        group B           bachelor's degree standard                    none         72
    reading.score writing.score
1              72            74
1.1            72            74
1.2            72            74
1.3            72            74
1.4            72            74
1.5            72            74
```

```
> table(sample.2$writing.score)

74
20
```



➢ Sampling method 3 — Stratified Sampling

- Use sample size of 20 and show the frequencies for gender variable in the dataset

```
Stratum 1

Population total and number of selected units: 518 10.36
Stratum 2

Population total and number of selected units: 482 9.64
Number of strata  2
Total number of selected units 20
```

```
> table(st.sample.3$gender)

female   male
    10      9
```

	gender	ID_unit	Prob	Stratum
43	female	43	0.02	1
141	female	141	0.02	1
146	female	146	0.02	1
274	female	274	0.02	1
346	female	346	0.02	1
446	female	446	0.02	1
463	female	463	0.02	1
582	female	582	0.02	1
609	female	609	0.02	1
919	female	919	0.02	1
128	male	128	0.02	2
148	male	148	0.02	2
276	male	276	0.02	2
296	male	296	0.02	2
414	male	414	0.02	2
550	male	550	0.02	2
592	male	592	0.02	2
717	male	717	0.02	2
728	male	728	0.02	2

Conclusion

- If samples are used instead of whole dataset, as the sample size grows up, the mean of the sample mean distribution reaches the mean of the parent data and the standard deviation of the sample mean distribution will decrease

- In addition, with sampling, it can save our time consuming on our large dataset, a sample is used by the suitable sampling methods or strategy can also yield valid and reliable information

Thank you!