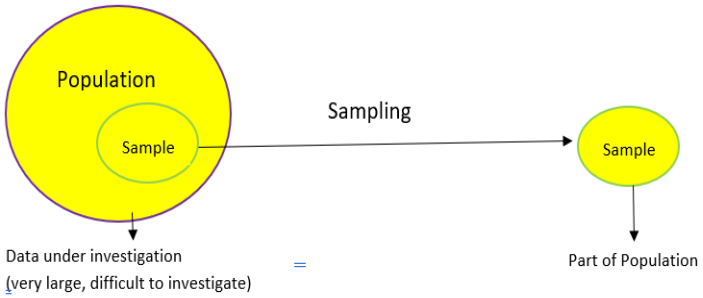
MAT 3103: Computational Statistics and Probability

Chapter 8: Sampling



Sampling:

It is a technique to select a representative part of population units, where units are investigated to study the characteristics of population units.

Explaining the need of sampling with real-life applications:

Sampling is useful as we can pair it with an inverse process known as generalization. To know a population, the steps we follow are: (1) select a sample from the population, (2) measure certain data or an opinion for all individuals in the sample and (3) project the result we observe in the sample onto the population. This projection or extrapolation is called generalization of results.

In cooking rice, we check the status of the rice by inserting a spoon until it touches the bottom of the pot, pull out the spoon, some rice will stick to it (sample), and taste the rice.

Blood specimen collection is performed routinely to obtain blood for laboratory testing. Specimens are often sent to help diagnose conditions such as electrolyte imbalances, to screen for risk factors like high cholesterol levels, and to monitor the effects of treatments and medications. Here, only a sample of blood is collected, not the entire amount of blood from the body is taken away.

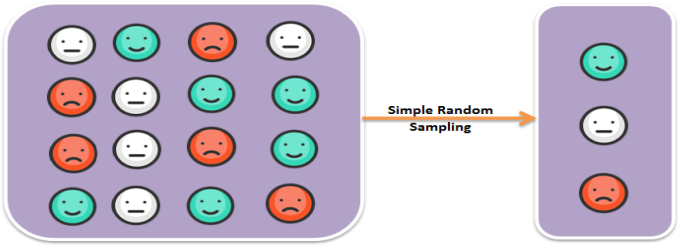
Different Methods of Sampling are:

There are several different sampling techniques available. In random sampling, we start with a complete sampling frame of all eligible individuals from which we select our sample. In this way, all eligible individuals have a chance of being chosen for the sample, and we will be more able to generalize the results from our study.

i) Simple random sampling, **ii)** Systematic sampling, **iii)** Circular systematic sampling iv) Stratified random sampling, **iv)** Cluster sampling, etc.

Simple Random Sampling:

Simple random sampling is a sampling technique where every unit in the population has an even chance and likelihood of being selected in the sample.



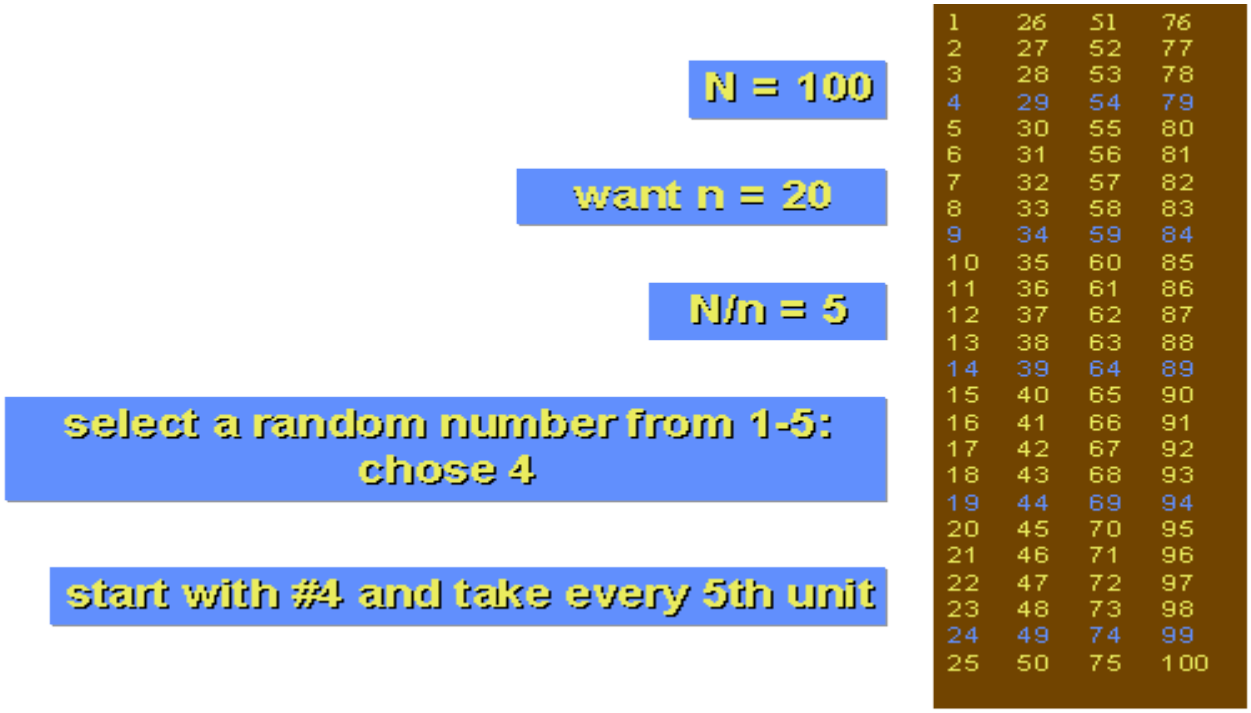
Let there be N units in a population. We need to select a random sample of size n ($n \leq N$). The possible number of samples, without replacement, are N_{c_n}. If any of these N_{c_n} samples are selected with equal probability $\frac{1}{N_{c_n}}$, then the sampling is simple random sampling. In other words, if every unit of N units is selected with equal probability $\frac{1}{N}$, then the sampling is simple random sampling.

Explaining simple random sampling with an example:

There are 40 students in a **Math** class at **AIUB**. The teacher wants to select a student as the class monitor. He makes 40 slips, write the IDs of the students on them distinctly and put them in a box. After shuffling the slips, he picks one up randomly and declare the student whose ID is there on the selected slip as the class monitor. Here, each and every single student has equal probability $\frac{1}{40}$ of being selected as the class monitor.

Systematic Sampling:

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point (R) but with a fixed, periodic interval. This interval, called the sampling interval $\left(k = \frac{N}{n}\right)$, is calculated by dividing the population size by the desired sample size.



Circular Systematic Sampling:

In this method, we assume the listings to be in a circle such that the last unit is followed by the first. A random start is chosen from 1 to *N*. We then add the intervals *k* until exactly n elements are chosen. If we come to the end of the list, you continue from the beginning.

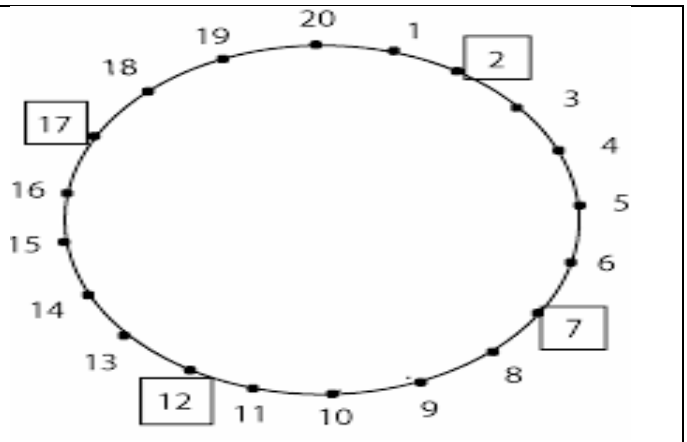
$N = 20$, $n = 4$, $k = \frac{N}{n} = 5$. Random start, R = 7. 7, 12, 17, and 2 are selected.	

Table of Random Numbers

36518 36777 89116 05542 29705 83775 21564 81639 27973 62413 85652 62817 57881
46132 81380 75635 19428 88048 08747 20092 12615 35046 67753 69630 10883 13683
31841 77367 40791 97402 27569 90184 02338 39318 54936 34641 95525 86316 87384
84180 93793 64953 51472 65358 23701 75230 47200 78176 85248 90589 74567 22633
78435 37586 07015 98729 76703 16224 97661 79907 06611 26501 93389 92725 68158
41859 94198 37182 61345 88857 53204 86721 59613 67494 17292 94457 89520 77771
13019 07274 51068 93129 40386 51731 44254 66685 72835 01270 42523 45323 63481
82448 72430 29041 59208 95266 33978 70958 60017 39723 00606 17956 19024 15819
25432 96593 83112 96997 55340 80312 78839 09815 16887 22228 06206 54272 83516
69226 38655 03811 08342 47863 02743 11547 38250 58140 98470 24364 99797 73498
25837 68821 66426 20496 84843 18360 91252 99134 48931 99538 21160 09411 44659
38914 82707 24769 72026 56813 49336 71767 04474 32909 74162 50404 68562 14088
04070 60681 64290 26905 65617 76039 91657 71362 32246 49595 50663 47459 57072
01674 14751 28637 86980 11951 10479 41454 48527 53868 37846 85912 15156 00865
70294 35450 39982 79503 34382 43186 69890 63222 30110 56004 04879 05138 57476
73903 98066 52136 89925 50000 96334 30773 80571 31178 52799 41050 76298 43995
87789 56408 77107 88452 80975 03406 36114 64549 79244 82044 00202 45727 35709
92320 95929 58545 70699 07679 23296 03002 63885 54677 55745 52540 62154 33314
46391 60276 92061 43591 42118 73094 53608 58949 42927 90993 46795 05947 01934
67090 45063 84584 66022 48268 74971 94861 61749 61085 81758 89640 39437 90044
11666 99916 35165 29420 73213 15275 62532 47319 39842 62273 94980 23415 64668
40910 59068 04594 94576 51187 54796 17411 56123 66545 82163 61868 22752 40101
41169 37965 47578 92180 05257 19143 77486 02457 00985 31960 39033 44374 28352
76418

Some formulas to estimate different statistic values:

The **estimate of sample means,** $\bar{x} = \frac{1}{n}\Sigma x$.

The estimate of **sample variance,** $s^2 = \frac{1}{n-1}\ [\Sigma x^2 - \frac{(\Sigma x)^2}{n}]$.

The **estimate of the variance of sample means**, $v(\bar{x}) = \frac{N-n}{Nn}s^2$.

The **estimated standard error of sample means**, s.e. $(\bar{x}) = \sqrt{v(\bar{x})}$.

The **estimate of population total**, $\hat{X} = N\bar{x}$

The **estimate of variance of the estimate of population total**, $v(\hat{X}) = N^2 v(\bar{x})$;

$$s.e.(\hat{X}) = \sqrt{v(\hat{X})}$$

Estimate of proportion: Let

N = Number of population units,

n= Number of sample units,

A = Number of units in the population possessing a particular character,

a = Number of sample units possessing that particular character,

P = ***A*** **/*N*** =Proportion of **population** units possessing that particular character,

P* = *a* /*n = Proportion of **sample** units possessing that particular character.

This ***p* is an unbiased estimate of *P***. The **estimate of variance of proportion** is given by

$$v(p) = \frac{N-n}{N(n-1)} pq \ , \ q = 1 - p$$

Problem 8.1: Number of pharmacies in various regions of a city are, *X*: 12, 20, 5, 25, 10, 35, 8, 15, 20, 13, 20, 18, 8, 9, 24, 25, 15, 30, 18, 22, 25, 17, 30, 25, 18, 20, 22, 20. Select a random sample of 6 regions by **(i)** simple random sampling, **(ii)** systematic sampling, **(iii)** circular systematic sampling.

1. Estimate mean number of pharmacies per region.
2. Estimate total number of pharmacies in the city.
3. Estimate standard error of estimated mean number of pharmacies.
4. Estimate standard error of estimated total number of pharmacies.
5. Find 95% confidence interval for mean number of pharmacies.
6. Suggest a sample of size *n* to estimate a population mean with margin of error 0.5 at 95% level of confidence, where variance of the population observations is 6.35.
7. Estimate the proportion of regions in which there are less than 18 pharmacies.
8. Estimate the variance of the estimated proportion.
9. Find sample size *n* to estimate proportion 0.75 with margin of error 0.2 at 95% confidence.

Observations (*x*)	12	20	5	25	10	35	8	15	20	13	20	18	8	9
Serial Number	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***
Observations (*x*)	24	25	15	30	18	22	25	17	3	25	18	20	22	20
Serial Number	***15***	***16***	***17***	***18***	***19***	***20***	***21***	***22***	***23***	***24***	***25***	***26***	***27***	***28***

i) We have $N = 28$. We need to select a sample of size $n = 6$ using Random Number Table. We can use any row or any column of the table. Let us use Column 1. Here $N = 28$, the last serial

number is of two digits. So, we need to select a random number of two digits. The selected random numbers and the selected number of pharmacies of different regions are shown below:

Random Numbers	16	11	10	19	17	9
Pharmacies of selected regions	25	20	13	18	15	20

ii) Let $N = nk$, $k = N/n$. In our case $k = 28/6 = 4.7 \sim 5$. We have to select first observation from first $k = 5$ observations using Random Number Table. After that every $k^{th} = 5^{th}$ observation is selected. Five is a number of one digit, so we need to select a random number of one digit first. The selected random numbers and the selected observations are shown below: [using column 2 of random number table]

Random Numbers	3	8	13	18	23	28
Pharmacies of selected regions	5	15	8	30	30	20

iii) Here also $k = N/n = 4.7 \sim 5$. First observation is selected from all observations using random number table. After that every $k^{th} = 5^{th}$ observation is selected moving through a circle. We have $n = 28$ observations and 28 is a number of two digits. So, we need to select a random number of two digits. The selected random number and the selected observations are shown below: [Using column 3 of random number table]

Random Numbers	10	15	20	25	2	7
Pharmacies of selected regions	13	24	22	18	20	8

1. Estimate of mean, $\bar{x} = \frac{1}{n}\Sigma x = \frac{111}{6} = 18.5$. [Calculation is from Simple random sample]
2. Estimate of total, $\hat{X} = N\bar{x} = 28 \times 18.5 = 518.0$.
3. The standard error of estimate of mean is, $\text{s.e.}(\bar{x}) = \sqrt{v(\bar{x})} = \sqrt{2.334} = 1.531$.

   The variance of sample mean is $v(\bar{x}) = \frac{N-n}{Nn}s^2 = \frac{28-6}{28\times 6}17.9 = 2.344$.

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \frac{1}{5}(2143 - \frac{(111)^2}{6}) = 17.9$$

4. The estimate of standard error of estimate of population total is given as follows:

$v(\hat{X}) = N^2 v(\bar{x}) = 28^2\ 2.344 = 14033.6$ and $s.e.(\hat{X}) = \sqrt{v(\hat{X})} = \sqrt{14033.60} = 118.463$

5. 95% confidence limits for mean are given as follows:

   $\bar{x}_l = \bar{x} - t_{n-1}$ s.e. $(\bar{x}) = 18.5 - 2.571\ (1.531) = 14.56$, Here, $t_{n-1} = t_5 = 2.571$

$\overline{x_u} = \bar{x} + t_{n-1}$ s.e. $(\bar{x}) = 18.5 + 2.571\ (1.531) = 22.43$

6. The sample size *n* is given by $n = \frac{z^2\ v(x)}{d^2} = \frac{(1.96)^2\,(6.35)^2}{(0.5)^2} = 97.57 \sim 98$

   Here *z* is the tabulated of normal distribution at 5% level = 1.96, *d* = margin of error.

7. The estimate of proportion of regions in which there are less than 18 pharmacies is given by $p = a / n = 2/6 = 0.33$. Here *a* = number of regions in the sample in which there are less than 18 pharmacies = 2.

8. The estimated variance of p is given by

   $v(p) = \frac{N-n}{N(n-1)} pq = \frac{28-6}{28(6-1)} 0.33 \times 0.67 = 0.0303$. Here $q = 1 - p = 1 - 0.33 = 0.67$

9. The sample size *n* is given by, $n = \frac{z^2\ pq}{d^2} = \frac{(1.96)^2 \times 0.75 \times 0.25}{(0.2)^2} = 18$

Unbiasedness of mean and variance in case of simple random sampling:

Let us consider that, in a population there are $N = 3$ units. The unit values are *x*: 2, 4, 6. We can select a sample of size $n = 2$. The possible number of samples, without replacement, are ${}_3C_2 = 3$.

Samples	**Sample Means,** $\bar{x} = \frac{1}{n}\Sigma x$	**Sample Variances,** $s^2 = \frac{1}{n-1}\ [\Sigma x^2 - \frac{(\Sigma x)^2}{n}]$
2, 4	$\bar{x}_1 = \frac{1}{n}\Sigma x = \frac{2+4}{2} = 3$	$s_1^2 = \frac{1}{2-1}[(2^2 + 4^2) - \frac{(2+4)^2}{2}] = 2$
2, 6	$\bar{x}_2 = \frac{2+6}{2} = 4$	$s_2^2 = \frac{1}{2-1}[(2^2 + 6^2) - \frac{(2+6)^2}{2}] = 8$
4, 6	$\bar{x}_3 = \frac{4+6}{2} = 5$	$s_3^2 = \frac{1}{2-1}[(4^2 + 6^2) - \frac{(4+6)^2}{2}] = 2$

$$\text{Population Mean, } \bar{X} = \frac{1}{N}\Sigma X = \frac{1}{3}(2+4+6) = 4,$$

$$\text{Population Variance, } S^2 = \frac{1}{N-1}[\Sigma x^2 - \frac{(\Sigma x)^2}{N}]$$

$$= \frac{1}{3-1}[(2^2 + 4^2 + 6^2) - \frac{(2+4+6)^2}{3}] = 4$$

Sample **mean** is an **unbiased** estimate of population **mean** as: $E(\bar{x}) = \frac{1}{3}(3+4+5) = 4 = \bar{X}$.

Sample **variance** is an **unbiased** estimate of population **variance** as: $E(s^2) = \frac{1}{3}(2+8+2) = 4 = S^2$.

Sampling Distribution: The distribution of sample means or sample variances or any function of these is known as sampling distribution. Some distributions are as:

(i) Student's *t* – Distribution, **(ii)** Chi-square (χ^2) Distribution,

(iii) *F* – distribution [Distribution of Variance Ratio].

Student's *t* – Distribution, Chi-square (χ^2) Distribution and F- Distribution:

$Let\ x_1, x_2, \ldots\ldots\ldots x_n\ be\ observations\ selected\ from\ N(\mu, \sigma^2).\ Then, we\ have$

$\bar{x} = \frac{1}{n}\Sigma x\ , and\ s^2 = \frac{1}{n-1}[\Sigma x^2 - \frac{(\Sigma x)^2}{n}].\ Here$

$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}\ is distributed as\ Student's\ t\ with\ (n-1) d.f.\ and$

$\chi^2 = \frac{(n-1)\ s^2}{\sigma^2} is\ distributed as chi-square with (n-1)\ d.f.$

$F = \frac{s_1^2}{s_2^2}\ is\ distributed as variance\ ratio\ with\ (n_1 - 1)\ and\ (n_2 - 1)\ d.f.$

$Here\ s_1^2\ is\ the\ variance\ of\ first\ sample\ and\ s_2^2\ is\ the\ variancer\ of\ second\ sample.$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

This *z* is similar to *t* and it is distributed as Normal distribution with mean zero and variance 1. It is used if sample size *n* is big (≥ 30) and /or σ is known. If σ is not known and sample is small ($n < 30$), then *t*-statistic is used.

Exercise 8

8.1 Define sampling and simple random sampling with example.

8.2 Show, by an example, mean of simple random sample is an unbiased estimate of population mean.

8.3 Show, by an example, variance of simple random sample is an unbiased estimate of population variance.

8.4 The number of signals received in a server in different days are, X: 5, 8, 7, 10, 7, 6, 9, 11, 4, 2, 7, 7, 12, 9, 11, 3, 7, 8, 5, 6, 7, 6, 9, 11, 4. Select 4 days by systematic sampling.

a) Estimate total number of signals received per day along with its estimated standard error.

b) Estimate the proportion of days in which less than 8 signals are received.

8.5 The following are the number of faded out signals sent from a station in different days:

X: 4, 3, 0, 2, 6, 7, 4, 3, 2, 0, 1, 0, 3, 0, 6, 8, 0, 1, 4 ,3, 2, 6, 3, 7, 5, 8, 0, 2, 3, 5.

Select a random sample of 5 days by simple random sampling method and estimate total number of faded out signals along with its estimated standard error

1

8.6 Suggest a sample size n to estimate a proportion 0.45 of the number of days in which more than 10 signals are faded, with margin of error 0.1 at 95% confidence.

8.7 The number of mails received in a server at Bashundhara residential area in different days are:

X: 10, 7, 6, 9, 11, 4, 2, 7, 7, 9, 11, 45, 8, 7, 10, 7, 6, 9, 11, 4, 2, 7, 7.

Select 4 days by simple random sampling. Estimate mean number of emails received per day along with its estimated standard error.

8.8 Suggest a sample size n to estimate a proportion 0.3 of the number of days in which more than 10 signals are faded, with margin of error 0.05 at 95% confidence.

8.9 The number of noisy bits produced by an electronic device in different attempts is given as,

X: 5, 8, 7, 10, 7, 6, 9, 11, 4, 2, 7, 7, 12, 9, 11, 3, 7, 8, 5, 6.

Select a random sample of 5 attempts by circular systematic sampling and find 95% confidence interval for the average of noisy bits.

Sample MCQs

1. Suggest a sample of size n to estimate a population mean with margin of error 0.3 at 95% level of confidence, where variance of the population observations is 8.

a) 341 b) 431 c) 414 d) 342

2. The number of miss calls received by a person in different days are randomly observed. Number of miss calls: 7, 3, 10, 6. These are selected from the record of a month.
Estimate variance of estimated mean.

a) 3.41 b) 1.80 c) 2.14 d) 0.42

3. The number of miss calls received by a person in different days are randomly observed. Number of miss calls: 7, 13, 10, 15, 20. These are selected from the record of a month.
Estimate the total number of miss calls.

a) 341 b) 331 c) 390 d) 342

4. The number of miss calls received by a person in different days are randomly observed. Number of miss calls: 7, 13, 10, 15, 20, 8, 12.

Estimate the variance of proportion of days in which the number of miss calls are less than 10.

a) 0.045 b) 0.031 c) 0.286 d) 0.026