Information Collection & Sampling techniques

- Information can be obtained from the following two methods.
  - **Census:**
    - obtaining information for the entire population of interest.
    - conducting a census may be time consuming, costly, impractical, or even impossible.
  - **Sampling:**
    - Process of selecting a sample.
    - Method for obtaining a sample from the population.
    - Because the sample will be used to draw conclusions about the entire population, it should be a **representative sample.**

what can happen when a sample is not representative?

To see what can happen when a sample is not representative, consider the presidential election of 1936. Before the election, the *Literary Digest* magazine conducted an opinion poll of the voting population. Its survey team asked a sample of the voting population whether they would vote for Franklin D. Roosevelt, the Democratic candidate, or for Alfred Landon, the Republican candidate.

Based on the results of the survey, the magazine predicted an easy win for Landon. But when the actual election results were in, Roosevelt won by the greatest landslide in the history of presidential elections! What happened?

- The sample was obtained from among people who owned a car or had a telephone. In 1936, that group included only the more well-to-do people, and historically such people tend to vote Republican.
- The response rate was low (less than 25% of those polled responded), and there was a nonresponse bias (a disproportionate number of those who responded to the poll were Landon supporters).

Probability Sampling

- Most modern sampling procedures involve the use of **probability sampling.**
- In probability sampling, a random device—such as tossing a coin, consulting a table of random numbers, or employing a random-number generator—is used to decide which members of the population will constitute the sample instead of leaving such decisions to human judgment.
- Probability sampling **eliminates unintentional selection bias** and permits the researcher to **control the chance of obtaining a non-representative sample**.

Simple Random Sampling

Simple Random Sampling; Simple Random Sample

Simple random sampling: A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained.

Simple random sample: A sample obtained by simple random sampling.

- There are two types of simple random sampling.
  - Sampling with replacement
  - Sampling without replacement

Example # 01: Sampling without replacement

- The top five state officials of Oklahoma are as shown: **(i)** Governor (G), **(ii)** Lieutenant Governor (L), **(iii)** Secretary of State (S), **(iv)** Attorney General (A), and **(v)** Treasurer (T). Consider these five officials a population of interest.

a. List the possible samples (without replacement) of two officials from this population of five officials.
b. Describe a method for obtaining a simple random sample of two officials from this population of five officials.
c. For the sampling method described in part (b), what are the chances that any particular sample of two officials will be the one selected?
d. Repeat parts (a)–(c) for samples of size 4.

Example # 02: Sampling with replacement

- Select random samples of size 2 from a population of 5 students: Ahmer, Bilal, Chand, Daniyal and Ejaz.

AA,	**BA,**	**CA,**	**DA,**	**EA**
AB,	**BB,**	**CB,**	**DB,**	**EB**
AC,	**BC,**	**CC,**	**DC,**	**EC**
AD,	**BD,**	**CD,**	**DD,**	**ED**
AE,	**BE,**	**CE,**	**DE,**	**EE**

- (a) P(Ahmer)

Example # 03:

- Assume students as a population and the marks obtained by them in a certain statistics class are 20, 15, 12, 16, and 18. Draw all possible random samples of two students when sampling is done
- (i) with replacement
- ii) without replacement
- Calculate the mean marks for each samples.

SAMPLING DISTRIBUTION

- The probability distribution of a **statistic** computed from all possible samples of the same size is called a **sampling distribution**.
- Since sampling distribution is a probability distribution, therefore the sum of all probabilities is always one.
- Sampling distribution of a statistics provide all the information one needs in making decisions about the values of the population parameters.
- The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples.

Standard Error (S.E) & Sampling Error

- The standard deviation of a sampling distribution of a sample statistic is called the standard error of the statistic.
  - It measures the dispersion of the value of a statistic, that might be computed from all possible samples.
- **Sampling error** is the error resulting from using a sample to estimate a population characteristic.

Sampling Distribution of Sample Mean

- Suppose a researcher selects a sample of 30 adult males and finds the mean of the measure of the triglyceride levels for the sample subjects to be 187 milligrams/deciliter.
- Then suppose a second sample is selected, and the mean of that sample is found to be 192 milligrams/deciliter.
- Continue the process for 100 samples.
- 187, 192, 184, . . . , 196
- What happens?
- The mean becomes a random variable.

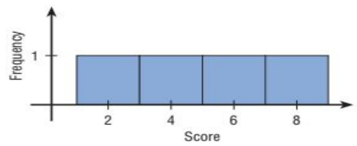
Sampling Distribution of Sample Mean

A **sampling distribution of sample means** is a distribution using the means computed from all possible random samples of a specific size taken from a population.

Properties of the distribution of Sample Means

1. The mean of the sample means will be the same as the population mean i.e.
$$\mu_{\overline{X}} = \mu$$
2. The standard deviation of the sample means will be smaller than the standard deviation of the population, and it will be equal to the population standard deviation divided by the square root of the sample size.
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Example # 03:

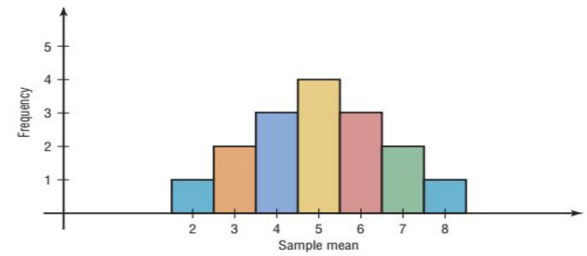


- Suppose a professor gave an 8-point quiz to a small class of four students. The results of the quiz were 2, 6, 4, and 8. For the sake of discussion, assume that the four students constitute the population.
- (a) Calculate population mean and standard deviation.
- (b) Draw the graph of original distribution.
- (c) Draw all possible samples of size 2 with replacement & Calculate mean of each sample.
- (d) Construct Frequency distribution of Sample Mean and repeat Part (a) & (b)

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

Frequency Distribution of Sample Mean

$\overline{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1



The Central Limit Theorem

As the sample size n increases without limit, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution. As previously shown, this distribution will have a mean μ and a standard deviation $\sigma/\sqrt{n}$.

It's important to remember two things when you use the central limit theorem:

1. When the original variable is normally distributed, the distribution of the sample means will be normally distributed, for any sample size n.
2. When the distribution of the original variable might not be normal, a sample size of 30 or more is needed to use a normal distribution to approximate the distribution of the sample means. The larger the sample, the better the approximation will be.

Central Limit Theorem (Contd.)

- If the sample size is sufficiently large, the central limit theorem can be used to answer questions about sample means in the same manner that a normal distribution can be used to answer questions about individual values. The only difference is that a new formula must be used for the *z* values. It is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Example # 04

- **Hours That Children Watch Television**: A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.
- **Age of Vehicle:** The average age of a vehicle registered in the United States is 8 years, or 96 months. Assume the standard deviation is 16 months. If a random sample of 36 vehicles is selected, find the probability that the mean of their age is between 90 and 100 months.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Summary of Formulas

Formula	Use
1. $z = \dfrac{X - \mu}{\sigma}$	Used to gain information about an individual data value when the variable is normally distributed.
2. $z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$	Used to gain information when applying the central limit theorem about a sample mean when the variable is normally distributed or when the sample size is 30 or more.

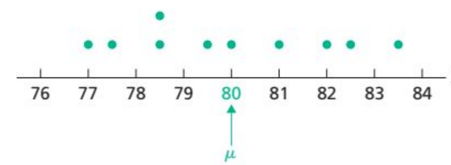
Example # 05:

- Suppose that the population of interest consists of the five starting players on a men's basketball team, who we will call A, B, C, D, and E. Further suppose that the variable of interest is height, in inches.

Player	A	B	C	D	E
Height	76	78	79	81	86

- **(a.)** Obtain the sampling distribution of the sample mean for samples of size 2.
- (**b.)** Make some observations about sampling error when the mean height of a random sample of two players is used to estimate the population mean height.

Sample	Heights	$\bar{x}$
A, B	76, 78	77.0
A, C	76, 79	77.5
A, D	76, 81	78.5
A, E	76, 86	81.0
B, C	78, 79	78.5
B, D	78, 81	79.5
B, E	78, 86	82.0
C, D	79, 81	80.0
C, E	79, 86	82.5
D, E	81, 86	83.5

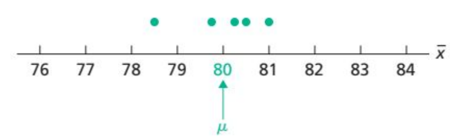


Example # 06:

a. Obtain the sampling distribution of the sample mean for samples of size 4.
b. Make some observations about sampling error when the mean height of a random sample of four players is used to estimate the population mean height.
c. Find the probability that, for a random sample of size 4, the sampling error made in estimating the population mean by the sample mean will be 1 inch or less; that is, determine the probability that $\bar{x}$ will be within 1 inch of μ.

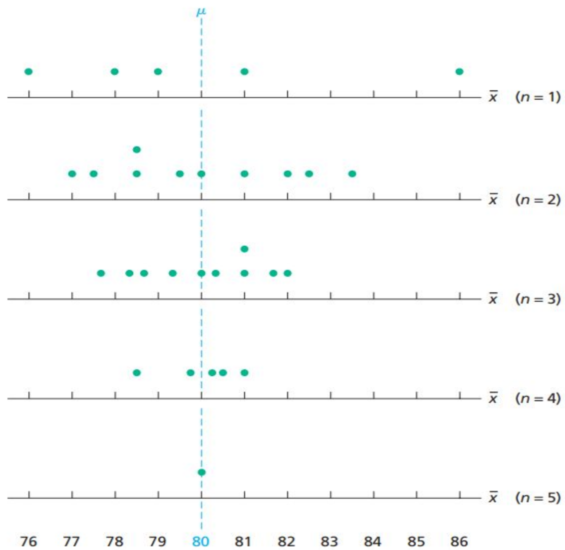
Player	A	B	C	D	E
Height	76	78	79	81	86

Sample	Heights	$\bar{x}$
A, B, C, D	76, 78, 79, 81	78.50
A, B, C, E	76, 78, 79, 86	79.75
A, B, D, E	76, 78, 81, 86	80.25
A, C, D, E	76, 79, 81, 86	80.50
B, C, D, E	78, 79, 81, 86	81.00



Interpretation There is an 80% chance that the mean height of the four players selected will be within 1 inch of the population mean.

Sample Size and Sampling Error



Finite Population Correction Factor

- The formula for the standard error of the mean is accurate when the samples are drawn with replacement Or are drawn without replacement from a very large or infinite population.
- **Since sampling with replacement is for the most part unrealistic**, a *correction factor* is necessary for computing the standard error of the mean for samples drawn **without replacement** from a finite population. $\sqrt{\frac{N-n}{N-1}}$

  where N is the population size and n is the sample size.
- This correction factor is necessary if relatively large samples are taken from a small population, because the sample mean will then more accurately estimate the population mean and there will be less error in the estimation.

Finite Population Correction Factor (Contd.)

- Therefore, the standard error of the mean must be multiplied by the correction factor to adjust for large samples taken from a small population. Finally, the formula for the *z* value becomes:

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}}$$

- When the population is large and the sample is small, the correction factor is generally not used, since it will be very close to 1.00.

Example # 07

- Consider the following population of size 5: 0, 3, 6, 9 & 12.
- (a) Draw all possible samples of size 3 w/o replacement and the find mean of each sample.
- Find sampling distribution of sample mean and verify that
- $V(\bar{X}) = \frac{\sigma^2}{n} . \frac{N-n}{N-1}$
- Find standard error of $(\bar{X})$

Sampling distribution of Difference between means

Example # 08: Sampling Difference b/w two means

- Consider the following finite populations:

  Population – I: 3, 4, 5 & **Population – II:** 0, 3

(a) Draw all possible random samples of size 2 from population – I with replacement.

(b) Draw all possible random samples of size 3 from population – II with replacement.

(c) Construct a sampling distribution of their differences $(\overline{x}_1 - \overline{x}_2)$.

(d) Verify that $\mu_{\overline{x}_1 - \overline{x}_2} = \mu_1 - \mu_2$ & $\sigma^2_{\overline{x}_1 - \overline{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$