Chapter 7
Sampling and Sampling Distribution of statistic

March,2023

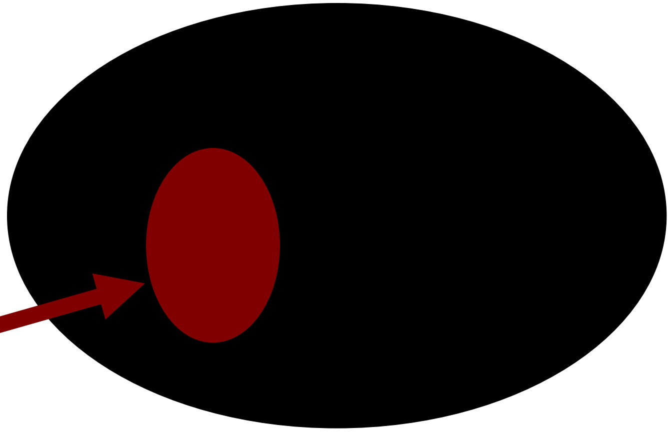
Sampling, sampling techniques and sampling distributions

Definitions:

Population:

a set which includes all measurements of interest to the researcher

(The collection of ***all*** responses, measurements, or counts that are of interest)



Sample:

A subset of the population

As the population is too large for us to consider collecting information from all its members, we obliged to take a representative sample(has all the important characteristic of the population from which is it drawn).

Sampling is a process of selecting some members of a given population as representatives of the entire population in terms of the desired characteristics.

- Sampling is simply the process of learning about the population on the basis of a sample drawn from it.

Common Terminologies

- ***Sampling fraction***:-The ratio of reference population to the number of units in the sample (N/n).
- ***Sampling frame***:-is the list of units from which the sample was selected.

- **Target Population**: The population to be studied/ to which the investigator wants to generalize his results
- **Sampling Unit**: smallest unit from which sample can be selected

Parameter: Characteristic or measure obtained from a population.

Statistic: Characteristic or measure obtained from a sample.

Sampling: The process or method of sample selection from the population.

Difference between Census and Sample Method

- **Census Survey Method**

Under the census or complete enumeration survey method, data are collected for each and every unit (person, household, field, shop, factory etc.), as the case may be of the **population** or **universe**, which is the complete set of items, which are of interest in any particular situation.

- **Sample Survey Method**

Sampling is a method used in statistical analysis in which a decided number of considerations are taken from a comprehensive population or a **sample survey**. Thus, in the sample survey instead of every unit of the population only a part of the population is studied and the conclusions are drawn on that basis for the entire universe.

Sampling…why?

Because	But…..
Reduced cost Greater speed Greater scope Greater accuracy Feasibility	There is always sampling error Sampling may create a feeling of discrimination in the population. Inadvisable where every unit in the population is legally required to have a record

When and Where sampling technique is appropriate?

- **Vast data**

  – No. of units is very large-S economizes money,

  time &effort

- **When utmost accuracy is not required**

  – suitable in those situations where 100%

  accuracy is not required

- **Where census is impossible**

  -- not enumerating all individuals

- **Homogeneity**

  – if all the units are alike. Sampling is very easy to use

Sampling…Methods/types

- Two broad categories of sampling procedures are: ***probability methods*** and ***non-probability methods***.

A. Probability sampling methods

- Involves random selection of a sample
- A **sample** is obtained in a way that ensures every member of the population to have **a known (non zero) probability** of being included in the sample.
- Let sampling frame be N & sample size be n, every individual has a known chance of being selected

Sampling….probability

- Generalization is possible from sample to population
- more complex, more time consuming and usually more costly
- The method chosen depends on a number of factors, such as
  - The available sampling frame,
  - How spread out the population is,
  - How costly it is to survey members of the population

Sampling….probability

- **Most common probability sampling methods**

1. Simple random sampling
2. Systematic random sampling
3. Stratified random sampling
4. Cluster sampling

Sampling….probability

1. ***Simple (Unrestricted) random sampling***

- Principle
  - Equal chance/probability of each unit being drawn
- Procedure
  - Take sampling population
  - Need listing of all sampling units ("sampling frame")
  - Number all units
  - Randomly draw units
- Then we can apply methods like
  - Lottery method (sample drawn from box)
  - Table of random numbers
  - Computer generated random numbers

Sampling….probability

Lottery method

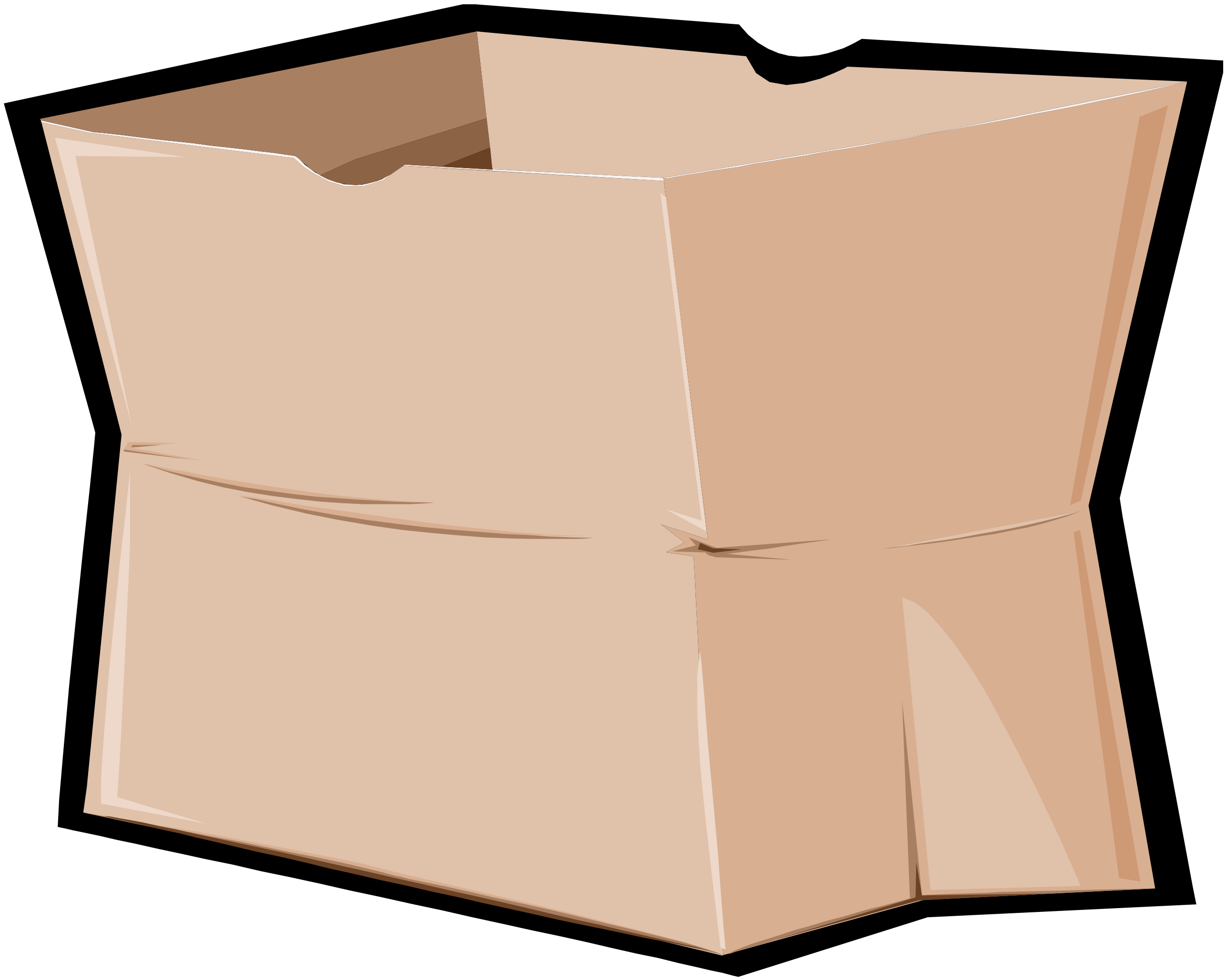


Table B.1: Random Numbers Table

	A	B	C	D	E	F	G	H	I	J
1	8450	6992	6563	0340	2649	6933	9446	6182	2601	7800
2	5952	1443	7100	8444	3904	0159	1849	2601	9763	9058
3	6711	6779	9388	9663	4167	1423	2744	4622	2179	8503
4	2681	8047	0494	7853	8411	5406	8127	9677	8530	2350
5	0739	3114	3997	3482	3226	2216	6874	0620	8521	2938
6	8985	2463	5054	3448	6357	0187	6342	4740	4064	5068
7	7644	9339	8376	4583	7715	6365	6827	2055	9328	3287
8	6277	6631	8797	3693	6370	1436	1599	6267	2758	0323
9	6355	7690	7628	9054	0022	4241	7499	3430	3644	6576
10	7828	0689	3075	1954	5972	2266	0055	1097	9706	9009
11	6026	4546	4119	1554	4895	3123	9849	2094	5062	6711
12	8416	1972	9345	1593	2943	2379	5062	4829	5952	8292
13	1433	8823	7706	5273	6160	2161	5510	8617	7894	0176
14	0622	4884	8113	4447	5736	6347	7280	2301	2330	0693
15	4104	7164	1184	3964	2119	6968	0469	3827	0845	8400
16	4272	4979	1471	0942	9573	4283	1557	0161	3957	2516
17	1225	4171	3433	8700	0042	5834	2508	3250	1520	6366
18	7442	6675	1927	7267	7182	3960	4341	0350	1126	5945
19	4911	9007	3048	0319	0916	3002	1466	4421	7246	7662
20	3143	7402	4486	0909	1858	7961	1211	6296	5545	4588
21	8055	9294	2578	0426	4322	6925	2487	5677	9491	4301
22	9240	5260	7134	8001	0140	3394	8437	4066	2856	0933
23	7923	8630	3654	2638	2868	1059	0903	3114	6351	8261
24	0020	5104	4344	3324	9214	6615	5926	7012	9052	9205
25	3312	6923	5469	9171	4877	5392	3394	5077	3750	5637
26	3466	4193	5330	4680	0456	5891	3175	5733	5678	0956
27	1677	1694	1697	8921	2520	2811	3597	1355	9606	3637
28	3846	6283	0969	0051	5867	1043	1671	2013	8956	7706
29	8084	2327	0550	7231	1037	4830	9742	5654	5458	8290
30	2715	2247	4504	1374	9236	7340	1773	0693	2749	1336
31	6537	6815	9312	1460	6593	7678	4312	7537	9360	7195
32	4263	8931	1642	6694	1925	2661	1274	7346	8234	3159
33	7468	4077	6691	3961	7640	2355	9938	8485	9398	8364
34	4884	3324	3690	7433	1245	0623	4483	5933	5634	0512
35	7222	7299	1346	8937	0933	1569	5562	3735	2982	5966
36	5040	0820	8606	4006	4743	6343	4873	1002	4757	1076
37	2980	4860	5894	1501	5791	9414	7246	1283	9766	7427
38	8660	5480	7436	9745	8869	3307	4916	6543	9830	6099
39	7627	4959	6417	3542	1877	0370	5464	9690	5184	7379
40	1890	7664	7144	3523	8465	0385	8174	4740	3654	5543
41	3175	2680	3919	7436	0796	1018	5565	1142	4577	0457
42	7616	9338	6304	0283	6502	9085	5443	1531	9724	4140
43	5223	4625	0896	9930	0050	2201	5270	6447	1850	2070
44	9384	9794	8418	0374	4119	2075	0067	4535	7769	4719
45	5862	9165	5302	9789	5771	9670	7523	9280	2604	0212
46	9450	9307	6597	7183	5243	8854	6735	2415	0364	3096

Sampling….probability

Computer generated Random numbers

OpenEpi | Start | Enter | Results | Examples | Help

Calculate

Clear

Random Number Generator		
Range	Lowest value	1
	Highest value	500
How many would you like?		25
Format into how many columns?		5
Omit text from output?		no

OpenEpi | Start | Enter | Results | Examples | Help

357	449	254	433	388
416	101	53	489	392
14	462	431	39	307
236	447	290	400	68
38	186	331	245	469

Print the numbers from the browser File menu, or copy and paste them to word processors, Excel, and other programs.

The numbers are generated by the JavaScript Math.random() function. Although these are pseudorandom numbers, the Math.random function in common browsers has been tested by many and found to generate high quality 'random' numbers. For more information, search the internet for 'random number quality' and related topics.

Results from OpenEpi, Version 2, open source calculator--Random

file:///C:/Program%20Files/OpenEpi/Random/Random.htm

Source file last modified on 11/09/2007 21:51:00

Sampling….probability

Simple random sampling....

- **Advantages**
  - No possibility of personal bias which affecting the results.
  - The analyst can easily assess the accuracy of this estimate because sampling errors follow the principles of chance.
  - Sampling error easily measured.
- **Disadvantages**
  - necessitates a completely catalogued universe from which to draw the sample.
  - The size of the sample required to ensure statistical reliability is usually larger under simple random sampling than stratified sampling.
  - Units may be scattered and poorly accessible
  - Heterogeneous population
    → important minorities might not be taken into account

- **Note:** *let N = population size, n = sample size*.

- Suppose simple random sampling is used
  - We have N^n possible samples if sampling is **with replacement**.
  - We have $\binom{N}{n}$ possible samples if sampling is **without replacement**

Sampling….probability

- ***Restricted Random Sampling***

1. Systematic sampling

- **Principle**
  - Select sampling units at regular intervals (e.g. every k^{th} unit)
- **Procedure**
  - Arrange the units in some kind of sequence
  - Divide total population by the designated sample size (i.e $N/n=k$)
  - Choose a random starting point (for k, the starting point will be a random number between 1 and k)
  - Select units at regular intervals (in this case, every k^{th} unit)

Sampling….probability

Example

N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Sampling....probability

Systematic sampling....

- **Advantages**
  - Ensures representativeness across list when the population size is very large.
  - Its design is simple and convenient to adopt.
- **Disadvantages**
  - Periodicity-underlying pattern may be a problem (characteristics occurring at regular intervals)
  - if the population is ordered in a systematic way with respect to the characteristics the investigator is interested in, then it is possible that only certain types of items will be included in the population.

More complex sampling methods

Sampling….probability

NB. Sampling error is reduced by two factors:

1. Large sample size produces smaller error than do small samples

2. Homogeneous population produce smaller errors than heterogeneous population

- Stratified sampling is based on the second factor ensure samples are drawn from homogeneous population

- The choice of stratifying variable depends on the investigator (variables you want to represent accurately)

Sampling….probability

2. Stratified sampling

- When to use
  - Population with distinct subgroups i.e. When the population is made up of groups with different characteristics and.
  - It is expected big differences in the feature under study.

Procedure;

- Divide (stratify) sampling frame into homogeneous subgroups (**strata**) e.g. minorities, urban/rural areas, occupations.
- Draw random sample within each **stratum.**

Sampling....probability

Stratified sampling....

- The sampling method can vary from one stratum to another
  - ***Proportionate allocation-*** if the same sampling fraction is used for each stratum
  - ***Non-proportionate allocation-*** the strata unequal in size and a fixed number of units is selected from each stratum

Sampling....probability

Advantages

- representativeness of the sample is improved.
- focuses on important subpopulations and ignores irrelevant ones
- improves the accuracy of estimation

Disadvantages

- can be difficult to select relevant stratification variables
- not useful when there are no homogeneous subgroups
- can be expensive
- Sampling error is difficult to measure

Example: A sample of 50 students is to be drawn from a population consisting of 500 students belonging to two institutions A and B. The number of students in the institution A is 200 and the institution B is 300. How will you draw the sample using proportional allocation?

Solution:

There are two strata in this case with sizes

$N1 = 200$ and $N2 = 300$ and the total population

$N = N1 + N2 = 500$ The sample size is 50.

If n1 and n2 are the sample sizes,

$$n_1 = \frac{n}{N} \times N_1 = \frac{50}{500} \times 200 = 20$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{50}{500} \times 300 = 30$$

The sample sizes are 20 from A and 30 from B. Then the units from each institution are to be selected by simple random sampling.

Sampling….probability

3. Cluster sampling

- Reference population (homogeneous) is divided into clusters – often **geographical units**.
- There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, again by some suitable method, which may be the same as, or different from the method employed for the first stage units. Further stages may be added as required.

Sampling….probability

Cluster sampling….

- To reduce costs, researchers may choose a cluster sampling technique from other types.
- **Principle**
  - Whole population divided into groups e.g. neighbourhoods
  - A type of multi-stage sampling where all units at the lower level are included in the sample
  - Random sample taken of these groups (“clusters”)
  - Within selected clusters, all units e.g. households included (or random sample of these units)

Sampling....probability

- Involves selection of groups called clusters followed by selection of individuals within each selected cluster.
- Can be used when it is either impossible or impractical to compile exhaustive list of individuals of the target population.
- Cluster sampling is recommended for its efficiency, however accuracy is less because it is subject to more than one sampling error unlike SRS.

Sampling….probability

Cluster sampling….

- **Advantages**
  - Simple as complete list of sampling units within population not required
  - Less travel/resources required
  - It introduces flexibility in the sampling method, which is lacking in the other methods.
  - It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the fieldwork to be concentrated and yet large area to be covered.

- **Disadvantages**
  - Cluster members may be more alike than those in another cluster (homogeneous)
  - this “dependence” needs to be taken into account in the sample size **and** in the analysis (“design effect”)

B. Non Probability Sampling Method

- Non-random sampling is a process of sample selection without the use of randomization.

- In other words, a non-random sample is selected basis other than the probability consideration.

- The most important difference between random and non-random sampling is that the pattern of sampling variability can be ascertained in case of random sampling. whereas In non-random sampling, there is no way of knowing the patterns of variability in the process.

Advantage	Disadvantage
- Cheaper - Used when sampling frame is not available - widely dispersed popn that cluster sampling would not be efficient - in exploratory studies	- Inability to generalize

Sampling….non probability

Includes

- Judgmental /Purposive
- Quota
- Convenience / haphazard
- Snow ball
- Voluntary/self selection …etc criterion….

Sampling….non probability

1. Judgmental /Purposive

- Researcher choose based on their thinking of appropriate for the study.
- Used during limited number of people
- Appropriate when the study subjects are difficult to locate.
- Used where randomization is not expected
- Reduced cost and time

For example. if sample of ten students is to be selected from a class of sixty for analyzing the spending habits of students, the investigator would select 10 students who, in his opinion, are representative of the class.

Sampling….non probability

2. Quota

- The population is first segmented in to mutually exclusive sub-groups as in stratified sampling.
- Select subjects until a specific number of units/quota/ for various sub-groups has been filled.
- No rules for selecting the subjects
- This is one of the most common forms of non-probability sampling.

Sampling….non probability

3. Convenience / haphazard

- Selection of subjects based on easily availability & accessibility

Examples :People who just happen walking

- Often used in face to face interviews
- very easy to carry out,
- Difficult to draw any meaningful conclusion.
- May not be representative

Sampling….non probability

4. ***Snowball***

- Involves a process of “chain referrals”
- Suitable for locating key informants.
- You start with one or two key informants and ask them if they know persons who know a lot about your topic of interest.
- Used when trying to interview *hard to reach groups.*

Sampling….non probability

5. ***Volunteer/self selection***

- Subjects selected are volunteers who show interest to the study.
- Common in trials demanding long duration.
- Payments for subjects some times be involved.
- Introduces strong bias/self selection bias.

***Errors in sample survey*:**

- There are two types of errors

1) Sampling error:

- It is the discrepancy between the population value and sample value.
- May arise due to inappropriate sampling techniques applied.
- Sampling error can be minimized by increasing the size of sample (i.e. when $n \to N$, sampling error $\to 0$).

2) Non-sampling error: are errors due to procedure bias such as:

- Due to incorrect responses
- Measurement.
- Errors at different stages in processing the data.

- **The Needs for Sampling**

- Reduced cost
- Greater speed
- Greater accuracy
- Greater scope
- More detailed information can be obtained.

Sampling Distribution

Definition: The probability distribution of a statistic is called a **sampling distribution.**

For example, the probability distribution of $\bar{X}$ is called the **sampling distribution of the mean.**

The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.

Sampling Distribution of the sample mean

Given a variable X, if we arrange its values in ascending order and assign probability to each of the values or if we present X_i in a form of relative frequency distribution the result is called *Sampling Distribution of X.*

- It is a theoretical probability distribution that shows the functional relation ship between all possible values of a given sample mean based on samples of size *n*.
- There are commonly three properties of interest of a given sampling distribution
  - Its Mean
  - Its proportion
  - Its Variance & Functional form.

Steps for the construction of Sampling Distribution of the mean

1. From a finite population of size N ,list all possible samples of size n.
2. Calculate the mean for each sample.
3. Summarize the mean obtained in step 2 in terms of frequency distribution or relative frequency distribution.

Example: Suppose we have a population of size N=5 , consisting of the age of five desktop computer (in years): 6, 8, 10, 12, and 14. Take samples of size 2 with replacement and construct sampling distribution of the sample mean.

$$\Rightarrow Population mean = \mu = 10$$

$$population\ Variance = \sigma^2 = 8$$

Solution:

$N = 5, \quad n = 2$

➔ We have $N^n = 5^2 = 25$ possible samples since sampling is with replacement.

Step 1: Draw all possible samples:

	6	**8**	**10**	**12**	**14**
6	(6, 6)	(6, 8)	(6, 10)	(6, 12)	(6, 14)
8	(8,6)	(8,8)	(8,10)	(8,12)	(8,14)
10	(10,6)	(10,8)	(10,10)	(10,12)	(10,14)
12	(12,6)	(12,8)	(12,10)	(12,12)	(12,14)
14	(14,6)	(14,8)	(14,10)	(14,12)	(14,14)

Step 2: Calculate the mean for each sample:

	6	**8**	**10**	**12**	**14**
6	6	7	8	9	10
8	7	8	9	10	11
10	8	9	10	11	12
12	9	10	11	12	13
14	10	11	12	13	14

Step 3: Summarize the mean obtained in step 2 in terms of frequency distribution.

$\bar{x}$	$Frequency$
6	1
7	2
8	3
9	4
10	5
11	4
12	3
13	2
14	1

- Find the mean of $\overline{X}$, say $\mu_{\overline{X}}$

$$\mu_{\overline{X}} = \frac{\sum \overline{X}_i f_i}{\sum f_i} = \frac{250}{25} = 10 = \mu$$

- Find the variance of $\overline{X}$, say $\sigma_{\overline{X}}{}^2$

$$\sigma_{\overline{X}}{}^2 = \frac{\sum (\overline{X}_i - \mu_{\overline{X}})^2 f_i}{\sum f_i} = \frac{100}{25} = 4 \neq \sigma^2$$

Remark:

- In general if sampling is with replacement

$$\sigma_{\bar{X}}^{2} = \frac{\sigma^{2}}{n}$$

- If sampling is with out replacement

$$\sigma_{\bar{X}}^{2} = \frac{\sigma^{2}}{n}\left(\frac{N-n}{N-1}\right)$$

- In any case the sample mean is unbiased estimator of the population mean.

$$\mu_{\bar{X}} = \mu \Rightarrow E(\bar{X}) = \mu$$

Example:

The standard deviation of measurements of a linear dimension of a mechanical part is 0.14 mm. What sample size is required if the standard error of the mean must be no more than (a) 0.04 mm, (b) 0.02 mm?

Answer: Since the dimension can be measured as many times as desired, the population size is effectively infinite. Then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(a) For $\sigma_{\bar{x}} = 0.04$ mm and $\sigma = 0.14$ mm,

$$\sqrt{n} = \frac{0.14}{0.04} = 3.50$$

$$n = 12.25$$

Then for $\sigma_{\bar{x}} \leq 0.04$ mm, the minimum sample size is 13.

(b) For $\sigma_{\bar{x}} = 0.02$ mm and $\sigma = 0.14$ mm,

$$\sqrt{n} = \frac{0.14}{0.02} = 7.00$$

$$n = 49$$

Then for $\sigma_{\bar{x}} \leq 0.02$ mm, the minimum sample size is 49.

- When sampling is from a normally distributed population, the distribution of $\bar{X}$ will possess the following property.

- The distribution of $\bar{X}$ will be normal
  - The mean of $\bar{X}$ is equal to the population mean , i.e. $\mu_{\bar{X}} = \mu$
    - The variance of $\bar{X}$ is equal to the population variance divided by the sample size, i.e. $\sigma_{\bar{X}}{}^2 = \dfrac{\sigma^2}{n}$

$$\Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

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 $\frac{\sigma}{\sqrt{n}}$. 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 here is that a new formula must be used for the *z* values. It is:

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

If a large number of samples of a given size are selected from a normally distributed population, or if a large number of samples of a given size that is greater than or equal to 30 are selected from a population that is not normally distributed, and the sample means are computed, then the distribution of sample means will look like the normal distribution.

Example:

A plant manufactures electric light bulbs with a burning life that is approximately normally distributed with a mean of 1200 hours and a standard deviation of 36 hours. Find the probability that a random sample of 16 bulbs will have a sample mean less than 1180 burning hours. ***(Exercise!)***

<u>Givens</u>: $\boldsymbol{population\ mean\ (\mu = 1200 hr.),}$

$\boldsymbol{population standard\ deviation (\sigma = 36 hr.),}$

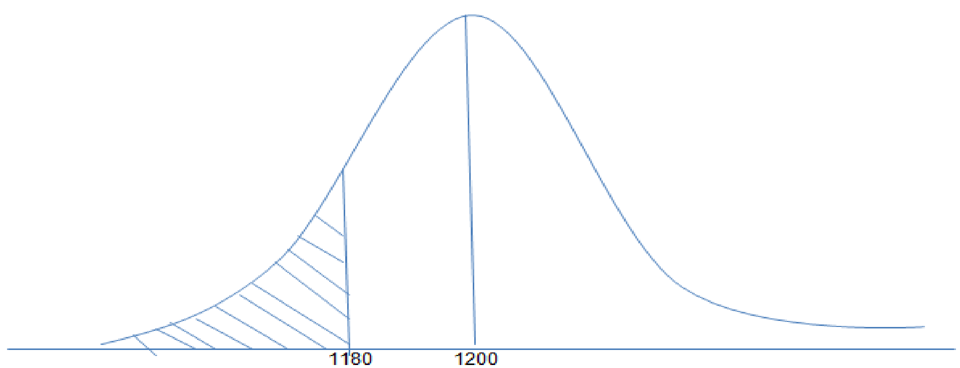
$\boldsymbol{and\ sample\ size\ n = 16}$

<u>Soln.</u>

Since the variable is approximately normally distributed, the distribution of sample means will be also approximately normal, with

- a mean $\mu_{\dot{x}} = \mu = 1200.$ and
- the standard error of the sample means $\sigma_{\dot{x}} = \frac{\sigma}{\sqrt{n}} = \frac{36}{\sqrt{16}} = \frac{36}{4} = 9$

Step 1: Draw a normal curve and shade the desired area.

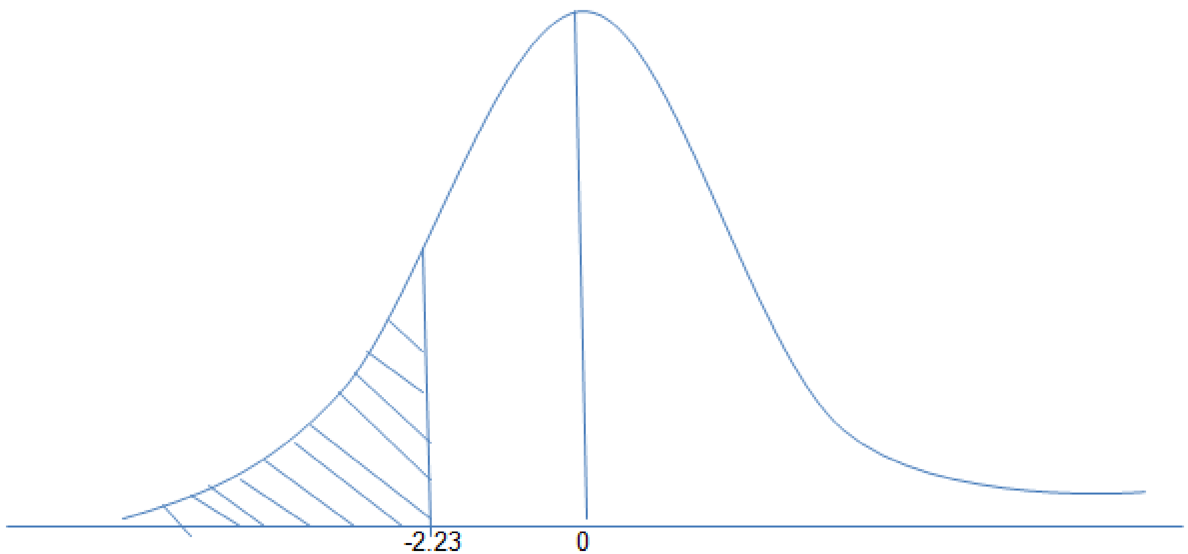


P($\bar{x}$<1180hr.)

Step 2: Convert the value to a z value. The z value is

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}} = \frac{1180 - 1200}{36/\sqrt{16}} = -2.23$$

$\Rightarrow P(\bar{x} < 1180hr) means\, P(Z < -2.23)$ which is graphically shown below. o, here we can use the standard normal distribution table to compute the value of the required probability.



Step 3: Find the corresponding area for the *z* value. The area to the left of -2.23 is = 0.0129 or 1.29%

Step 4: Conclusion

One can conclude that the probability that a random sample of 16 bulbs will have a sample mean less than 1180 burning hours. is 1.29% [that is, $P(\bar{x} < 1180\text{hr.}) = 0.0129$].

THANK YOU!