



BITS Pilani
Pilani Campus

Sampling and Estimation

Akanksha Bharadwaj
Asst. Professor, BITS Pilani



BITS Pilani
Pilani Campus



SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS Contact Session 5



BITS Pilani
Pilani Campus



Quick Review of last session

Exercise



The U.S. Environmental Protection Agency publishes figures on solid waste generation in the United States. One year, the average number of waste generated per person per day was 3.58 pounds. Suppose the daily amount of waste generated per person is normally distributed, with a standard deviation of 1.04 pounds. Of the daily amounts of waste generated per person, 67.72% would be greater than what amount?

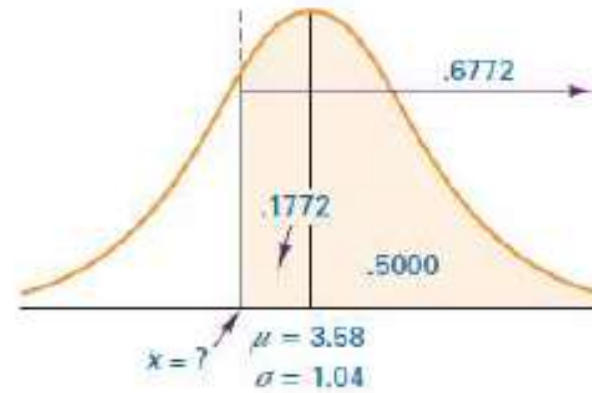
Solution



$$\text{area} = 0.1772$$

$$\downarrow$$
$$z = 0.46$$

left side of curve
z will be -ve.



$$-0.46 = \frac{x - 3.58}{1.04}$$

$$\underline{\underline{x = 3.10}}$$

Approximate Binomial Distribution Problems



- As sample sizes become large, binomial distributions approach the normal distribution in shape regardless of the value of p .
- This phenomenon occurs faster (for smaller values of n) when p is near .50.
- To work a binomial problem by the normal curve requires a translation process.
- The first part of this process is to convert the two parameters of a binomial distribution, n and p , to the two parameters of the normal distribution, μ and σ .

$$\mu = n \cdot p \text{ and } \sigma = \sqrt{n \cdot p \cdot q}.$$

continued



- After completion of this, a test must be made to determine whether the normal distribution is a good enough approximation of the binomial distribution:
Does the interval $\mu \pm 3\sigma$ lie between 0 and n ?
- For a normal curve approximation of a binomial distribution problem to be acceptable, all possible x values should be between 0 and n , which are the lower and upper limits, respectively, of a binomial distribution.
- Another rule of thumb for determining when to use the normal curve to approximate a binomial problem is that the approximation is good enough if both $np > 5$ and $nq > 5$.

Continuity correction factor



- It is used when you use a continuous probability distribution to approximate a discrete probability distribution. For example, when you want to use the normal to approximate a binomial.
- When you use a normal distribution to approximate a binomial distribution, you're going to have to use a continuity correction factor. **It's as simple as adding or subtracting .5 to the discrete x-value:** use the following table to decide whether to add or subtract.

If $P(X=n)$ use $P(n - 0.5 < X < n + 0.5)$

If $P(X > n)$ use $P(X > n + 0.5)$

If $P(X \leq n)$ use $P(X < n + 0.5)$

If $P(X < n)$ use $P(X < n - 0.5)$

If $P(X \geq n)$ use $P(X > n - 0.5)$

Example

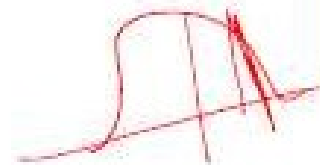


Work the following binomial distribution problem by using the normal distribution.

$$P(x = 12 | n = 25 \text{ and } p = .40) = ?$$

Solution

$$n=25 \quad p=0.4$$



$$\checkmark \mu = np = 25 \times 0.4 = 10$$

$$\checkmark \sigma = \sqrt{npq} = \sqrt{25 \times 0.4 \times 0.6} = 2.45$$

[this $np > 5$
so we can use
normal approx.]

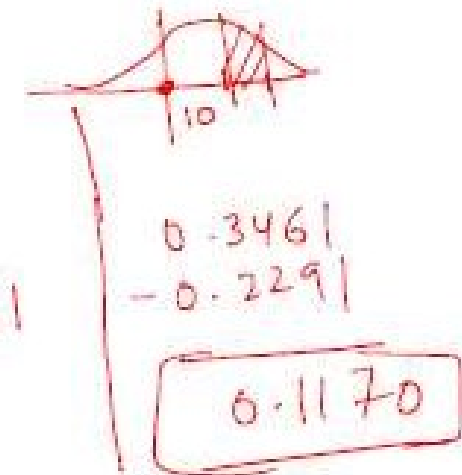
$$q = 1 - p$$

$\mu \pm 3\sigma = 10 \pm (3 \times 2.45)$
 \Rightarrow range is 2.65 to 17.35
 this lies 0 to 25? (0 to 25)
 yes

$$P(X=12) = P(11.5 < X < 12.5)$$

for 11.5
 $z = \frac{11.5 - 10}{2.45} = 0.61$
 \downarrow
 prob
 0.2291

for 12.5
 $z = \frac{12.5 - 10}{2.45} = 1.02$
 prob
 0.3461



Exercise (HW)

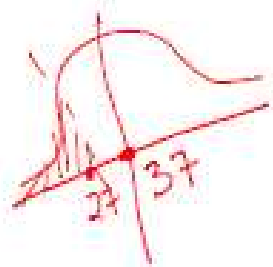


Solve the following binomial distribution problem by using the normal distribution

$$P(x < 27 | n = 100 \text{ and } p = .37) = ?$$

Solution

Homework



$$P(X < 27 | n = 100 \text{ \& } p = 0.37) = ?$$

$$\mu = np = 37, \sigma = 4.83$$

$\therefore np > 5$ we can use normal distribution for this

$$\mu \pm 3\sigma = 37 \pm 14.49$$

So, range is 22.51 to 51.49

\therefore this lies b/w 0 to n i.e. 0 to 100

we can use normal distribution approximation.

$$P(X < 27) = P(X < 26.5) \quad [\text{Based on continuity correction factor}]$$

$$Z = \frac{26.5 - 37}{4.83} = -2.17$$

prob using z-table is 0.4850

$$\text{Ans} = 0.5 - 0.4850 = 0.0150$$

continued



Had this problem been solved by using the binomial formula, the probabilities would have been the following.

x Value	Probability
26	.0059
25	.0035
24	.0019
23	.0010
22	.0005
21	.0002
20	.0001
$x < 27$.0131

The answer obtained by using the normal curve approximation (.0150) compares favorably to this exact binomial answer. The difference is only .0019.

Sampling



- Sampling is widely used in business as a means of gathering useful information about a population.
- Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process
- A sample provides a reasonable means for gathering useful decision-making information that might be otherwise unattainable and unaffordable.



Reasons for Sampling

- Taking a sample instead of conducting a census offers several advantages
 1. The sample can save money.
 2. The sample can save time.
 3. For given resources, the sample can broaden the scope of the study.
 4. If accessing the population is impossible, the sample is the only option.

Random Versus Non-random Sampling



- In **random** sampling every unit of the population has the same probability of being selected into the sample.
- In **non-random** sampling not every unit of the population has the same probability of being selected into the sample.

innovate

achieve

lead

BITS Pilani



Random Sampling



Simple random sampling

- With simple random sampling, each unit of the frame is numbered from 1 to N (where N is the size of the population).
- Next, a table of random numbers or a random number generator is used to select n items into the sample.
- A random number generator is usually a computer program that allows computer-calculated output to yield random numbers.

Stratified Random Sampling



- In this the population is divided into nonoverlapping **subpopulations** called **strata**.
- The researcher then extracts a random sample from each of the subpopulations.
- The main reason for using stratified random sampling is that it has the potential for reducing sampling error.
- With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups.

Systematic Random Sampling

- With systematic sampling, every k th item is selected to produce a sample of size n from a population of size N .
- The value of k , sometimes called the **sampling cycle**, can be determined by the following formula.

$$k = \frac{N}{n}$$

where

n = sample size

N = population size

k = size of interval for selection



innovate

achieve

lead

BITS Pilani



Non-random Sampling



Convenience Sampling

- In convenience sampling, elements for the sample are selected for the convenience of the researcher.
- The researcher typically chooses elements that are readily available, nearby, or willing to participate
- The sample tends to be less variable than the population because in many environments the extreme elements of the population are not readily available.
- The researcher will select more elements from the middle of the population.



Quota Sampling

- It appears to be similar to stratified random sampling. Certain population subclasses, such as age group, gender, or geographic region, are used as strata.
- However, instead of randomly sampling from each stratum, the researcher uses a nonrandom sampling method to gather data from one stratum until the desired quota of samples is filled.

Snowball Sampling



- Another nonrandom sampling technique is **snowball sampling**, in which *survey subjects are selected based on **referral from other survey respondents***.
- The researcher identifies a person who fits the profile of subjects wanted for the study.
- The researcher then asks this person for the names and locations of others who would also fit the profile of subjects wanted for the study.
- Through these referrals, survey subjects can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate.

Sampling Error



- **Sampling error** occurs *when the sample is not representative of the population.*
- When random sampling techniques are used to select elements for the sample, sampling error occurs by chance.

innovate

achieve

lead

BITS Pilani



Sampling Variation

Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490



Select different samples of varied sizes

Sample 1

3000 2486 820 1678 2070 2638 2490 1865 1000 2090 596 3200

Sample 2

2840 2858 3000 2490 2998 3050 2070 2896 3200 2490 3280

Sample 3

2858 3240 2497 2865 656 2093 934 1861 868 795

Sample 4

2086 1000 2497 596 656 875 2085 934 1313

Sample 5

820 1313 3000 2640 596 2640 2600 2495 934 2500



Select different samples of varied sizes

Sample 6

2840 2499 1327 1861 2495 3024 3038 2497

Sample 7

2858 2490 868 1670 1480 2643 1480 1680 2085 2490

Sample 8

2495 2858 1861 2092 2499 3000 2660 1000 1679 926 2660

Sample 9

795 791 3200 2085 2638 2497 2486 1159 2640

Sample 10

3019 3240 3200 3050 3000 3015 2900 2896 2998



Compute sample mean of these samples

Sample No.	Sample size	Mean	SD
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
Overall	100	2162.24	732.26

Sampling Variability



- The term "sampling variability" refers to the fact that the statistical information from a sample (called a *statistic*) will vary as the random sampling is repeated.
- **Sampling variability will decrease as the sample size increases.**
- the samples must be randomly chosen, must be of the same size (not smaller than 30), and the more samples that are used, the more reliable the information gathered will be.

innovate

achieve

lead

BITS Pilani



Sampling Distribution

Do you consider these sample means and sample SDs as variable?

If yes, should we not describe the distribution of these variables?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

Definition

The probability distribution of a statistic (sample estimate) is called sampling distribution.

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 \bar{x}

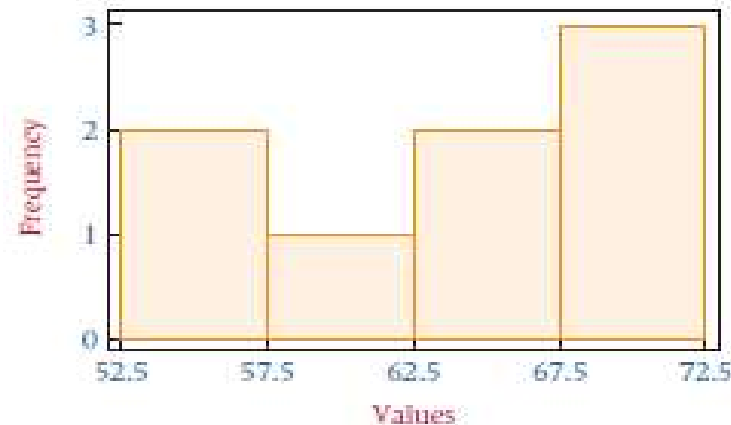


- The sample mean is one of the more common statistics used in the inferential process.
- The **distribution** of the values of the sample mean (\bar{x}) in repeated **samples** is called the **sampling distribution of \bar{x}**
- One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Example



- Suppose a small finite population consists of only $N = 8$ numbers:
54 55 59 63 64 68 69 70
- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



- Suppose we take all possible samples of size $n = 2$ from this population with replacement.

Example



The result is the following pairs of data.

(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

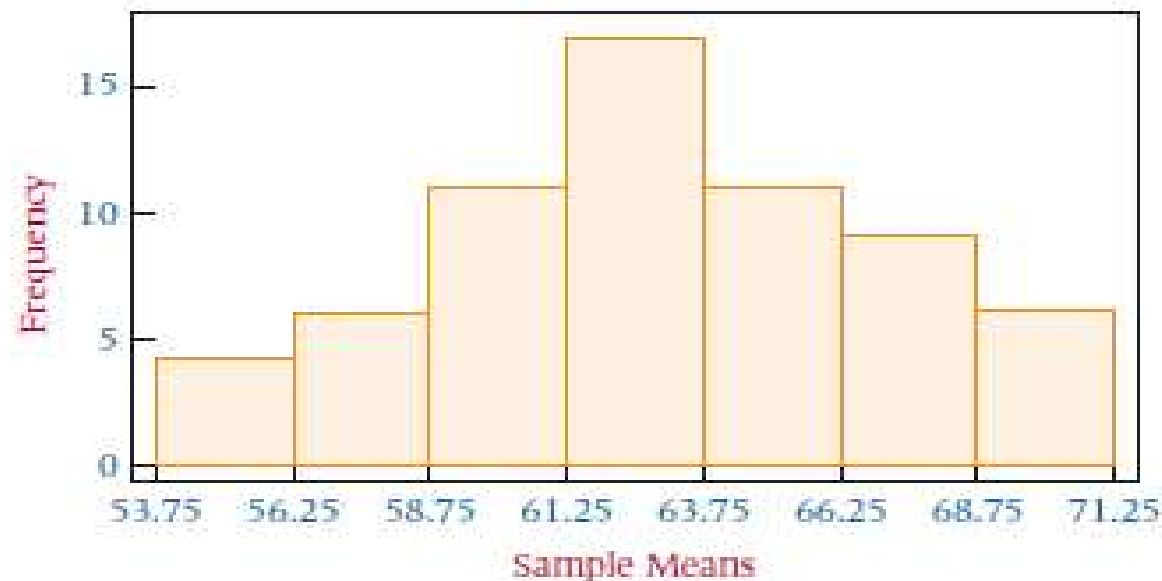
The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Example



- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.

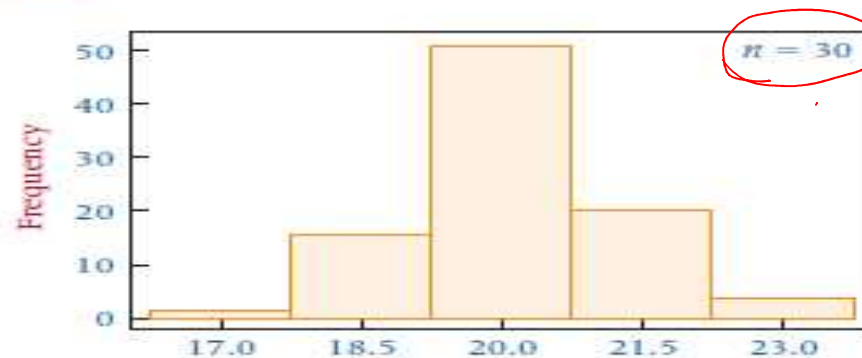
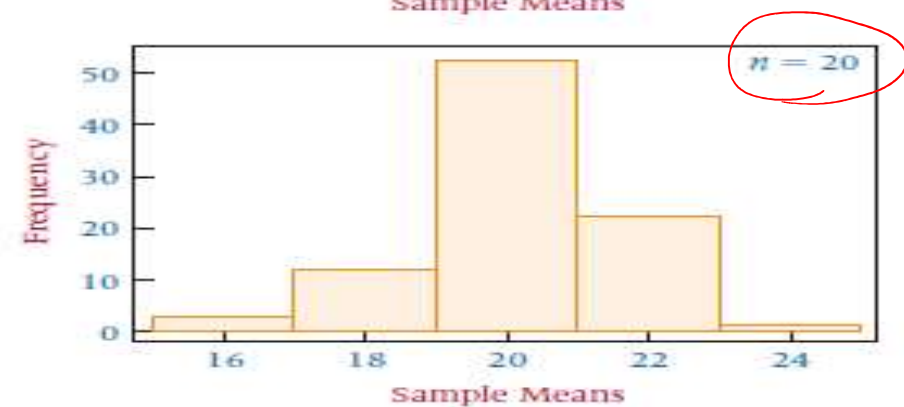
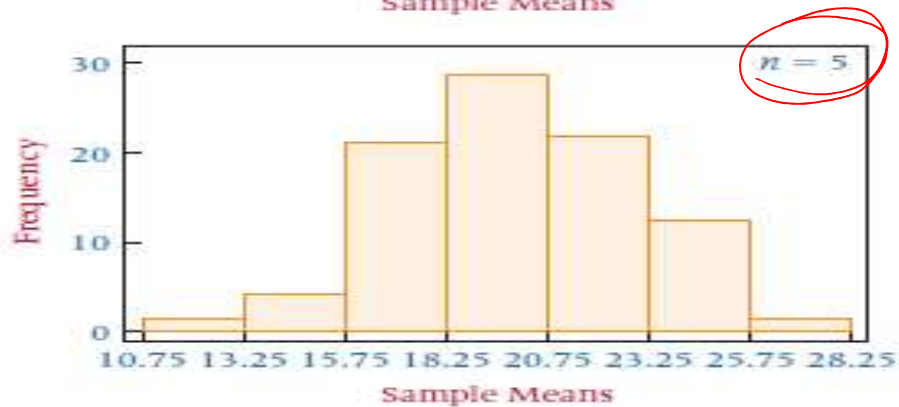
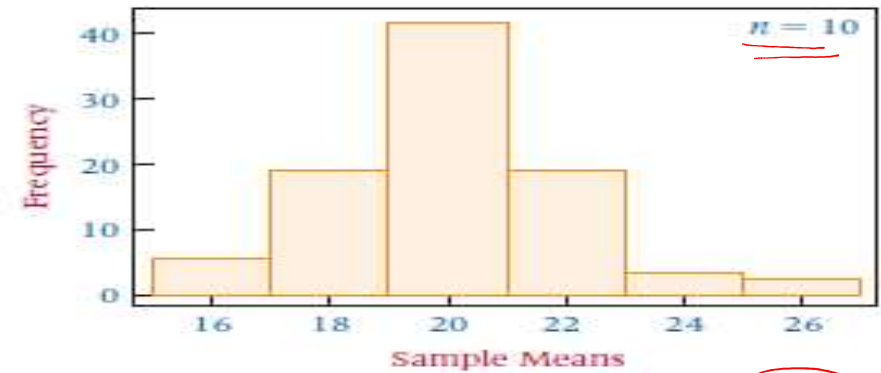
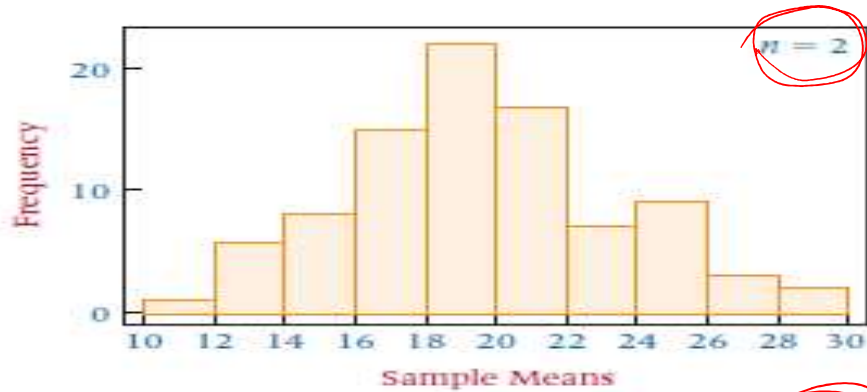


Conclusions

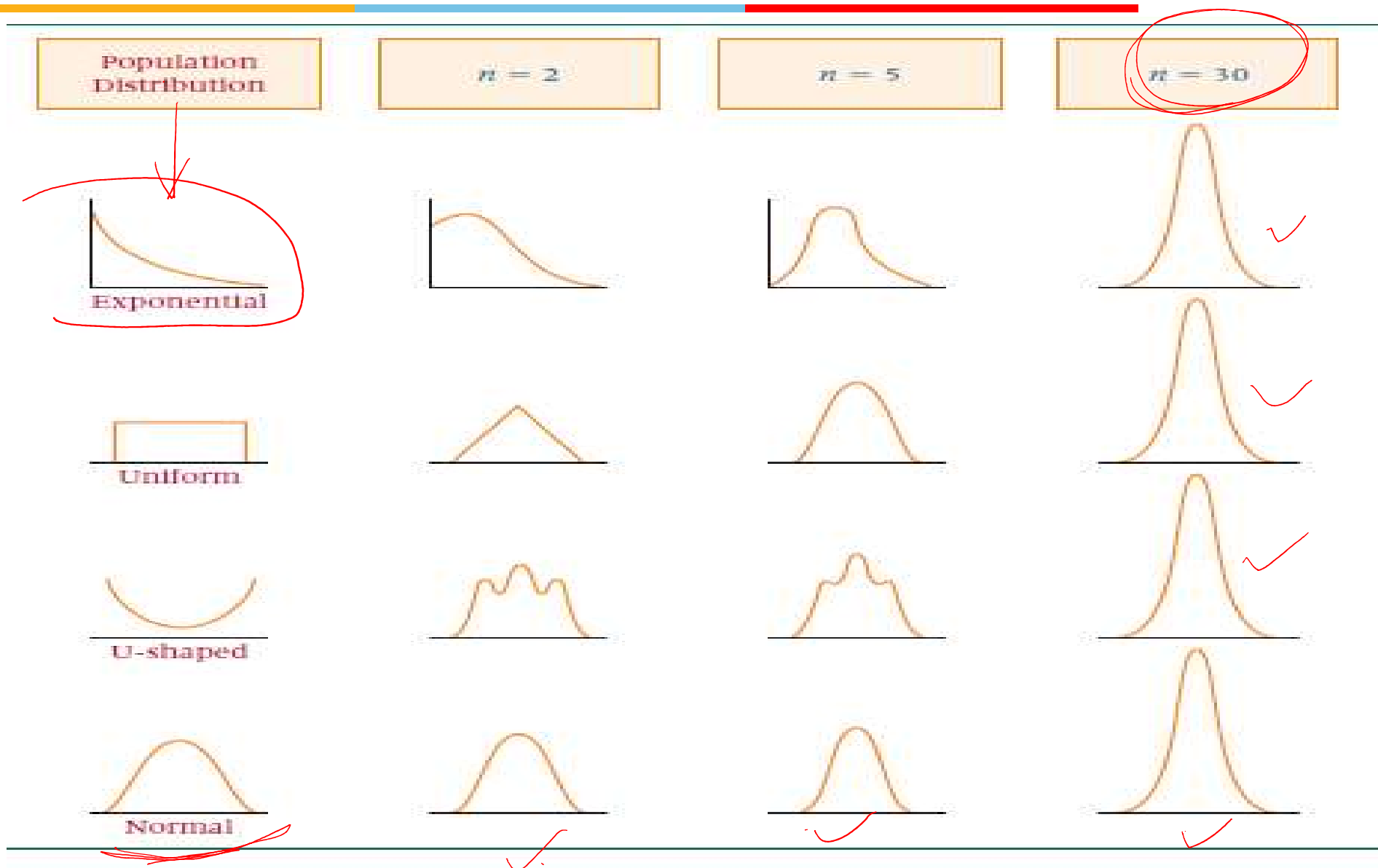


- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
- The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
- As sample sizes become much larger, the sample mean distributions begin to approach a normal distribution and the variation among the means decreases.

Sample Means from 90 Samples Ranging in Size from $n = 2$ to $n = 30$ from a Uniformly Distributed Population with $a = 10$ and $b = 30$



Shapes of the Distributions of Sample Means



Central Limit Theorem



- If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ , the sample means, \bar{x} , are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution.
- If the population is normally distributed, the sample means are normally distributed for any size sample.
- From mathematical expectation

$$\mu_{\bar{x}} = \mu$$

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

← sample size



Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using **z scores**
- The formula to determine z scores for individual values from a normal distribution:
$$z = \frac{x - \mu}{\sigma}$$
- If sample means are normally distributed, the z score formula applied to sample means would be
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$
- The standard deviation of the statistic of interest is $\sigma_{\bar{x}}$, sometimes referred to as the **standard error of the mean**.



Z score for sample means

- The researcher would randomly draw out all possible samples of the given size from the population, compute the sample means, and average them. This task is virtually impossible to accomplish in any realistic period of time.
- Similar activity for variation
- the mean of the sample means is the population mean.
- the standard deviation of the sample means is the standard deviation of the population divided by the square root of the sample
- Using central limit theorem:

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

Example

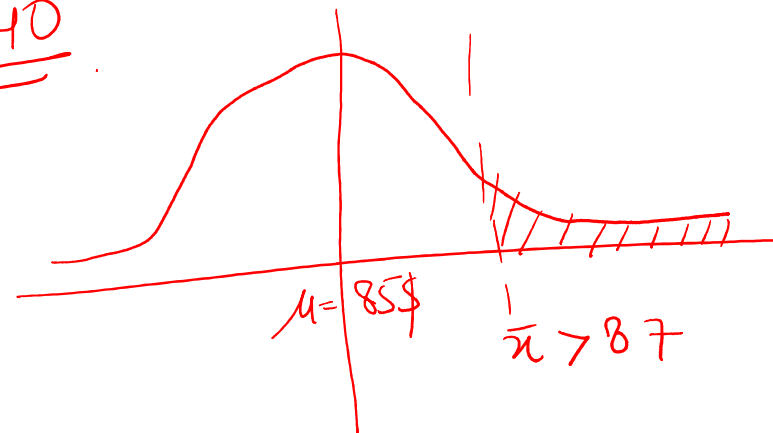


Suppose the mean expenditure per customer at a tire store is \$85.00, with a standard deviation of \$9.00. If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

$$P(\bar{x} \geq 87)$$

$$n = 40$$

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



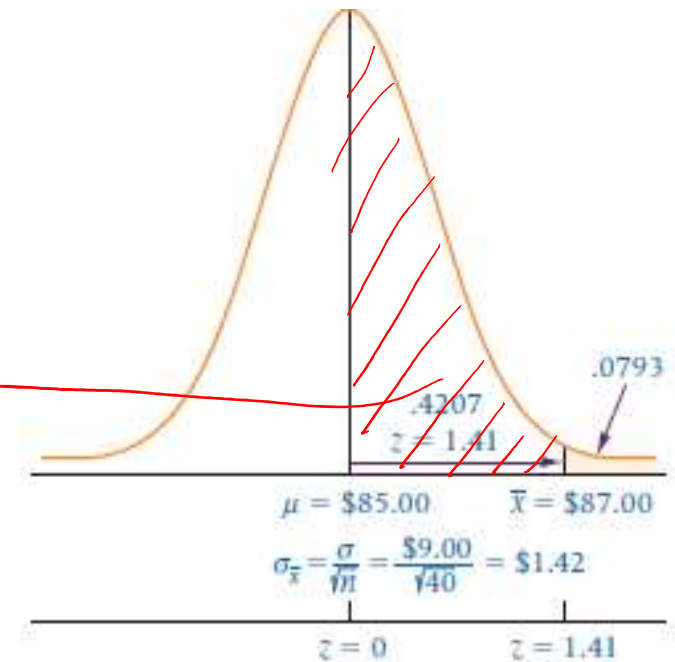
Solution



$$Z = \frac{87 - 85}{\frac{9}{\sqrt{40}}} = \frac{2}{1.42} \approx 1.41$$

area under curve for 1.41
= 0.4207

$$P(\bar{x} \geq 87) = 0.5 - 0.4207 \\ = \underline{\underline{0.0793}}$$



Exercise



- Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

$$P(441 \leq \bar{X} \leq 446) = ?$$

$\because n > 30$ we can say sample means are normally distributed

Solution

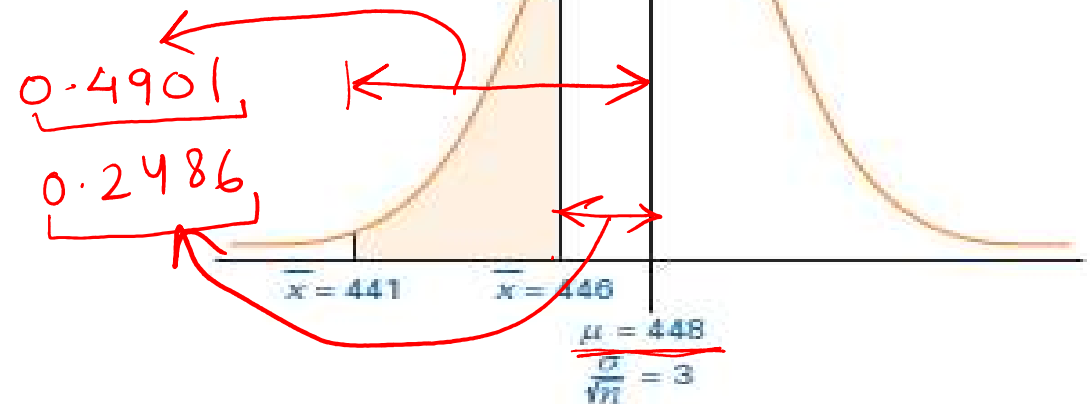


For this problem, $\mu = 448$, $\sigma = 21$, and $n = 49$. The problem is to determine $P(441 \leq \bar{x} \leq 446)$. The following diagram depicts the problem.

$$\text{for } 441, z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{441 - 448}{21/\sqrt{49}} = \frac{-7}{3} = -2.33$$

$$\text{for } 446, z = \frac{446 - 448}{21/\sqrt{49}} = \frac{-2}{3} = -0.67$$

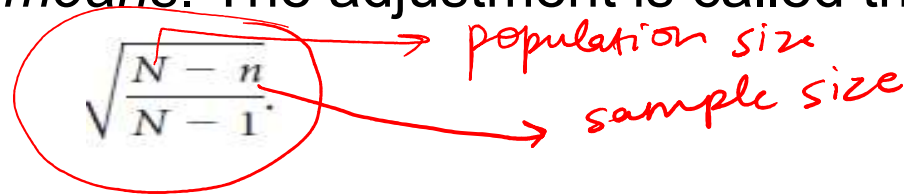
prob for $z = -2.33$ is
prob. for $z = -0.67$ is



$$\begin{aligned} \text{Ans} &= 0.4901 - 0.2486 \\ &= 0.2415 \end{aligned}$$

Sampling from a Finite Population

- The earlier example was based on the assumption that the population was infinitely or extremely large.
- In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor**
- Following is the z formula for sample means when samples are drawn from finite populations.



The diagram shows the finite correction factor formula: $\sqrt{\frac{N-n}{N-1}}$. A red circle is drawn around the entire fraction. A red arrow points from the handwritten text "population size" to the variable N in the numerator. Another red arrow points from the handwritten text "sample size" to the variable n in the numerator.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$



Rules for finite population

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.
- In theory, whenever researchers are working with a finite population, they can use the finite correction factor.
- A rough rule of thumb for many researchers is that, if the sample size is **less** than 5% of the finite population size or **$n/N < 0.05$** , the finite correction factor does **not** significantly modify the solution.

$$\sqrt{\frac{N-n}{N-1}}$$

Exercise



A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

$$n > 30$$

$$P(\bar{X} < 40)$$

$$\mu = 37.6$$

$$\sigma = 8.3$$

$$N = 350$$

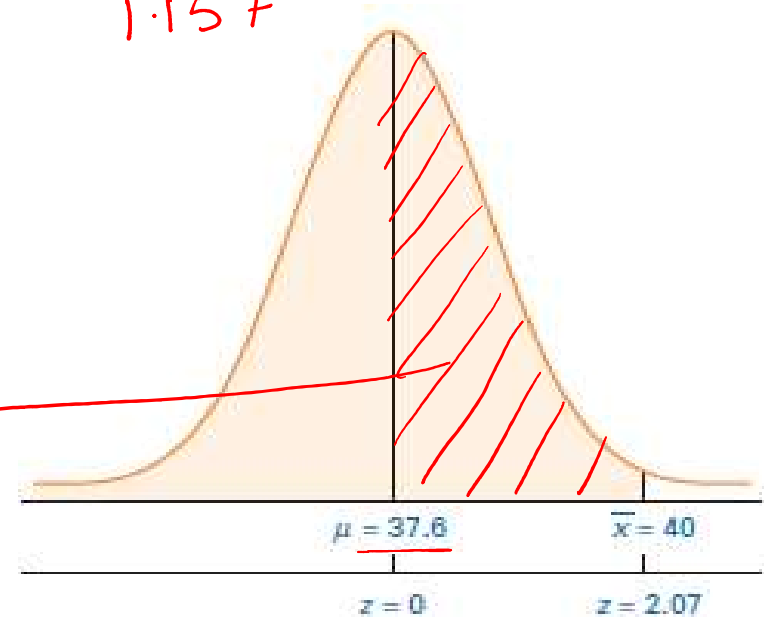
$$n = 45$$

Solution



$$\text{for } \bar{x} = 40, Z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}} \sqrt{\frac{350 - 45}{350 - 1}}} = \frac{2.4}{1.157} = 2.07$$

$$\begin{aligned} \text{area under curve for } 2.07 \\ = 0.4808 \end{aligned}$$



$$\begin{aligned} \text{Ans} \rightarrow & 0.5000 \\ & + 0.4808 \\ \hline & 0.9808 \end{aligned}$$

Sampling Distribution Of Sample Proportion



- If research results in **countable** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

SAMPLE PROPORTION

$$\hat{p} = \frac{x}{n}$$

where

x = number of items in a sample that have the characteristic

n = number of items in the sample

Example



- In a sample of 100 factory workers, 30 workers might belong to a union.
- The value of sample proportion for this characteristic, union membership, is

$$30/100 = .30$$

How does a researcher use the sample proportion in analysis?



- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions
- If $n \cdot p > 5$ and $n \cdot q > 5$ (p is the population proportion and $q = 1 - p$).
- The mean of sample proportions for all samples of size n randomly drawn from a population is p (the population proportion) and the **standard deviation of sample proportions** is $\sqrt{\frac{p \cdot q}{n}}$
- sometimes referred to as the **standard error of the proportion**

Z Formula For Sample Proportions



For $n \cdot p > 5$ and $n \cdot q > 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

\hat{p} = sample proportion

n = sample size

p = population proportion

$q = 1 - p$

Example



- Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire?

$$P(\hat{p} \leq 0.5)$$

$$np = 120 \times 0.6 = 72.0$$

$$nq = 120 \times 0.4 = 48$$

$$np > 5 \text{ \& } nq > 5$$

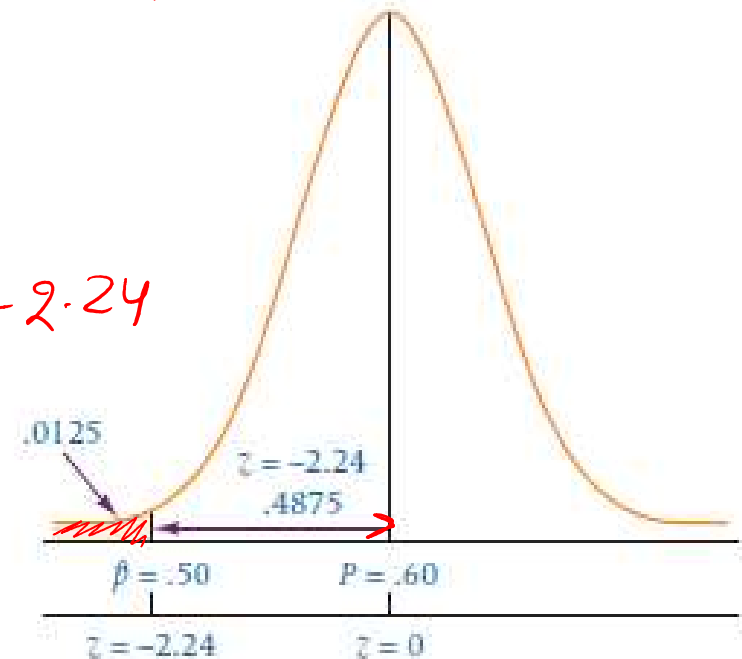
Solution

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$



for 0.5, $Z = \frac{0.5 - 0.6}{\sqrt{\frac{0.6 \times 0.4}{120}}} = -2.24$

area under the curve/prob. of $Z = -2.24$
0.4875



$$P(\hat{p} \leq 0.5) = 0.5 - 0.4875 = 0.0125$$

Exercise



If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

$$p = 0.1$$

$$n = 80$$

$$\hat{p} = ?$$

$$\frac{12}{80} = 0.15$$

$$P(\hat{p} \geq 0.15)$$

Solution

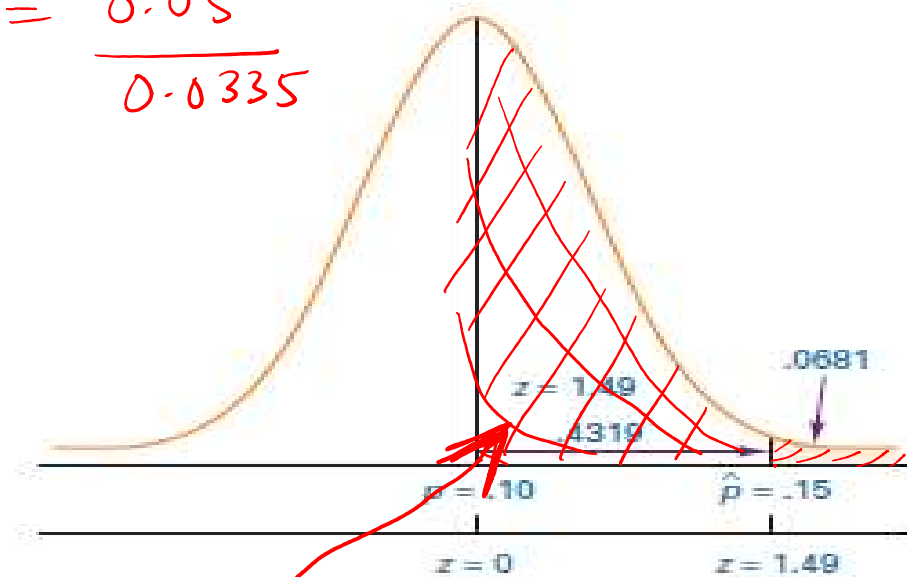


for $\hat{p} = 0.15$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.15 - 0.10}{\sqrt{\frac{0.1 \times 0.9}{80}}} = \frac{0.05}{0.0335}$$

$$= 1.49$$

area under curve for 1.49
0.4319



$$\begin{aligned} \text{Ans} &= 0.5 - 0.4319 \\ &= 0.0681 \end{aligned}$$