Estimation

S C Agarkar

VN BRIMS, Thane

Preamble

- Estimation is an everyday phenomenon as we have to estimate many things in day to day activities. For example, time taken for the journey between home and college, time required to cross the road, etc.
- Managers need to estimate quite often. The outcomes of their estimates could be serious. For example, sell of houses, people registering for the tour, loan takers from the bank, etc.
- Statistical inference is the branch in statistics concerned with using probability concepts to deal with uncertainty in decision making.

Types of Estimates

- We can make two types of estimates: Point estimate and interval estimate.
- Point estimate is a single number that is used to estimate an unknown population parameter. For example, department head estimates that there would be 120 students next year for the MMS course.
- An interval estimate is a range of values used to estimate a population parameter. For example, the department head estimates that the enrolment of students next year would be between 100 to120.

Estimator and estimates

- An estimator is a sample statistic used to estimate a population parameter. For example, a sample mean (x) can be an estimator of the population mean (μ). Similarly, a sample proportion can be an estimator of population proportion.
- An estimate is a specific observed value of a statistic. For example, suppose we calculate the mean odometer reading of a sample of used taxis as 98,000 kilometers. Using this value we can estimate that mileage of all the fleet of used taxis would be 98,000 kilometers.

Tabular summary

Population	Parameter	Sample Statistic	Estimate
Employees in a furniture factory	Mean turnover per year	Mean turnover for period of one month	8.9% turnover per year
Applicants for town manager	Mean formal education years	Mean formal education of every fifth applicant	17.9 years of formal education
Teenagers in a given community	Proportions whohave criminal records	Proportion of a sample of 50 teenagers who have criminal record	2 % have criminalrecords

Point Estimates of mean

- The sample mean is the best estimator of population mean. It is unbiased, consistent, and the most efficient estimator.
- In a medical supply company disposable hypodermic syringes are wrapped in a sterile package and then jumble packed in a large corrugated carton. Jumble packing causes the carton to contain differing number of syringes. As the syringes are sold on per unit basis the company needs to estimate number of syringes per carton.
- 35 Cartons are chosen at random and the number of syringes in each were recorded. The sample mean was then calculated which was 102. This calculation helped the manufacturer to estimate the number of syringes in each carton and price them accordingly.

Point Estimate of SD

- Sometimes, the variance or standard deviation of a population is estimated. For this purpose too variance or standard deviation of the sample is calculated. From this information, these values for the entire population is estimated.
- Suppose the management of a medical company wants to estimate the variance and standard deviation of the distribution of number of packaged syringes per carton. Using the data from the selected sample of 35 Cartons one can calculate these values.

Point Estimate of proportion

- The proportion of units that have a particular characteristics in a given population is symbolized as p. If we know the proportion of units in a sample that has same characteristics then we can use it as an estimator of p.
- Suppose the management wishes to estimate the number of carton of syringes that will arrive damaged. They can check a sample of 50 cartons and find the value say 0.8. This value can be used to estimate the proportion of damaged cartons in the entire population (0.8).

Problems to solve

- A meteorologist for a local television station would like to report average rainfall for that day in the evening newscast. The following are the rainfall measurements (in inches) for that date for 16 randomly chosen past years, Determine the sample mean rainfall and help the meteorologist to estimate the rainfall for that day.
- 0.470.270.130.540.000.080.750.060.001.050.340.260.170.420.500.86
- In a sample of 400 textile workers, 184 expressed extreme dissatisfaction regarding a prospective plan to modify working conditions. Because this dissatisfaction was strong enough to allow management to interpret plan reaction as being highly negative, they were curious about the proportion of total workers harbouring this sentiment. Give a point estimate.

Ans: 0.46

Interval Estimate

- An interval estimate describes a range of values within which a population parameter is likely to lie.
- A marketing research director wants to know the average life of a car battery made by his factory. He selects a random sample of 200 batteries and finds that a mean life is 36 months. Then he concludes that this must be true for the entire population.
- SE in this case would be $\sigma \div \sqrt{n} = 0.707$ months assuming that the SD of the population is 10. Thus actual mean life of batteries may lie between 35.292 to 36.77 months (i.e. 35 to 37 months).

Problems to solve

- For a population with known variance of 185, a sample of 64 individuals leads to 217 as an estimate of the mean.

a) Find the standard error of the mean

b) Establish an interval estimate that should include thepopulation mean 68.3 percent.

Ans: a) 1.70 b) 215.3 $ 218.7

- From a population known to have a standard deviation of 1.4, a sample of 60indivudualsis taken. The mean for this sample is found to be 6.2.

a) Find the standard error of the mean

b) Establish an interval estimate around the sample mean using one standard error of the mean.

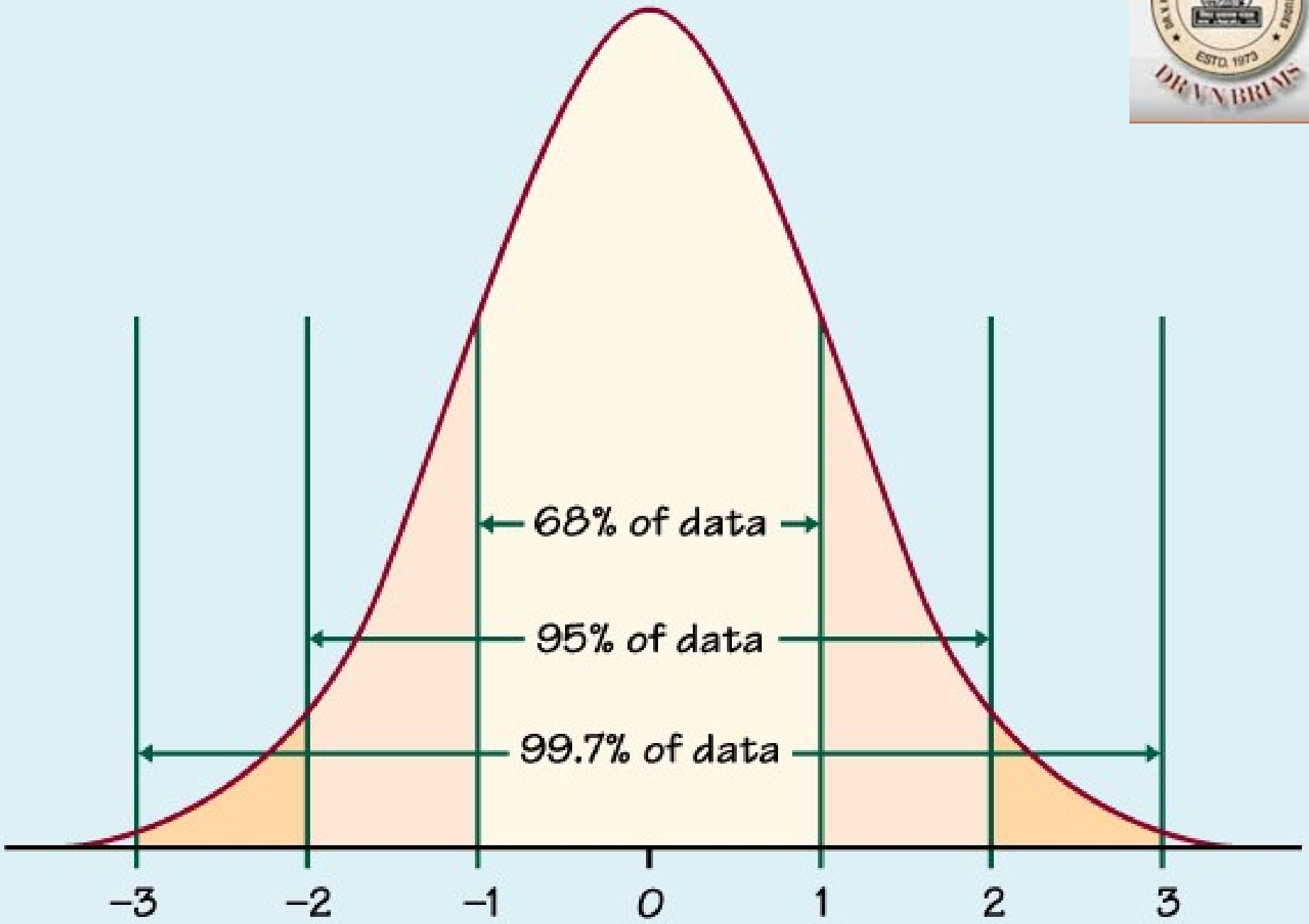
Ans: a) 0.181 b) 6.019 $ 6.381

Characteristics of normal curve

- The curve has a single peak, it is unimodal.
- The mean of the population lies at the centre.
- The mean, median and mode all coincide.
- Two tails extends indefinitely and never touch the horizontal axis.
- To define a particular normal probability distribution we need only two parameters, the mean and the standard deviation.

Areas under normal curve

- Approximately 68 percent of all values in a normally distributed population lie within 1 SD (plus or minus) from the mean.
- Approximately 95.5 percent of all the values in a normally distributed population lie within 2 SD (plus or minus) from the mean.
- Approximately 99.7 percent of all the values in a normally distributed population lie within 3 SD (plus or minus) from the mean.
- Every time it is not necessary to look at the area under the normal curve. We can refer to statistical tables constructed showing the portions of the area covered.


68% of data
95% of data
99.7% of data
−3
−2
−1
0
1
2
3

Distances under the normal curve

- The value of z is derived from the formula $z = (x-\mu)/\sigma$
- x = value of random variable with which we are concerned
- μ= mean of the distribution of this random variable
- σ= standard deviation of the distribution
- z= number of standard deviations from x to the mean of this distribution.
- Normally distributed random variable take on many different units (Rs, meter, degrees, etc). To avoid these units we talk in terms of unitless entity called z.

Interval estimates & confidence interva

- According to the z value table plus and minus 1.64 SE includes about 90 percent of the area under the curve (it includes .4495 of the area on either side of the mean in a normal distribution). Similarly, plus and minus 2.58 SE includes about 99 percent of the area or 49.5 percent on each side of the mean.
- In statistics the probability that we associate with interval level estimate is called the confidence level. The confidence interval is the range of estimate we are making.
- Upper limit= x+1.64SE and lower limit= x-1.64SE for confidence level of 90 percent.

Confidence level & confidence interval

- A high confidence level signify a high degree of accuracy in estimate. In practice, however, high confidence levels produce large confidence intervals and such large intervals are not precise.
- Confidence level of delivering an item in a short time is low. Instead it is high if the time interval is large. Similarly, the confidence level of completing a task quickly is short while it is large if the time duration is more.

Sampling & confidence interval

- In a battery example discussed earlier if we say “we are 95 percent confident that the mean battery life of the population lies between 30 and 42 months.
- This statement does not mean that the chance is .95 that the mean life falls between the limits. Instead, it means that if we select many random samples of this sample size and if we calculate a confidence interval for each of these samples, then in about 95 percent of these cases, the population mean would lie within that interval.

Interval estimates of the mean

- Standard error is given by SE=σ/√n
- The wholesaler of wiper blades have selected a sample of 100 blades and calculated sample mean as 21 months and SD of the population as 6 months.
- Hence SE=6/√100 = 0.6 months
- At the 95 percent confidence level will include 47.5 percent of the area on either side of the mean of the sampling distribution. We find that .475 of the area under the normal curve is contained between the mean and a point 1.96 SE.
- Upper limit= x+1.96 SE= 22.18 months
- Lower limit= x-1.96 SE= 19.82 months
- We can report that we estimate the mean life of the population of wiper blades to be between 19.82 and 22.18 months with 95 percent confidence.

When the population SD is unknown

- A social service agency is interested in estimating the mean annual income of 700 families in a locality. By taking a sample of 50 we get sample mean $4800 and SD $950.
- SE =σ/√n *√N-n/N-1
- Substituting the values we get SE= $129.57
- At 90 percent confidence level
- Upper confidence limit: x+1.64SE= $5,012.50
- Lower confidence limit: x-1.64SE= $4,587.50
- We report that with 90 percent confidence we estimate the average annual income of all 700 families living in the locality falls between $4,587.50 and $5,012.50.

Interval estimates of the proportion

- SE of proportion=√pq/n
- n= number of trials, p= probability of success and
- q = probability of a failure (1-p).
- An industry wants to estimate what proportion of employees prefer to provide their own retirement benefit in lieu of company sponsored plan.
- Taking a random sample of 75 employees they have found out that .4 of them are interested in providing their own benefit. Thus n=75, p= .4 and q= 1-.4=.6.

Continued

- SE= =√pq/n = .057
- A 99 percent confidence level would include 49.5 percent of the area on either side of the mean in the area under the normal curve. This refers to 2.58 standard error from the mean.
- Upper confidence limit = p+2.58 SE= .547
- Lower confidence limit = p-2.58 SE = .253
- Thus, we estimate from the sample of 75 employees that with 99 percent confidence we believe that the proportion of the total population of employees who wish to establish their own retirement plans lies between .253 and .547.

Interval estimate using t distribution

- When the sample size is small t distribution is used .
- Early theoretical work on t distribution was done by W. S. Gossett in the early part of 1900. Due to the restrictions from his company (Guinnes Brewery in Dublin) he published his work under the pen name 'Student'. Hence it is called student's t distribution. It is used when the sample size is less than 30.
- Apart from smallness of the sample t distribution is used when population standard deviation is not known.

Characteristics of t distribution

- In general the t distribution is flatter than normal distribution.
- A t distribution is lower at the mean and higher at tails than normal distribution.
- Hence t distribution has proportionally more of its area in its tails than normal distribution.
- It is necessary to go farther out from mean of a t distribution to include the same area under the normal curve.

Degrees of freedom

- We can define degrees of freedom as the number of values we can choose freely.
- If we are dealing with two numbers whose sum is 10. Once we choose the values of one number the value of another number is decided. Hence degree of freedom in this case is 1,i. e. (n-1).
- If we are dealing with seven numbers having an average of 16 we can specify six variables. Thus, degree of freedom is 6 i.e. (n-1).

Family of Normal Curves

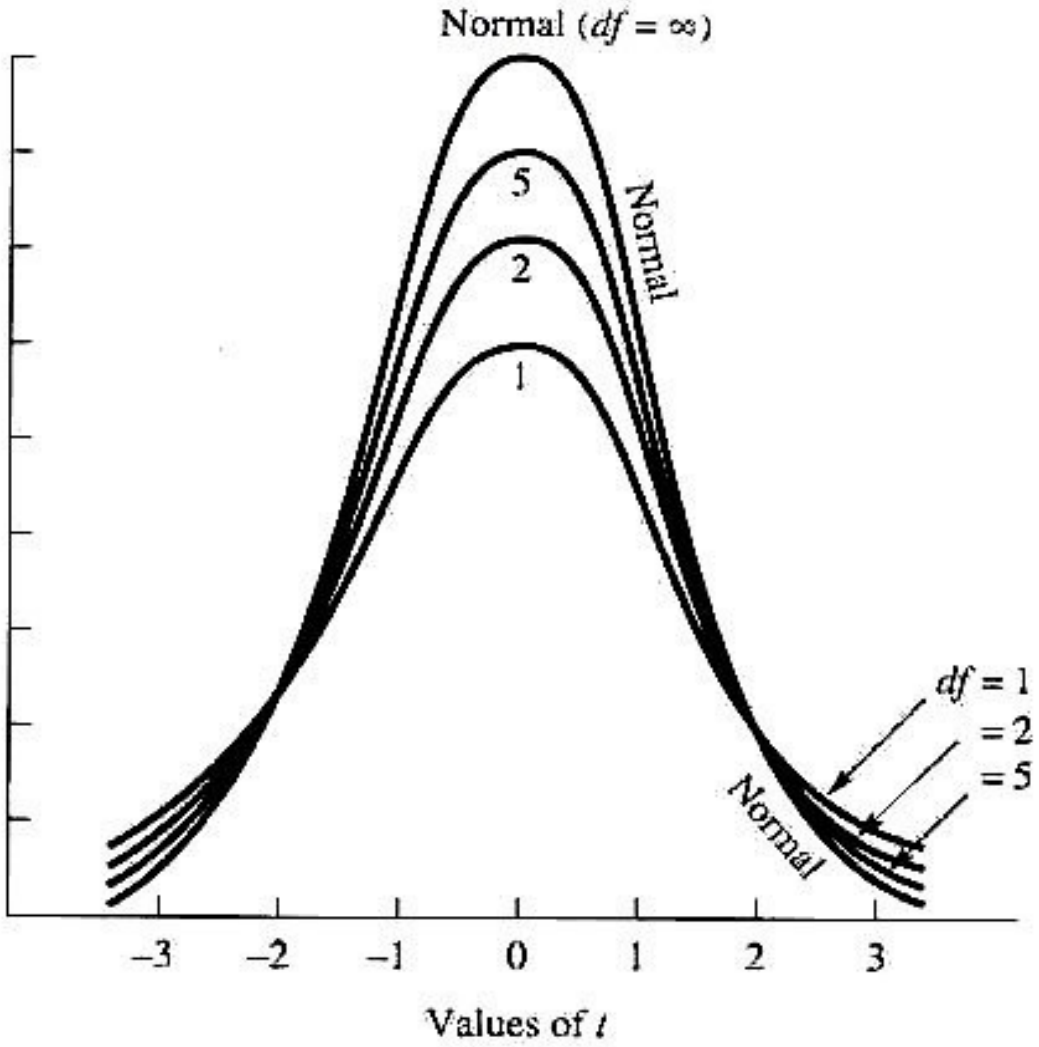
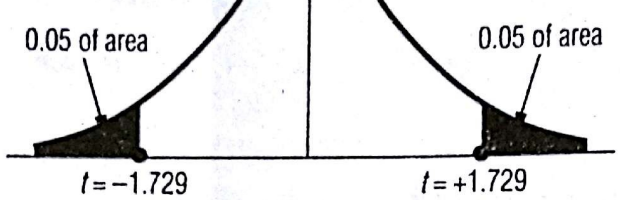


FIGURE 6.8 The Family of *t* Distributions

Example

- The plant manager wants an interval estimate of the mean coal consumption per week. He took a sample by measuring coal usage for ten weeks, found average to be 11,400 tons and standard deviation as 700 tons.
- Given $\sigma = 700$ and $n=10$, SE $= \sigma/\sqrt{n} = 221.38$ tons
- At the 95 percent confidence level and at 9 degrees of freedom the value in the table is 2.262.
- Upper confidence limit$=x+2.262$ SE$=$ 11,901 tons
- Lower confidence limit$=$ x-2.262 SE$=$ 10,899 tons
- We can report that with 95 percent confidence the mean weekly usage of coal lies between 10.899 and 11,901 tons.



Areas in Both Tails Combined for Student's *t* Distribution

Example: To find the value of *t* that corresponds to an area of 0.10 in both tails of the distribution combined, when there are 19 degrees of freedom, look under the 0.10 column, and proceed down to the 19 degrees of freedom row; the appropriate *t* value there is 1.729.

Degrees of Freedom	Area in Both Tails Combined			
	0.10	0.05	0.02	0.01
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
60	1.671	2.000	2.390	2.660
120	1.658	1.980	2.358	2.617
Normal Distribution	1.645	1.960	2.326	2.576

Problem to solve

- Sandra Cummins, the financial manager of Fike Lumber Company wanted to evaluate the receivable collection policy she had recently implemented. She sampled 24 accounts and the average collection period was 27.3 days with SD of 1.9. Construct a 98 percent confidence interval for the mean of the population (t value at 98% cl, with 23 df is 2.5).
- Hint: Given SD $\sigma=1.9$ $n=24$ $x= 27.3$ days

Calculate $SE=\sigma \div \sqrt{n}$

$= 1.9 \div \sqrt{24} = 1.9 \div 4.90 = 0.38775$

Upper limit= 27.3 + 2.5 * 0.38775 = 27.3 + 0.96 = 28.26 days

Lower limit = 27.3 – 2.5 * 0.38775 = 27.3 + 0.96= 26.34 days

With 98 percent confidence the limit lies between 26.3 days and 28.26 days.

Determining sample size

Problem continued

- zSE= 500
- 1.96* SE = 500
- $1.96*\sigma/\sqrt{n} = 500$
- $1.96*1500/500 = \sqrt{n}$
- $5.882 = \sqrt{n}$
- Hence n = 34.6 that is approximately 35.
- The university should take a sample of 35 business school graduates to get precision it wants in estimating the class's mean annual earnings.

Problem to solve

- If the population standard deviation is 200, find the sample size necessary to estimate the true mean within 100 points for a confidence level of 90 percent.
- Given SD = 200, zSE=100, z value at confidence level of 90 percent is 1.64, we need to calculate n. For that we need to use the formulaz SE=100.
- zSE= 100

1.64* SE= 100

$1.64*\sigma/\sqrt{n}= 100$

Sample size for proportion

- We want to know in an university the proportion of students that favours the new grading system. We would like a sample size that will enable us to be 90 percent certain in estimating the true proportion of the students within the plus and minus of .02.
- zSE= .02 1.64 SE =.02 1.64 $\sqrt{(pq/n)}$ = .02
- Taking p and q as 0.5 we get n = 1680
- To be 90 percent certain of estimating the true proportion within 0,02 we should pick up a simple random sample of 1,680 students to interview.

Problem to solve

For a test market, find the sample size needed to estimate the true proportion of consumers satisfied with a certain new product within plus or minus .03 percent at the 95 percent confidence interval. Assume you have no strong feeling about what the proportion is.

Given z*SE= .03, z at 95 percent confidence level is 1.96, p= .5, hence q=.5 we need to calculate n using the formula

$1.96 \sqrt{(pq/n)} = .03$

Summary Table

	When the populationis finite	When the population is infinite
Estimatingμ(the population mean)whenσ(population SD) is known	Upperlimit:x+zσ÷√n* √N-n/N-1 Lower limit: x-zσ÷√n* √N-n/N-1	Upper Limit=x+zσ÷√n Lower Limit=x-zσ÷√n
Whenσ(the population SD) is not known When n(sample size) is larger than 30	Upperlimit:x+zs÷√n* √N-n/N-1 Lower limit: x-zs÷√n* √N-n/N-1	Upper Limit=x+zσ÷√n Upper Limit=x-zσ÷√n
When n (sample sizeis less than 30	Upperlimit:x+ts÷√n* √N-n/N-1 Lower limit: x-ts÷√n* √N-n/N-1	Upper Limit=x+ts÷√n Upper Limit=x-ts÷√n
Estimating p (the population proportion)when n (sample size) is	Upperlimit:x+zσ÷√n* √N-n/N-1 Lower limit: x-zσ÷√n* √N-n/N-1	Upper Limit=p+zSE

Thank you.

Can you estimate the number of birds next morning in the garden?