

IIT KHARAGPUR



NPTEL ONLINE CERTIFICATION COURSES

Business Analytics for Management Decision

Sampling and Estimation

Rudra P. Pradhan
Vinod Gupta School of Management
IIT KHARAGPUR

NPTEL

Sampling and Estimation

- Statistical Sampling
- Estimating Population Parameters
- Sampling Error
- Sampling Distributions
- Interval Estimates
- Confidence Intervals
- Using Confidence Intervals for Decision Making
- Prediction Intervals
- Confidence Intervals and Sample Size

NPTEL

Statistical Sampling

- Sampling is the foundation of statistical analysis.
- Sampling plan - a description of the approach that is used to obtain samples from a population
- A sampling plan states:
  - its objectives
  - target population
  - population frame
  - operational procedures for data collection
  - statistical tools for data analysis

NPTEL

Statistical Sampling

Example 6.1 A Sampling Plan for a Market Research Study

- A company wants to understand how golfers might respond to a membership program that provides discounts at golf courses.
- Specify 5 components for a sampling plan.

NPTEL

Statistical Sampling

Example 6.1 (continued)

A Sampling Plan for a Market Research Study

- Objective - estimate the proportion of golfers who would join the program
- Target population - golfers over 25 years old
- Population frame - golfers who purchased equipment at particular stores
- Operational procedures - e-mail link to survey or direct-mail questionnaire
- Statistical tools - PivotTables to summarize data by demographic groups and estimate likelihood of joining the program

NPTEL

Statistical Sampling

Sampling Methods

- Subjective Methods
  - judgment sampling
  - convenience sampling
- Probability Sampling
  - simple random sampling
    involves selecting items from a population so that every subset of a given size has an equal chance of being selected

NPTEL

Statistical Sampling

Example 6.2 Simple Random Sampling with Excel

- Sample from the Excel database *Sales Transactions*

  Data
  Data Analysis
  Sampling

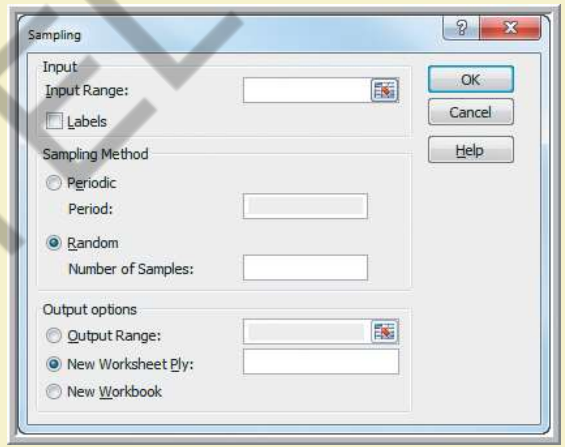


Figure 6.1

- *Periodic* selects every n^{th} number
- *Random* selects a simple random sample

NPTEL

Statistical Sampling

Example 6.2 (continued)

Simple Random Sampling with Excel

- Samples generated by Excel
- Sorted by customer ID
- Sampling is done *with replacement* so duplicates may occur.

	A
1	Sample of Customer IDs
2	10039
3	10089
4	10156
5	10157
6	10193
7	10213
8	10226
9	10232
10	10246
11	10257
12	10322
13	10335
14	10342
15	10357
16	10360
17	10360
18	10371
19	10431
20	10443
21	10448

Figure 6.2

NPTEL

Statistical Sampling

Additional Probability Sampling Methods

- Systematic (periodic) sampling
- Stratified sampling
- Cluster sampling
- Sampling from a continuous process

NPTEL

Statistical Sampling



Analytics in Practice: Using Sampling Techniques to Improve Distribution

- MillerCoors brewery wanted to better understand distributor performance
- Defined 7 attributes of proper distribution
- Collected data from distributors using stratified sampling based on market share
- Developed performance rankings of distributors and identified opportunities for improvement

NPTEL

Estimating Population Parameters

- Estimators are measures used to estimate unknown population parameters
- A point estimate is a single number derived from a sample that is used to estimate the value of a population parameter.
- Examples:

  Mean: $\overline{x}$ is a point estimate of μ

  Standard deviation: s is a point estimate of σ

  Proportion: $\hat{p}$ is a point estimate of π

NPTEL

Estimating Population Parameters

- Unbiased Estimators - the expected value of the estimator equals the population parameter
- Using $n - 1$ in the denominator of the sample variance s^2 results in an unbiased estimator of σ^2.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

is an unbiased estimator of

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

NPTEL

Sampling Error

- Sampling (statistical) error occurs because samples are only a subset of the total population
- Sampling error depends on the size of the sample relative to the population.
- Nonsampling error occurs when the sample does not adequately represent the target population.
- Nonsampling error usually results from a poor sample design or choosing the wrong population frame.

NPTEL

Sampling Error

Example 6.3 A Sampling Experiment

- A population is uniformly distributed between 0 and 10.
- Mean = (0 + 10)/2 = 5
- Variance = $(10 - 0)^2/12 = 8.333$
- Use Excel to generate 25 samples of size 10 from this population. Compute the mean of each.
- Prepare a histogram of the 25 sample means.
- Prepare a histogram of the 250 observations.

NPTEL

Sampling Error

Example 6.3 (continued) A Sampling Experiment

Observation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 22	Sample 23	Sample 24	Sample 25
1	5.3935	0.8756	9.9338	4.3294	7.1908	8.2244	8.4655	1.9404	9.9133
2	2.8282	2.5047	6.4480	5.9877	7.3946	3.3000	0.3632	9.8871	5.1079
3	5.2715	0.6949	1.5015	8.3935	3.1559	7.1023	9.3628	1.7844	7.3937
4	5.4912	0.7739	8.1466	5.5205	2.4586	8.7262	9.1598	7.5820	1.8513
5	9.3158	4.4591	6.3573	3.8679	1.1493	3.3854	1.2482	1.9391	5.4405
6	7.9745	6.9784	7.9962	2.3157	7.8564	8.9032	3.8716	5.8525	5.4164
7	6.7043	8.4039	5.1088	9.1098	1.1802	2.7732	2.4815	9.0817	4.3889
8	1.3041	2.5678	6.1794	7.8396	6.2709	0.5692	2.5800	1.1911	7.2430
9	0.9870	6.3964	8.2269	9.6112	6.6814	2.8306	4.6004	9.0274	6.1232
10	9.9493	9.3936	4.5015	5.2385	0.6970	3.7074	5.9062	0.6592	7.5021
Sample mean	5.5219	4.3048	6.4400	6.2214	4.4035	4.9522	4.8039	4.8945	6.0380

Average	5.0108
Standard Dev.	0.816673



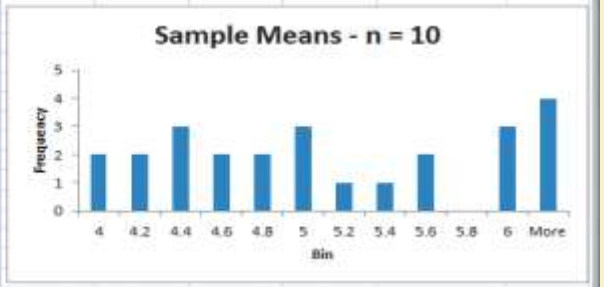


Figure 6.3

NPTEL

Sampling Error

Example 6.3 (continued) A Sampling Experiment

- Repeat the sampling experiment for samples of size 25, 100, and 500

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

Table 6.1

NPTEL

Sampling Error

Example 6.3 (continued) A Sampling Experiment

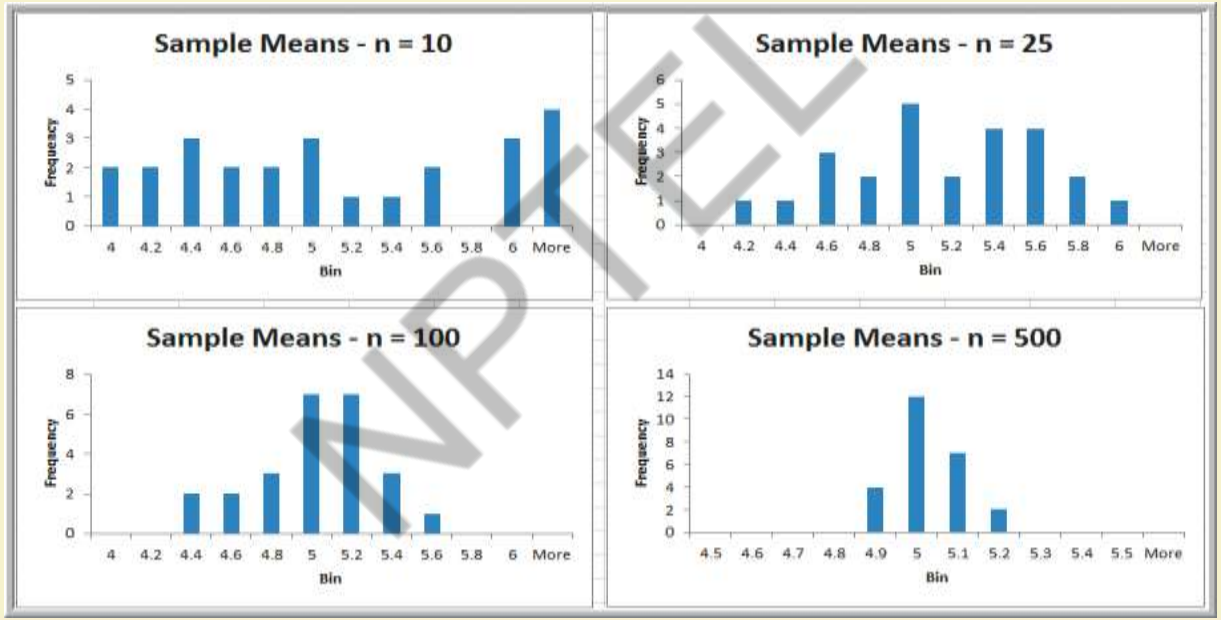


Figure 6.4

NPTEL

Sampling Error

Example 6.4 Estimating Sampling Error Using the Empirical Rules

- Using the empirical rule for 3 standard deviations away from the mean, ~99.7% of sample means should be between:

[2.55, 7.45] for n = 10

[3.65, 6.35] for n = 25

[4.09, 5.91] for n = 100

[4.76, 5.24] for n = 500

Sample Size	Average of 25 Sample Means	Standard Deviation of 25 Sample Means
10	5.0108	0.816673
25	5.0779	0.451351
100	4.9173	0.301941
500	4.9754	0.078993

Table 6.1

NPTEL

Sampling Distributions

Sampling Distribution of the Mean

- Formed by the means of *all possible* samples of a fixed size *n* from some population

Standard Error of the Mean

- The standard deviation of the sampling distribution of the mean
- Standard Error of the mean = $\sigma/\sqrt{n}$
- Larger sample sizes have less sampling error and provide greater accuracy in estimating the true population mean.

NPTEL

Sampling Distributions

Example 6.5

Computing the Standard Error of the Mean

- For the uniformly distributed population, we found $\sigma^2 = 8.333$ and, therefore, $\sigma = 2.89$
- Compute the standard error of the mean for sample sizes of 10, 25, 100, 500.

Sample Size, *n*	Standard Error of the Mean
10	0.914
25	0.577
100	0.289
500	0.129

Standard Deviation of 25 Sample Means
0.816673
0.451351
0.301941
0.078993

For comparison from Table 6.1

NPTEL

Sampling Distributions

Central Limit Theorem

- If the sample size is large enough, then the sampling distribution of the mean is:
  - approximately normally distributed *regardless* of the distribution of the population
  - has a mean equal to the population mean
- If the population is normally distributed, then the sampling distribution is also normally distributed for *any* sample size.
- This theorem is one of the most important practical results in statistics.

NPTEL

Sampling Distributions

Example 6.6

Using the Standard Error in Probability Calculations

- The purchase order amounts for books on a publisher's Web site is normally distributed with a mean of $36 and a standard deviation of $8.
- Find the probability that:

  a) someone's purchase amount exceeds $40

  b) the mean purchase amount for 16 customers exceeds $40

NPTEL

Sampling Distributions

Example 6.6 (continued)

Using the Standard Error in Probability Calculations

a) $P(x > 40)$ = 1− NORM.DIST(40, 36, 8, true)

= 0.3085

b) $P(\bar{x} > 40)$ = 1− NORM.DIST(40, 36, 2, true)

= 0.0228

NPTEL

Interval Estimates

Example 6.7 Interval Estimates in the News

- A Gallup poll might report that 56% of voters support a certain candidate with a margin of error of ± 3%.
- We would have a lot of confidence that the candidate would win.

- If, instead, the poll reported a 52% level of support with a ± 4% margin of error, we would be less confident in predicting a win for the candidate.

NPTEL

Confidence Intervals

Interval Estimates

- Provide a range for a population characteristic based on a sample.
- A confidence interval of $100(1-\alpha)\%$ is an interval $[A, B]$ such that the probability of falling between A and B is $1-\alpha$.
- $1-\alpha$ is called the level of confidence.
- 90%, 95%, and 99% are common values for $1-\alpha$.
- Confidence intervals provide a way of assessing the accuracy of a point estimate.

NPTEL

Confidence Intervals

Confidence Interval For the Mean with Known Population Standard Deviation

- Sample mean ± margin of error
- Sample mean ± $z_{\alpha/2}$ (standard error)

$$\bar{x} \pm z_{\alpha/2}\,(\sigma/\sqrt{n})$$

where $z_{\alpha/2}$ is the value of the standard normal random variable for an upper tail area of $\alpha/2$ (or a lower tail area of $1 - \alpha/2$).

NPTEL

Confidence Intervals

Example 6.8 Computing a Confidence Interval with a Known Standard Deviation

- A production process fills bottles of liquid detergent.
- The standard deviation in filling volumes is constant at 15 mls.
- A sample of 25 bottles revealed a mean filling volume of 796 mls.
- Give a 95% confidence interval estimate of the mean filling volume for the population.

NPTEL

Confidence Intervals

Example 6.8 (continued) Computing a Confidence Interval with a Known Standard Deviation

$$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$$

= Mean ± CONFIDENCE.NORM(*alpha, stdev, size*)

= 796 ± CONFIDENCE.NORM(.05, 15, 25)

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	25				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	5.879892				
10	Lower	790.1201				
11	Upper	801.8799				

Figure 6.5

NPTEL

Confidence Intervals

The *t*-Distribution

- Used for confidence intervals when the population standard deviation in unknown.
- Its only parameter is the degrees of freedom *(df).*

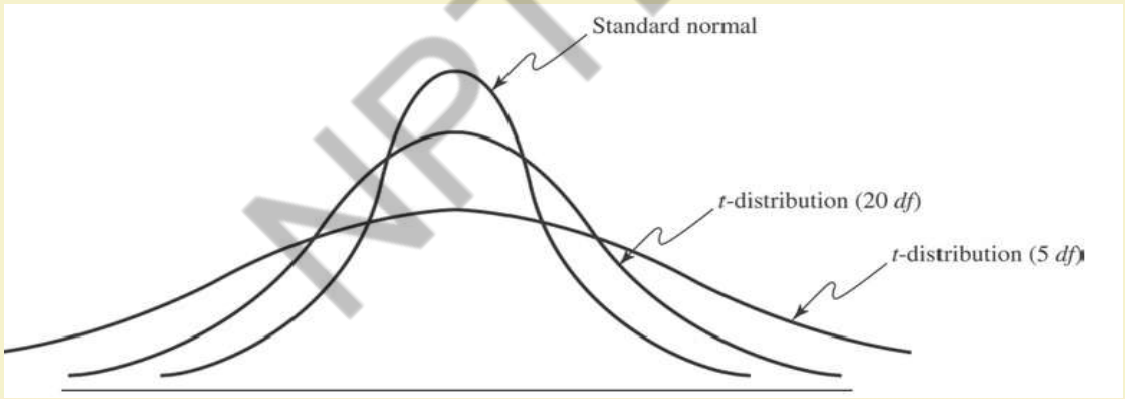


Figure 6.6

NPTEL

Confidence Intervals

Confidence Interval for a Population Mean with an Unknown Standard Deviation

- Sample mean ± margin of error
- Sample mean ± $t_{\alpha/2}$ (estimated standard error)

$$\bar{x} \pm t_{\alpha/2,n-1}\,(s/\sqrt{n})$$

where $t_{\alpha/2}$ is the value of the *t*-distribution with $df = n - 1$ for an upper tail area of $\alpha/2$.

- *t* values are found in Table 2 of Appendix B.

NPTEL

Confidence Intervals

Example 6.9 Computing a Confidence Interval with an Unknown Standard Deviation

- A large bank has sample data used in making credit decisions.
- Give a 95% confidence interval estimate of the mean revolving balance of homeowner applicants.

	A	B	C	D	E	F
1	Credit Approval Decisions					
2						
3	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
4	Y	725	20	$ 11,320	25%	Approve
5	Y	573	9	$ 7,200	70%	Reject
6	Y	677	11	$ 20,000	55%	Approve
7	N	625	15	$ 12,800	65%	Reject
8	N	527	12	$ 5,700	75%	Reject
9	Y	795	22	$ 9,000	12%	Approve
10	N	733	7	$ 35,200	20%	Approve

Figure 6.7

NPTEL

Confidence Intervals

Example 6.9 (continued) Computing a Confidence Interval with an Unknown Standard Deviation

$$\bar{x} \pm t_{\alpha/2,n-1}(s/\sqrt{n})$$

= Mean ± T.INV(*confidence level, df*)*s/SQRT(n)

= Mean ± CONFIDENCE.T(*alpha, stdev, size*)

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Unknown					
2						
3	Alpha	0.05				
4	Sample standard deviation	5393.38				
5	Sample size	27				
6	Sample average	12630.37				
7						
8	Confidence Interval	95%				
9	t-value	2.056				
10	Error	2133.55				
11	Lower	10496.82				
12	Upper	14763.92				

Figure 6.8

NPTEL

Confidence Intervals

Confidence Interval for a Proportion

- Sample proportion ± margin of error
- Sample proportion ± $z_{\alpha/2}$ (standard error)

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = x/n$ is an unbiased estimator of the population proportion π, and x is the number of successful outcomes in a sample of size n.

NPTEL

Confidence Intervals

Example 6.10 Computing a Confidence Interval for a Proportion (of those willing to pay a lower health insurance premium for a lower deductible)

	A	B	C	D	E	F	G
1	Insurance Survey						
2							
3	Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
4	36	F	Some college	Divorced	4	4	N
5	55	F	Some college	Divorced	2	1	N
6	61	M	Graduate degree	Widowed	26	3	N
7	65	F	Some college	Married	9	4	N
8	53	F	Graduate degree	Married	6	4	N
9	50	F	Graduate degree	Married	10	5	N
10	28	F	College graduate	Married	4	5	N
11	62	F	College graduate	Divorced	9	3	N
12	48	M	Graduate degree	Married	6	5	N
13	31	M	Graduate degree	Married	1	5	N
14	57	F	College graduate	Married	4	5	N
15	44	M	College graduate	Married	2	3	N
16	38	M	Some college	Married	3	2	N
17	27	M	Some college	Married	2	3	N
18	56	M	Graduate degree	Married	4	4	Y
19	43	F	College graduate	Married	5	3	Y
20	45	M	College graduate	Married	15	3	Y
21	42	F	College graduate	Married	12	3	Y
22	29	M	Graduate degree	Single	10	5	N
23	28	F	Some college	Married	3	4	Y
24	36	M	Some college	Divorced	15	4	Y
25	49	F	Graduate degree	Married	2	5	N
26	46	F	College graduate	Divorced	20	4	N
27	52	F	College graduate	Married	18	2	N
28							
29	*Measured from 1-5 with 5 being highly satisfied						
30	**Would you be willing to pay a lower premium for a higher deductible?						

Figure 6.9

NPTEL

Confidence Intervals

Example 6.10 (continued) Computing a 95% Confidence Interval for a Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample proportion ± NORM.S.INV((alpha/2)* (standard error of the sample proportion))

	A	B
1	Confidence Interval for a Proportion	
2		
3	Alpha	0.05
4	Sample proportion	0.25
5	Sample size	24
6		
7	Confidence Interval	95%
8	z-value	1.96
9	Standard error	0.088388
10	Lower	0.076762
11	Upper	0.423238

Figure 6.10

NPTEL

Using Confidence Intervals for Decision Making

Example 6.11 Drawing a Conclusion about a Population Mean Using a Confidence Interval

- In Example 6.8 we obtained a confidence interval for the bottle-filling process as [790.12, 801.88]
- The required volume is 800 and the sample mean is 796 mls.
- Should machine adjustments be made?

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	25				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	5.879892				
10	Lower	790.1201				
11	Upper	801.8799				

Figure 6.5

NPTEL

Using Confidence Intervals for Decision Making

Example 6.12 Using a Confidence Interval to Predict Election Returns

- An exit poll of 1300 voters found that 670 voted for a particular candidate in a two-person race; that is $\hat{p} = 0.515$.
- The 95% confidence interval for the population proportion of voters is [0.488, 0.543].
- Would it be wise to predict the winner based on this exit poll?

NPTEL

Prediction Intervals

Computing a Prediction Interval

- Confidence intervals estimate the value of a parameter such as a MEAN or PROPORTION.
- Prediction intervals provide a range of values for a new OBSERVATION from the same population.
- Prediction intervals are wider than confidence intervals.
- Confidence interval: $\bar{x} \pm t_{\alpha/2,n-1}\,(s/\sqrt{n})$
- Prediction interval: $\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1+\frac{1}{n}}\right)$

NPTEL

Confidence Intervals and Sample Size

Example 6.13 Computing a Prediction Interval

Compute a 95% prediction interval for the revolving balances of customers (*Credit Approval Decisions*)

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s\sqrt{1+\frac{1}{n}}\right)$$

$$12{,}630.37 \pm 2.056(5{,}393.38\sqrt{1+\frac{1}{27}}$$

$$12{,}630.37 \pm 2.094$$

$$12{,}630.37 \pm 11{,}292.27$$

$$[1{,}338.10,\ \ 23{,}922.64]$$

Prediction interval width = 22,585

	A	B
1	Confidence Interval for Population Mean	
2		
3	Alpha	0.05
4	Sample standard deviation	5393.38
5	Sample size	27
6	Sample average	12630.37
7		
8	Confidence Interval	95%
9	t-value	2.056
10	Error	2133.55
11	Lower	10496.82
12	Upper	14763.92

From Example 6.9

Confidence interval width = 4,267

NPTEL

Confidence Intervals and Sample Size

Sample Size Determination

- For estimating a Mean $n \geq (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$

- For estimating a Proportion $n \geq (z_{\alpha/2})^2 \frac{\pi(1-\pi)}{E^2}$

where E is the desired margin of error

and π is an estimate of the proportion successes.

One can use a conservative value of 0.5 for π if no estimate is available.

NPTEL

Confidence Intervals and Sample Size

Example 6.14

Sample Size Determination for the Mean

- In the liquid detergent example, the margin of error was 2.985 mls.
- What is sample size is needed to reduce the margin of error to at most 3 mls?

$$n \geq (z_{\alpha/2})^2 \frac{(\sigma^2)}{E^2}$$

$$= (1.96)^2 \frac{(15^2)}{3^2} = 96.04$$

Round up to 97 samples.

	A	B	C	D	E	F
1	Confidence Interval for Population Mean, Standard Deviation Known					
2						
3	Alpha	0.05				
4	Standard deviation	15				
5	Sample size	97				
6	Sample average	796				
7						
8	Confidence Interval	95%				
9	Error	2.985063				
10	Lower	793.0149				
11	Upper	798.9851				

Figure 6.11

NPTEL

Confidence Intervals and Sample Size

Example 6.15

Sample Size Determination for a Proportion

- In a voting poll in which there is no prior estimate of π, what size sample is needed to ensure a margin of error of at most ± 2% when using a 95% confidence interval?

$$n \geq (z_{\alpha/2})^2 \frac{\pi(1-\pi)}{E^2}$$

$$= (1.96)^2 \frac{(0.5)(1-0.5)}{0.02^2} = 2{,}401$$

NPTEL

Chapter 6 - Key Terms

- Central limit theorem
- Cluster sampling
- Confidence interval
- Convenience sampling
- Degrees of freedom
- Estimation
- Estimators
- Interval estimate
- Judgment sampling
- Level of confidence
- Nonsampling error
- Point Estimate
- Population frame
- Prediction interval
- Probability interval
- Sample proportion
- Sampling (statistical) error

NPTEL

Keywords (continued)

- Sampling distribution of the mean
- Sampling plan
- Simple random sampling
- Standard error of the mean
- Stratified sampling
- Systematic (or periodic) sampling
- *t*-distribution

NPTEL

Case Study: Performance Lawn Equipment (6)

- Recall that PLE produces lawnmowers and a medium size diesel power lawn tractor.
- Determine the probability a customer is highly satisfied for each geographic region.
- Compute a confidence interval estimate of customer service response times.
- Find the required sample size for a confidence interval estimate of blade weights.
- Write a formal report summarizing your results.



NPTEL

Books to read

- Business Analytics by J R Evans
- Business Analytics for Managers by H N Laursen

NPTEL

Thank You!!

NPTEL