air academy ASSOCIATES

Confidence Intervals and Sample Size

www.airacad.com

User Agreement and Copyright Information

- This recording and the accompanying guide contain copyrighted and proprietary content of Air Academy Associates, LLC. You are authorized to use this material for personal reference, but not for any commercial use. You may not modify, license, sub-license, distribute, copy, translate or create derivative works based on this guide, in part or in whole, without permission from Air Academy Associates.

- Other copyright information:

  – Six Sigma is a service mark of Motorola, Inc. Microsoft® and Excel® are registered trademarks of Microsoft Corporation in the United States and in other territories.

  – SPC XL™ and DOE Pro XL™ are copyright SigmaZone.com and Air Academy Associates, LLC. You may not copy, modify, distribute, display, license, reproduce, sell or use commercially any screen shots or any component contained therein without the express written permission of SigmaZone.com and Air Academy Associates, LLC. All rights reserved. SigmaZone.com may be contacted at www.SigmaZone.com. Air Academy Associates may be contacted at www.airacad.com.

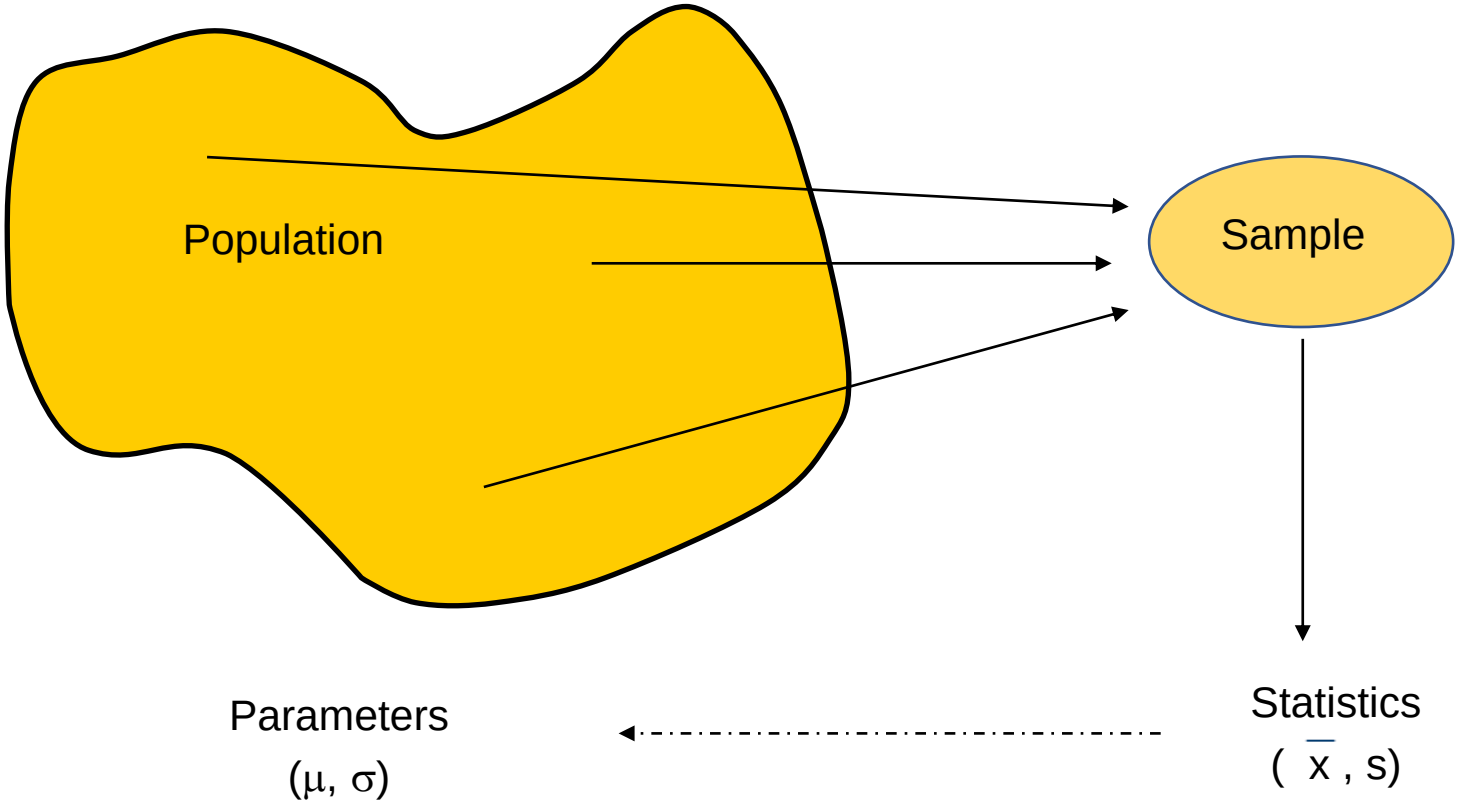
Confidence Intervals and Sample Size

- In this session, we will discuss:
  - Sampling and the Central Limit Theorem
  - Confidence Intervals for variables data
  - Confidence Intervals for attribute data
  - Sample Size calculations for variables data
  - Sample Size calculations for attribute data

- A list of supplemental material and additional practice/review questions for this session are provided at the end of this presentation
- You can download the pdf of this presentation, along with any supporting data files, on the site where you are accessing this course

Sampling a Population

Sampling … “the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population”



The What and Why of Sampling

- Sampling is done to help us understand and make inferences

- Sampling involves gathering data which is representative of the process being studied

- We may use sample data to:
  - baseline a process (describe performance)
  - estimate parameters (e.g., the average particle size, the percent defective rate)
  - help make comparisons during a pilot test or trial
  - help draw accurate conclusions

- Determining how much data we need (sample size) and then interpreting the results of the sample (using confidence intervals) helps us make good decisions

Exercise: Illustrating the Effect of Averaging Data

Rolling a Single Die.xlsx

- After listening to my instructions, pause the video and go through the exercise
- Open up the data file named *Rolling a Single Die*. This file contains the data captured after rolling a single die four thousand times. There are 1000 rows and 4 columns. **This will be our sample of the population**
- Using SPC XL, draw a histogram of all the data making sure you check the box for "survey data". For a video on how to use SPC XL to draw a histogram go to: https://airacad.com/our-insights/training-videos/spc-xl/
- What is the shape of the histogram? Is this what you expected? Take note of the mean and standard deviation of the data. Now restart the video

Parent Data (roll of a single die)
Partial view of the data

Single Die Rolled 4000 times

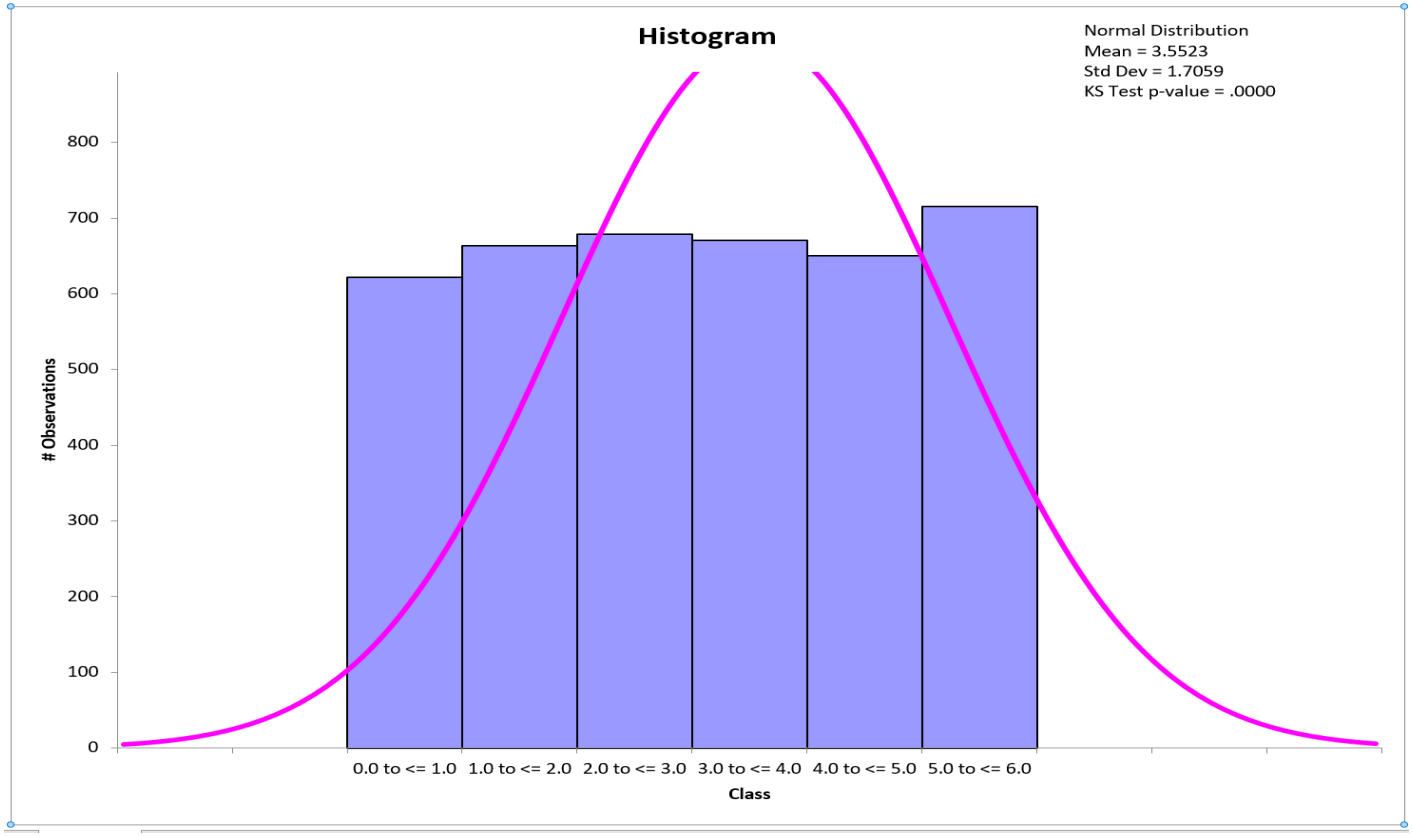
6	4	1	1
6	5	3	5
1	5	3	4
3	1	6	5
3	2	1	3
2	4	3	6
4	4	2	2
3	3	5	1
5	4	5	6
5	5	5	3
3	6	5	2
5	3	6	6
1	2	2	6
6	6	5	2

Exercise: Illustrating the Effect of Averaging Data (cont.)

- Drawing a histogram of the individual data points gives us a relatively flat graph. This data is uniformly distributed. A uniform distribution makes sense since we rolled a fair die 4000 times and each number is equally likely to occur

- Note the average of the data is 3.5523. You probably expected 3.5. Our result is very close to what we expected

- Note the standard deviation of the data is 1.7059



Exercise: Illustrating the Effect of Averaging Data (cont.)

Rolling four die and averaging.xlsx

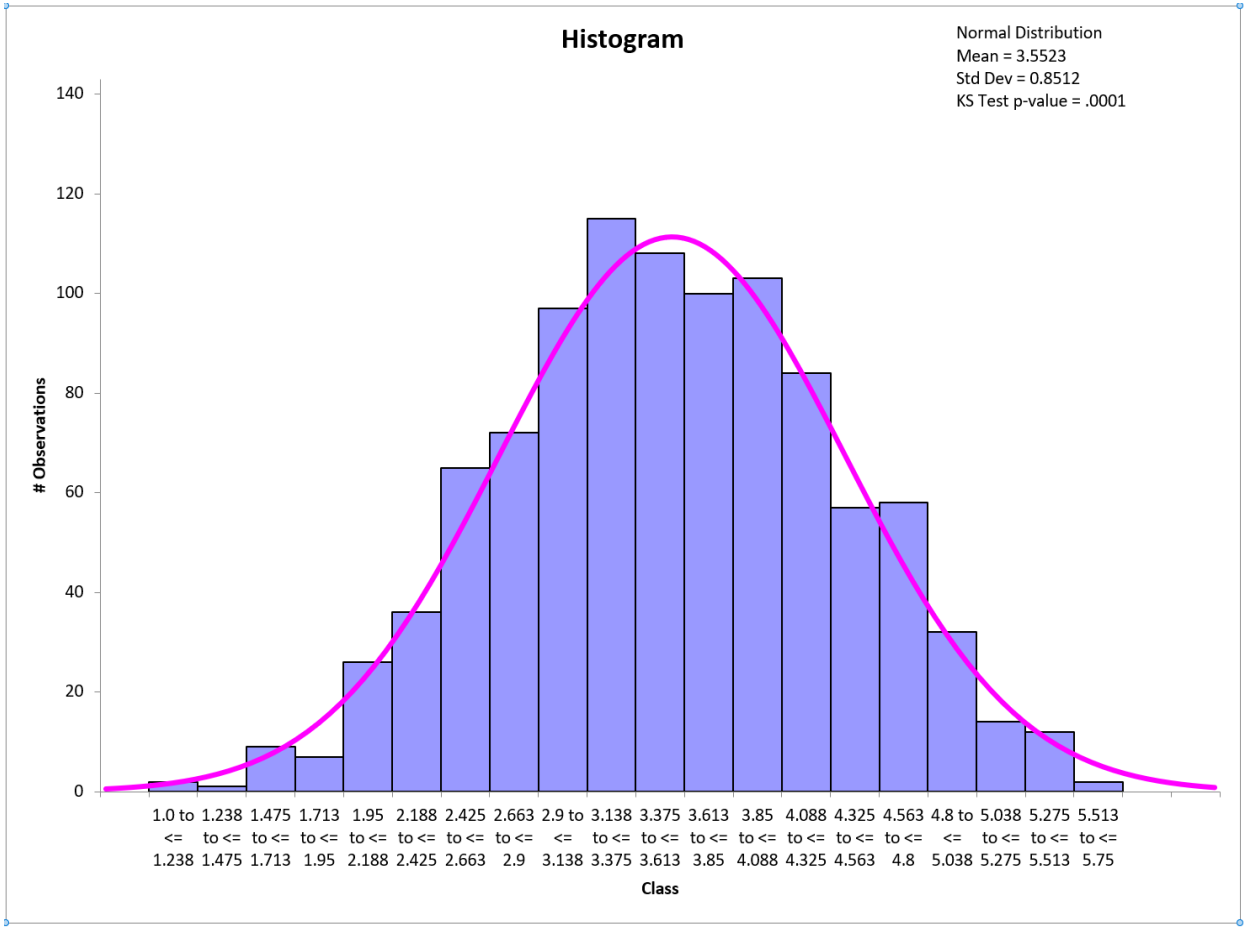
- After listening to my instructions, pause the video and follow the instructions below
- Open the data file named *Rolling Four Die and Averaging*. This time we will draw a histogram of the average of each row of four die
- Using SPC XL, draw a histogram of the data using the average column ONLY. DO NOT check the box for "survey data", but specify 20 "classes" of data so there are no "missing" columns in the graph
- What is the shape of the histogram? Is this what you expected? Take note of the mean and standard deviation of the data. Now restart the video

Child Data (Average of four rolls)
Partial view of the data

Single Die Rolled 4000 times				Average
6	4	1	1	3
6	5	3	5	4.75
1	5	3	4	3.25
3	1	6	5	3.75
3	2	1	3	2.25
2	4	3	6	3.75
4	4	2	2	3
3	3	5	1	3
5	4	5	6	5
5	5	5	3	4.5
3	6	5	2	4
5	3	6	6	5
1	2	2	6	2.75
6	6	5	2	4.75

Exercise: Illustrating the Effect of Averaging Data (cont.)

- Drawing a graph of the average of four die results in a normally distributed graph. Is that what you expected? Remember, the first histogram was flat
- The average is 3.5523, the same as the average of the individual x's
- The standard deviation is 0.8512. This is different from the other histogram



Central Limit Theorem

- What you just experienced is referred to as “The Sampling Distribution of the Mean” or the Central Limit Theorem
  - The shape of the first histogram (individual data points) is “flat” or uniform
  - The shape of the averages is normal
  - Because the averages are normally distributed, we can use the 68/95/99% “rule”
- The averages of the two histograms are exactly the same. They should be since it is the same data!!!!

Parent Data (roll of a single die)

Mean = 3.5523
Std Dev = 1.7059

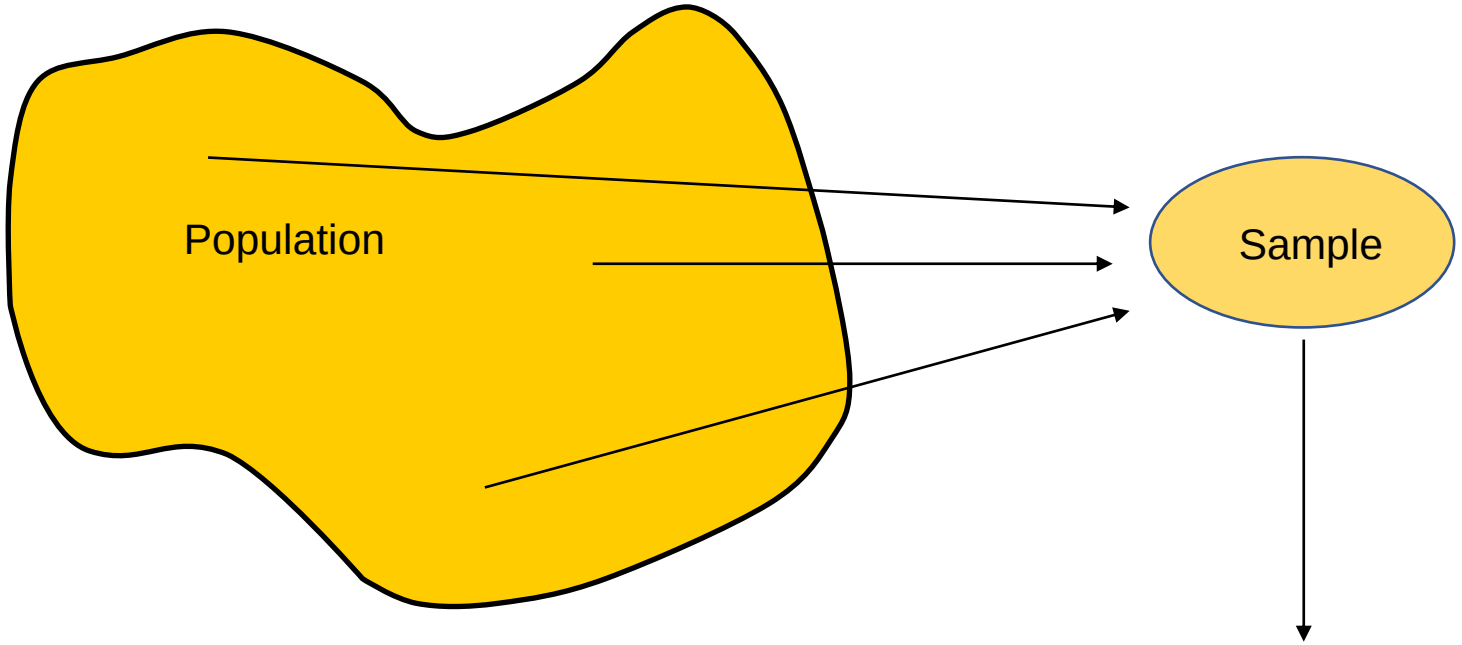
Child Data (Average of four rolls)

Mean = 3.5523
Std Dev = 0.8512

- The standard deviations are NOT the same. Why not?
  - Because, averaging pushes the values “towards” the center, hence the name: Central Limit Theorem
  - The ratio of the two sigma's is: 1.7059 / 0.8512 = 2
  - Why? Because the ratio of the Sigma's is related to sample size. Remember, the child data set averaged the sum of four die, So:

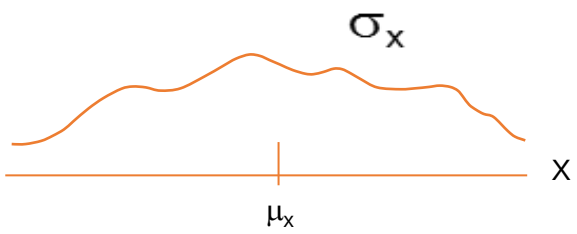
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Sampling, Confidence Intervals and Sample Size

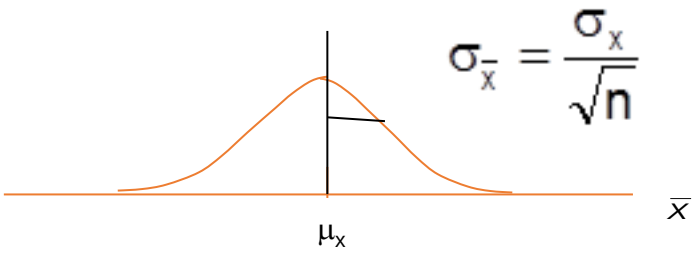


Parameters
(μ, σ)

Statistics
($\bar{x}$, s)



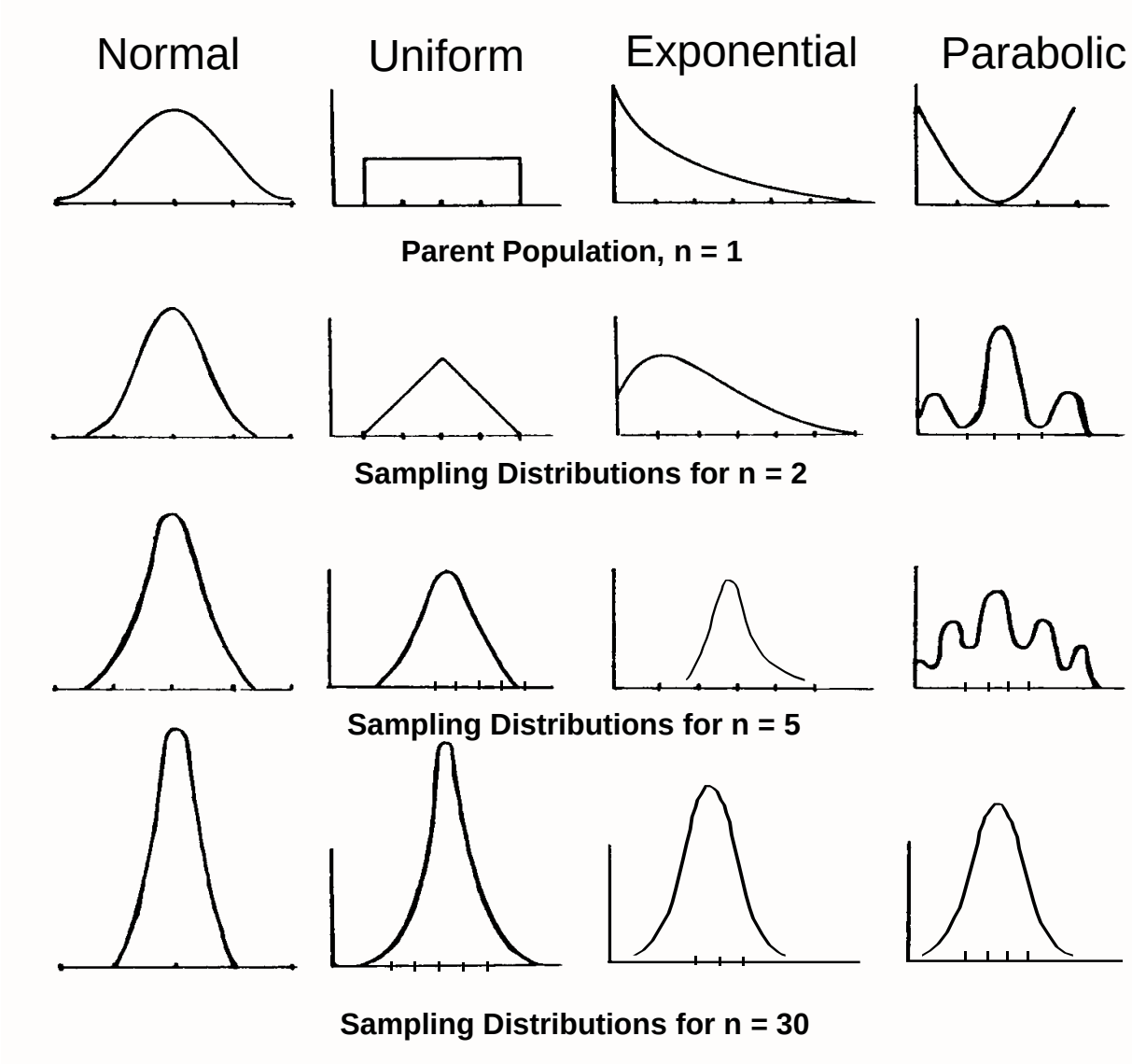
Parent



Child

Central Limit Theorem

For almost all populations, the sampling distribution of the mean can be approximated closely by a normal distribution, provided the sample size is sufficiently large



Sampling Distributions of $\bar{X}$ for Various Sample Sizes

Sampling and Confidence Intervals

- Suppose we measure the time it takes for a customer service representative to answer a call

- We use the data from our sample to estimate the average call time. How good is our estimate? How close is $\bar{x}$ to μ?

- The answer is based on the 68/95/99% rule, because the $\bar{x}$'s are normally distributed. The normal distribution allows us to generate a margin of error

- Confidence intervals provide error bounds of uncertainty based on the data

Confidence Interval =

Sample Estimate ± Margin of Error

Confidence Interval for Population Mean (Continuous Data)

$$\begin{pmatrix} U \\ L \end{pmatrix} = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

Where

U	**=**	**upper confidence limit**
L	**=**	**lower confidence limit**
$\bar{x}$	**=**	**sample average**
Z	**=**	**2 (for 95% confidence) or 3 (for 99% confidence)**
s	**=**	**sample standard deviation**
n	**=**	**sample size**

Computational Template for Confidence Limits

EXAMPLE:

Over the course of a week, we randomly select and time 25 customer service calls. We find that the average time is 18.2 minutes with a standard deviation of 2.5 minutes. What is a 99% confidence interval for the true average service time?

Confidence Interval for Population Mean (Continuous Data)

Stop the video and calculate the upper and lower confidence bounds based on the values given in the table. One decimal place accuracy is adequate

EXAMPLE:

Over the course of a week, we randomly select and time 25 customer service calls. We find that the average time is 18.2 minutes with a standard deviation of 2.5 minutes. What is a 99% confidence interval for the true average service time?

$$\begin{pmatrix} U \\ L \end{pmatrix} = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

Where

U	**=**	**upper confidence limit**
L	**=**	**lower confidence limit**
$\bar{x}$	**=**	**18.2 minutes**
Z	**=**	**3 (for 99% confidence)**
s	**=**	**2.5 minutes**
n	**=**	**25 data points (calls)**

Computational Template for Confidence Limits

N	25
X-bar	18.2
Std Dev	2.5
Z	3
Upper Limit	19.7
Lower Limit	16.7

Using SPC XL for Confidence Intervals for the Population Mean, μ (Continuous Data)

- SPC XL will produce a more exact interval, which will differ slightly from your manual calculations because the software uses more exact values from a t-distribution
- Stop the Video and replicate this output. From the SigmaZone (SPC XL) ribbon:
  - Analysis Tools > Confidence Interval > Mean (Normal)
  - Type in the appropriate values (in the yellow area)

Normal Confidence Interval (Mean)	
User defined parameters	
Sample Size (n)	25
Sample Avg	18.2
Sample Standard Dev	2.5
Confidence Level	99.00%
Confidence Interval	
Lower Limit	Upper Limit
16.80153025	19.59846975

- Interpreting the interval is critical. The confidence interval means we are 99% confident that μ falls in the given confidence interval
- For more instruction on using SPC XL for generating Confidence Intervals, go to: https://airacad.com/our-insights/training-videos/spc-xl/

Using SPC XL for Confidence Intervals for the Population Mean, μ (Continuous Data)

- Exercise: Suppose we're studying the average cycle time to pay invoices. We randomly sample 30 invoices, and find the average is 7.8 days with a standard deviation of 1.4 days. Using SPC XL, construct a 95% confidence interval for the true average cycle time for paying these invoices
- Stop the Video and use SPC XL to generate the following output:

Normal Confidence Interval (Mean)	
User defined parameters	
Sample Size (n)	30
Sample Avg	7.8
Sample Standard Dev	1.4
Confidence Level	95.00%
Confidence Interval	
Lower Limit	**Upper Limit**
7.277231409	8.322768591

Confidence Interval for Population Proportion π (Binary Data)

$$\begin{pmatrix} U \\ L \end{pmatrix} = p \pm Z\sqrt{\frac{pq}{n}}$$

Where

U	**=**	**upper confidence limit**
L	**=**	**lower confidence limit**
p	**=**	**proportion of "defectives" (or category of interest) in the sample**
q	**=**	**1 – p (q is the proportion of "non- defectives")**
Z	**=**	**2 (for 95% confidence) or 3 (for 99% confidence)**
n	**=**	**sample size**

Computational Template for Confidence Limits

EXAMPLE:

You work in a finance office and are in charge of processing travel vouchers submitted by several different organizations. You sample 100 vouchers and find 8 to have discrepancies or errors. Find a 95% confidence interval for the true but unknown proportion of vouchers containing errors

Using SPC XL for Confidence Intervals for Population Proportion, π (Binary Data)

- We'll skip doing the calculations by hand and use SPC XL to calculate the answer
- After listening to my instructions, pause the video and generate the following calculations using SPC XL
- From the SigmaZone (SPC XL) ribbon:

  Analysis Tools > Confidence Interval > Proportion (Binomial)

Binomial Confidence Interval (Proportion)

User defined parameters	
Sample Size (n)	100
Number Defective(x)	8
Confidence Level	95.00%

Confidence Interval		
Lower Limit	< p <	Upper Limit
0.035171509	0.08	0.151557446

- For more instruction on using SPC XL for generating Population Proportion Confidence Intervals, go to: https://airacad.com/our-insights/training-videos/spc-xl/

Using SPC XL for Confidence Intervals for Population Proportion, π (Binary Data)

- Exercise: Suppose in the previous example we had sampled 1,000 travel vouchers and found 80 to have discrepancies or errors. Find a 95% confidence interval for the true but unknown proportion of vouchers containing error
- Pause the video. Use SPC-XL to generate the confidence interval shown below

Binomial Confidence Interval (Proportion)

User defined parameters	
Sample Size (n)	1000
Number Defective(x)	80
Confidence Level	95.00%

Confidence Interval		
Lower Limit	< p <	Upper Limit
0.063941983	0.08	0.098579779

- Now, restart the video
- What effect did the sample size have on the confidence interval?
- Larger sample sizes produce much tighter confidence intervals. This is a key take-away of the central limit theorem
- Now we'll move on to calculating the appropriate sample sizes

Calculating the Proper Sample Size

- Sample size calculations follow directly from the confidence interval formulas

$$\begin{pmatrix} U \\ L \end{pmatrix} = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

$$\begin{pmatrix} U \\ L \end{pmatrix} = p \pm Z\sqrt{\frac{pq}{n}}$$

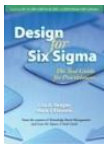
- How much data is needed to estimate a parameter with a certain desired margin of error?

Sample Size Considerations

- Practical Considerations
  - What is your timeframe?
  - How expensive is data collection?

- Statistical Considerations
  - What is the population variation?
  - How much precision do you want in your estimate? (your desired margin of error, or width of the confidence interval)
  - What level of confidence do you desire?

- These are the factors that determine the appropriate sample size
  - Variation of the population
  - Precision: Width of the confidence interval
  - Confidence level



Sample Size

Sample Size Formulas

- Solve the confidence interval equations for “n”, sample size
- For estimating a mean value

$$\begin{pmatrix} U \\ L \end{pmatrix} = \bar{x} \pm \underbrace{Z\left(\frac{s}{\sqrt{n}}\right)}_{= h} \quad \Rightarrow \quad n = \left\lceil \left(\frac{Z\hat{\sigma}}{h}\right)^2 \right\rceil$$

- For estimating a proportion

$$\begin{pmatrix} U \\ L \end{pmatrix} = p \pm \underbrace{Z\sqrt{\frac{pq}{n}}}_{= h} \quad \Rightarrow \quad n = \left\lceil \frac{Z^2 pq}{h^2} \right\rceil$$

Example: Sample Size for Estimating a Mean

- A human resources specialist is preparing a survey to send out to a randomly selected sample of employees from her organization. The organization has over 10,000 employees from which she can select the sample. The questions on the survey require responses on a 5 point Likert scale (i.e., 1 = strongly agree, 2 = moderately agree, etc.)

- The specialist would like to determine the sample size needed to estimate the true mean for each of the survey questions. She would like to estimate the true mean to within +/- 0.25 with 95% confidence (She expects the standard deviation to be around 1)
- Pause the video and replicate the analysis shown below:

Sample Size to Estimate the Mean of a Normal Distribution	
User defined parameters	
Estimated Standard Dev	1
Half Interval Width	0.25
Confidence Level	95.00%
Results	
Estimated Sample Size (n)	62

For more instruction on using SPC XL to calculate sample size, go to: https://airacad.com/our-insights/training-videos/spc-xl/

Sample Size for Estimating a Mean (cont.)

- In this example, the standard deviation is given to us. In real life that will not be the case. How do you "estimate" the standard deviation?

- Methods for estimating the standard deviation

  - Use historical data, if available
  - Use data from a similar process
  - Take a small pre-sample of data
  - Make an estimate based on the expected range
    - Use Range/ 6 (or range/ 4 to be more conservative)
    - Example: estimated std deviation = (5-1)/4 = 4/4 = 1

Example: Sample Size for Estimating a Proportion

- Suppose we're interested in estimating the true proportion of users who are dissatisfied with our customer support
- We believe that the true proportion is around 0.10 (10%) or less, and we'd like to estimate the true proportion to within +/- 0.03 (3%) with 95% confidence
- Pause the video and replicate this analysis. Then restart the video
- From the SigmaZone (SPC XL) ribbon:
  - Analysis Tools > Sample Size > Binomial Conf. Interval (proportion)

Binomial Sample Size	
User defined parameters	
Proportion defectives (p)	0.1
Half Interval Width	0.03
Confidence Level	95.00%
Results	
Estimated Sample Size (n)	385

From historical data, or a small pre-sample, or an estimate (if unknown, p=0.5 can be used to produce a conservative (worst case) estimate of sample size)

Sample Size Practice

- Pause the video and use SPC XL to answer these questions. You can compare your responses to the correct answers on the next slide when you resume the video
- You wish to collect data from a survey that has two types of questions. For each of the question types below, you wish to achieve the given error with 95% confidence. What sample size is required?

1) Several questions will be scaled from 1 to 10. You expect a standard deviation of about 1.5 on this scale and you want an error of no more than +/- 0.5

2) The remaining questions will be true/false. You wish to estimate the true proportion responding “true” with an error of no more than +/- 5%

Sample Size Practice: Answers

- Here are the correct responses to the previous questions

1) Several questions will be scaled from 1 to 10. You expect a standard deviation of about 1.5 on this scale and you want an error of no more than +/- 0.5

Sample Size to Estimate the Mean of a Normal Distribution	
User defined parameters	
Estimated Standard Dev	1.5
Half Interval Width	0.5
Confidence Level	95.00%
Results	
Estimated Sample Size (n)	35

2) The remaining questions will be true/false. You wish to estimate the true proportion responding "true" with an error of no more than 5%

Binomial Sample Size	
User defined parameters	
Proportion defectives (p)	0.5
Half Interval Width	0.05
Confidence Level	95.00%
Results	
Estimated Sample Size (n)	385

Key Takeaways

- As a review, you may want to pause the video at this point and summarize the key learnings from this session, at least from a high-level view. When you are finished, resume the video.

Key Takeaways

- The practical importance of the Central Limit Theorem is that when we are working with averages, we can use the normal distribution, no matter the distribution of the individual data points
- Sampling a population allows us to determine the parameters of the entire population within an error bound and a specified confidence interval
- A confidence interval allows us to use the sample data to generate an uncertainty around an estimate with a specified level of confidence
- A population is the entire set of products, people, or parts
- A sample is a subset of the population
- Sample size is affected by level of confidence, variation of the population, and error interval width
- Confidence interval width is affected by level of confidence, variation of the population, and sample size used

Supplemental Material

- Suggested Reading:
  - ***Basic Statistics – Tools for Continuous Improvement*** by Kiemele, Schmidt and Berdine, 4th edition (Chapter 5)
  - ***Design for Six Sigma: The Tool Guide for Practitioners*** by Reagan and Kiemele (pp. 63 – 68, 263 - 267)

- SPC XL™ software training tutorials:
  - https://airacad.com/our-insights/training-videos/spc-xl/

- The data files for this session can be downloaded from the site where you are accessing this course

Additional Practice / Review Questions

1) Assume when sampling a continuous population, we sampled 16 data points and the resulting confidence interval for the mean was (14.0 - 17.2). If we sampled 64 data points instead of 16 (assuming no change in the sample average and sample standard deviation), how would this affect our confidence interval?

2) Random sample of 100 bank customers were asked to rate their satisfaction with the bank on a scale from 0 to 10, where 10 represented perfectly satisfied, Of the 100 respondents, the average rating was 6.4 with a standard deviation of 2.1. Generate a 95% confidence interval for the true population mean for customer satisfaction.

3) The average cycle time for a heat treat process is 79.5 minutes with a standard deviation of 11.2 minutes based on a sample of 18 cycles. Generate a 99% confidence interval for this heat process.

4) For the heat treat process above, what sample size would be required to generate a confidence interval with a half-width of 5?

5) You are the quality engineer in a company that makes computer chips. You sample 120 chips randomly from a lot and discover 12 defective chips. Find the 95% confidence interval for the true proportion of defective chips.

Additional Practice / Review Questions

6) What is the appropriate sample size to generate a confidence interval of +/- 3% for the chip engineer in the previous question. Assume a defect rate of no more than 0.10 or 10% and 95% confidence.

7) In the previous question, if the historical defect rate was unknown, what would the sample size be?

We can help…
Connect With Us

Remote Project Coaching

There are times when help outside your organization is needed. When that time comes, benefit from a partner that is experienced, tested, and trusted.

Expert coaching is one of the Top Five Best Practices for generating step change in project execution, as well as enhanced return on investment. We can work remotely with your organization to provide coaching support.

Air Academy Associates
Phone: (719) 531-0777
Email: aaa@airacad.com
https://airacad.com/
https://sixsigmaproductsgroup.com/

Download on the App Store

There's an app for that!
Six Sigma Quick Tools