Foundations for Inferences

Sravani Vadlamani

QMB

09/17/2020

Review

- Transformations
- Z-Scores
- Normal Distributions
- Standard Normal Distribution
- Z-Table

Parameter Estimation

Population vs Sample

- Total votes for a presidential candidate
- Random polling data from random states

- Number of deaths from lung cancer related to smoking
- Registry deaths of lung cancer from random hospitals

- Number of volumes within the Library of Congress
- Volumes within random metropolitan libraries

- Average “grit” score for all students at Florida Poly
- Mean “grit” score from random students within Mechanical Engineering

Parameter Estimation

- We are often interested in Population Parameters
- Census or complete populations are difficult (or impossible) to collect.
- We therefore use Sample Statistics as Point Estimates for the unknown population parameters of interest.
- Best point estimate of our Sample statistics is the Mean
- Sample statistic vary from sample to sample – why?

Parameter Estimation

- Sample statistics vary from sample to sample because every sample from a population (if random) is not exactly like the last sample.

Parameter Estimates

- Quantifying how sample statistics vary provides a way to estimate the Margin of Error associated with our point estimate.

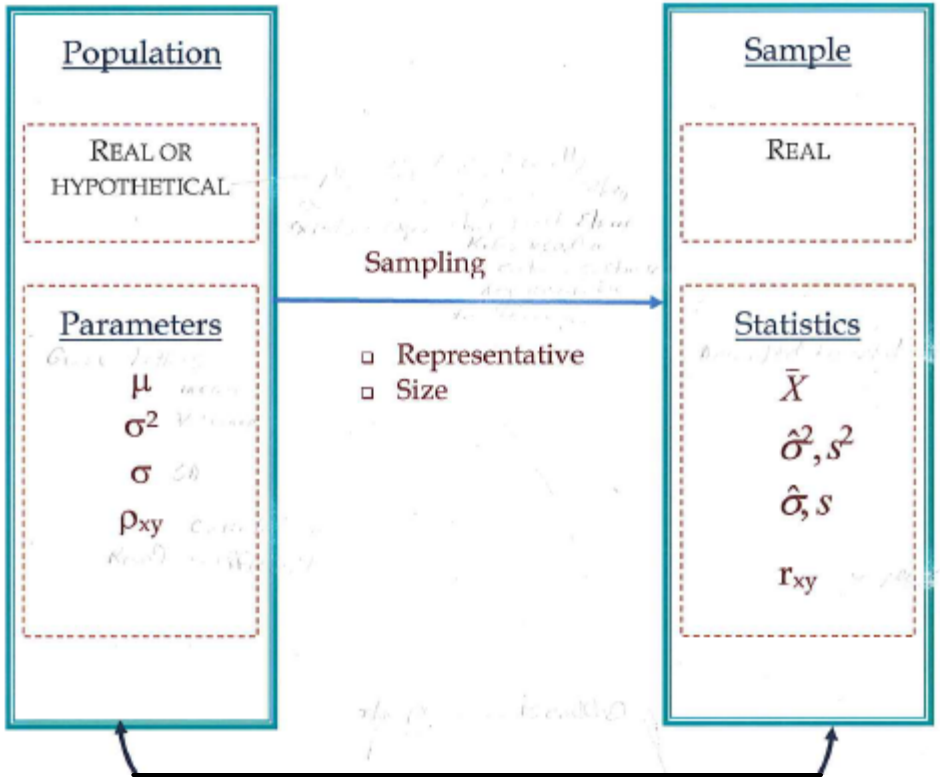
Suppose we randomly sample 1,000 adults from each state in the U.S.

Would you expect the sample means of their heights to be the same, somewhat different, or very different?

Process of Research

Hypothetical : medical experiments, software on all game users, etc.

Can't study all because not enough resources



Estimation or Inference

We want to not only understand the sample, but what the sample tells us about (or infers) about the population.

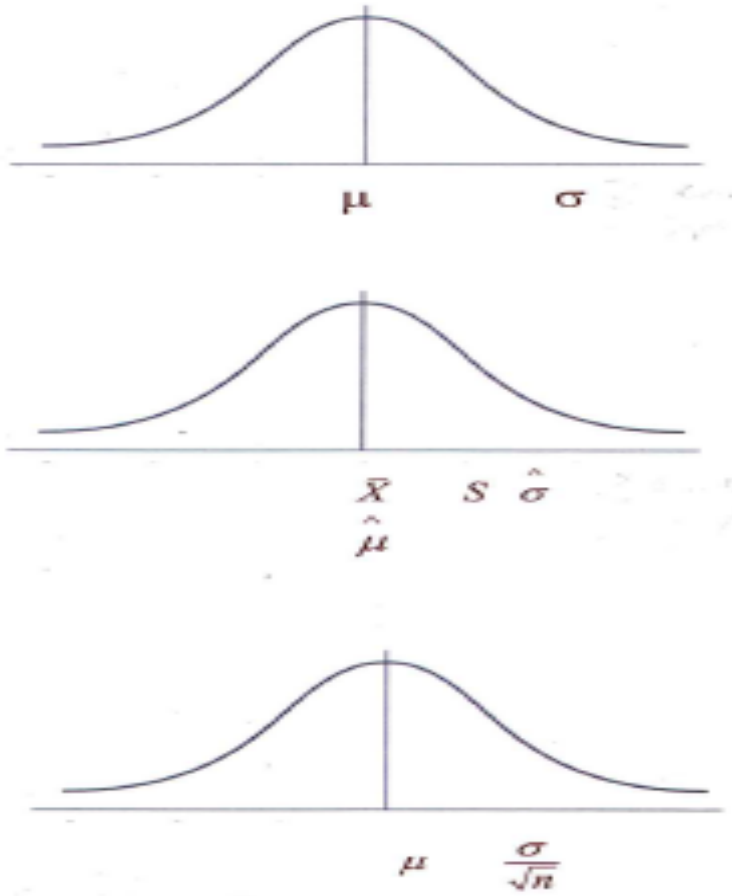
- How representative &
- How large is the sample?

Distributions

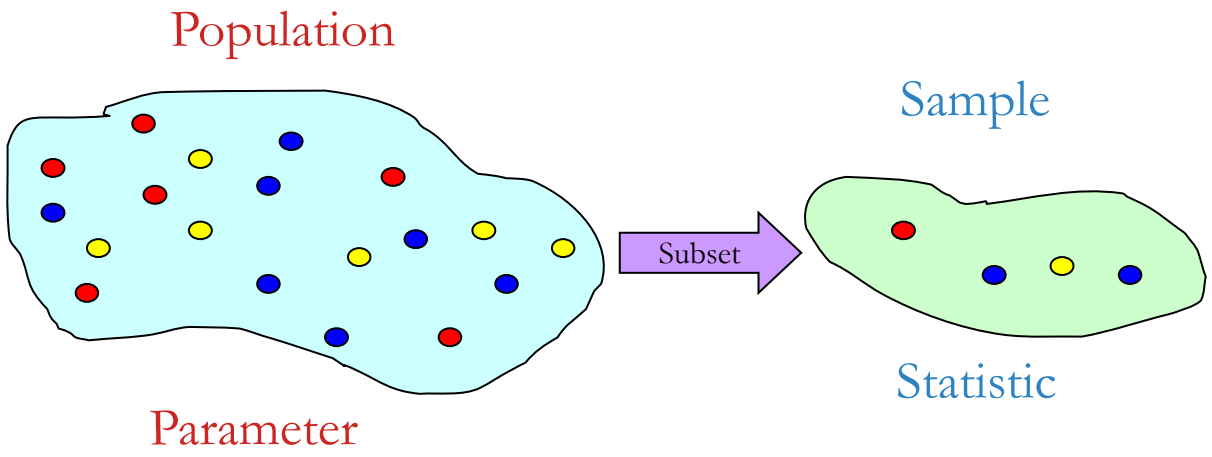
1. Population; what we really care about

2. Sample; the data we typically see

3. Sampling distribution; many samples from the population



Sampling Distribution



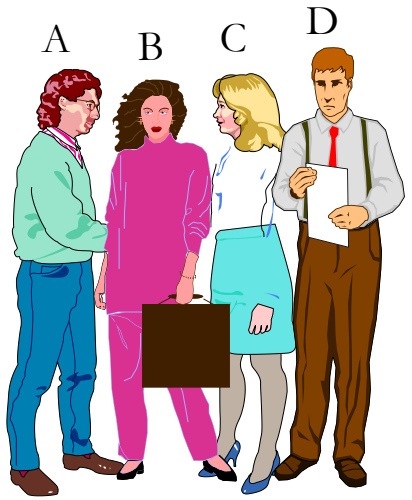
The sampling distribution of a statistic is its probability distribution.

Sampling distribution mean is **expected value.**

The **standard error** of the statistic is the standard deviation of its sampling distribution.

Developing a Sampling Distribution

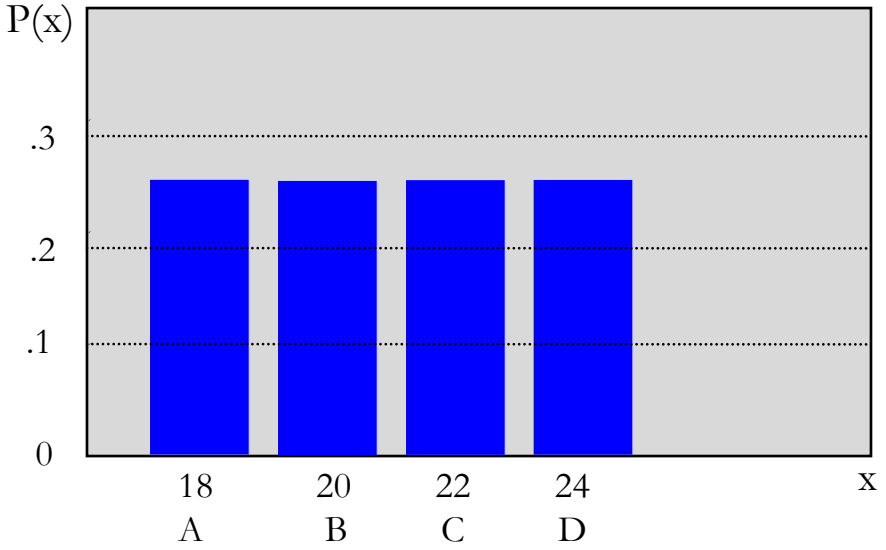
- **Assume there is a population …**
- Population size N=4
- Random variable, X, is age of individuals
- Values of X: 18, 20, 22, 24 (years)



$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Developing a Sampling Distribution

Now consider all possible samples of size n=2

1st Obs	2nd Observation			
	18	20	22	24
18	18, 18	18, 20	18, 22	18, 24
20	20, 18	20, 20	20, 22	20, 24
22	22, 18	22, 20	22, 22	22, 24
24	24, 18	24, 20	24, 22	24, 24

16 possible samples (sampling with replacement)

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

16 Sample Means

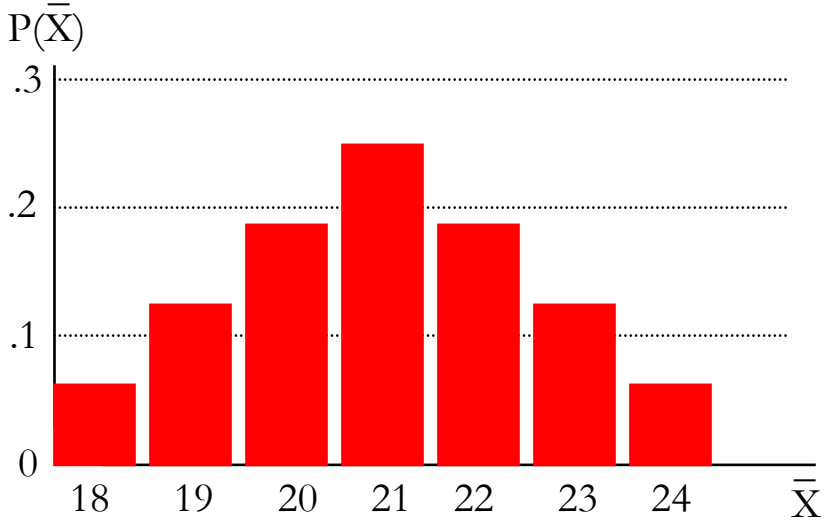
Developing a Sampling Distribution

Sampling Distribution of All Sample Means

16 Sample Means

1st	2nd Observation			
Obs	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution

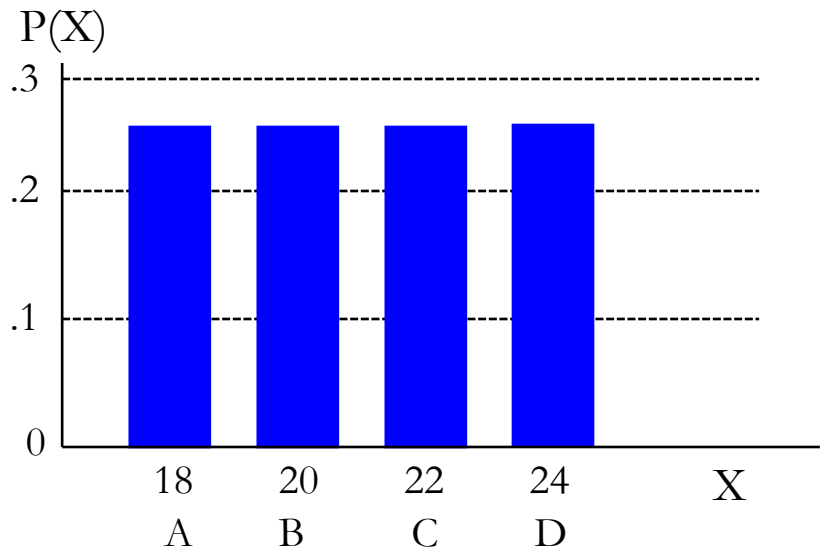


(no longer uniform)

Developing a Sampling Distribution

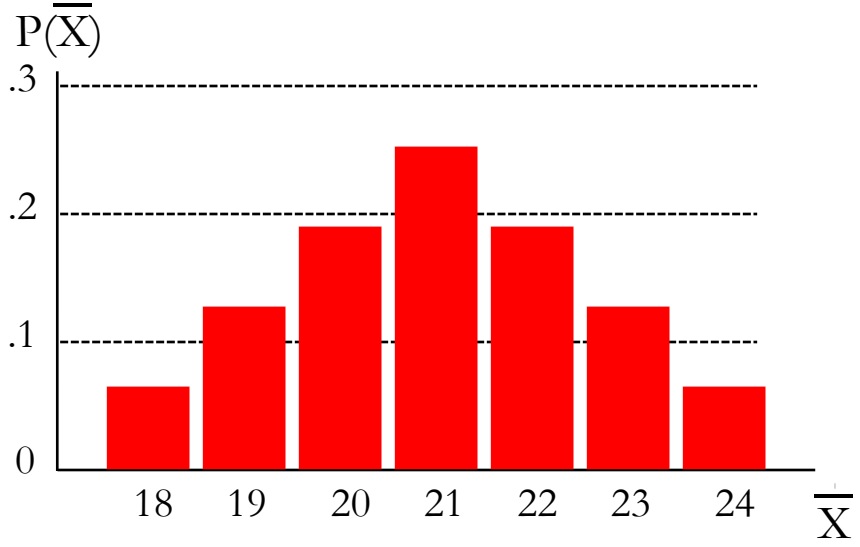
Population

$\mu = 21 \quad \sigma = 2.236$



Sample Means Distribution

$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$



Sampling Distributions

- Any statistic calculated on a sample has a certain sampling distribution.

- Imagine drawing a sample of size “n” from the defined population and calculating the statistic of interest (e.g., the sample mean), then drawing another sample of size “n” from the same population and calculating the statistic again. Repeat this process many times.

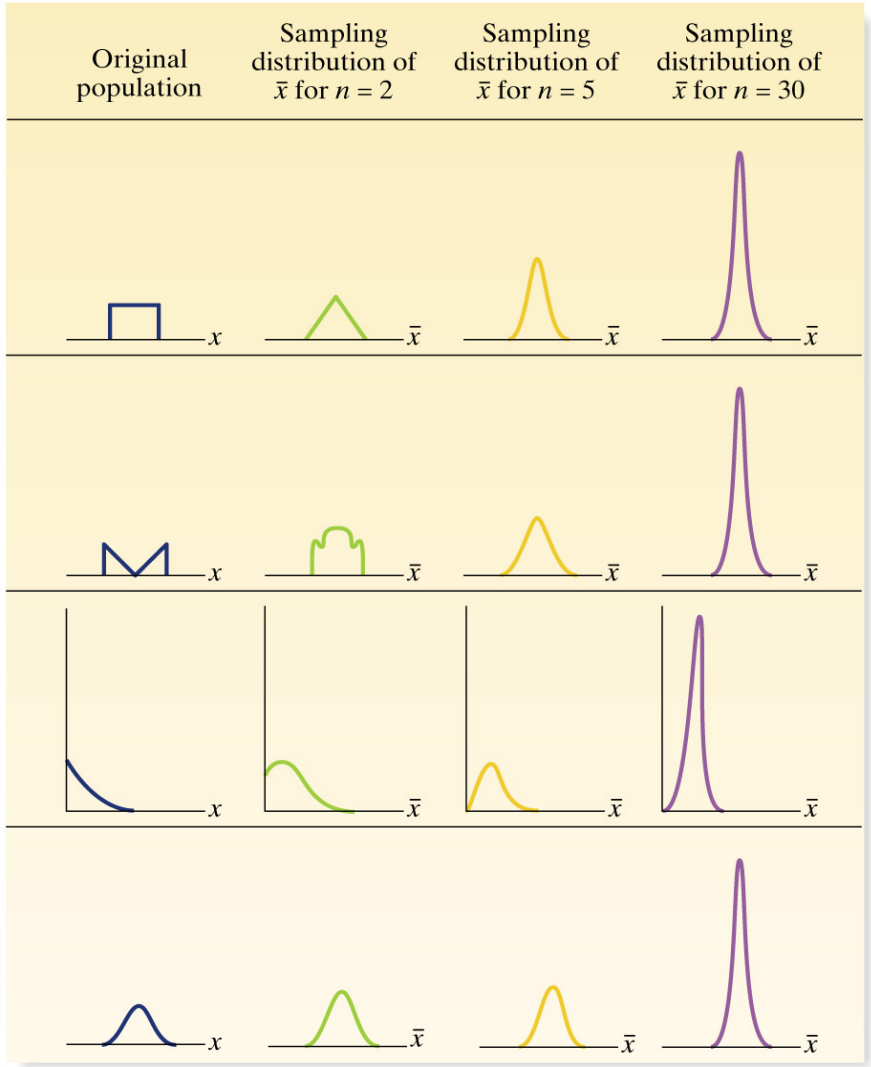
- We could construct a frequency distribution of the sample statistics – this is the values of the statistic that we observed and how often those values were observed. This frequency distribution is called a sampling distribution.

Central Limit Theorem

- The sampling distribution of the mean of a random sample drawn from any population is ***approximately normal*** for a ***sufficiently large*** sample size ($n \geq 30$).
- The larger the sample size, the more closely the sampling distribution of X will resemble a normal distribution.

If a random sample of n observations $X_1, X_2, .., X_n$ is drawn from a population with finite mean μ and variance σ^2, then when n is sufficiently large, the sampling distribution of the sample mean X can be approximated by a normal density function.

Central Limit Theorem



Central Limit Theorem - Rules

Describes the sampling distribution of the mean:

1. $\mu_{\bar{x}} = \bar{X}$

The mean of the sampling distribution is the same as the mean of the population (or the expected value of the sample mean is the same as the population mean).

2. $Б\bar{x} = \frac{б\bar{x}}{\sqrt{n}}$ What do we call this?

The variance of the sampling distribution is proportional to the variance of the population, and inversely related to the sample size (as n increases SE decreases)

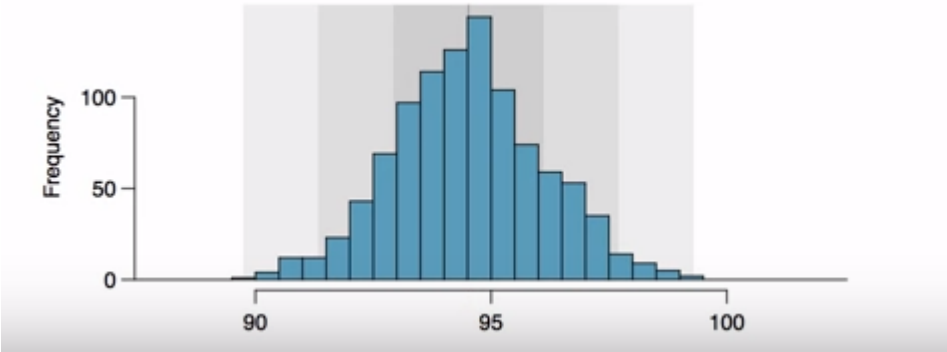
3. As n ↑ the sampling distribution approaches the shape of a normal distribution, regardless of the shape of the population distribution.

Central Limit Theorem

- As the sample size increases we would expect the samples to yield more consistent sample means, hence the variability among the sample means would be lower.
- CLT Conditions
- All 3 rules must apply but not always easy to verify
- Independence: Sampled observations must be independent. This is more likely if
  - Random sampling/ assignment is used and
  - If sampling without replacement, n < 10% of the population.
- Sample size (pop. skewness): Either the population is normal, or if the population is skewed, the sample size is large
  - The more skewed the population distribution, the larger sample size we need for the CLT to apply
  - For moderately skewed distributions n > 30% is a widely used rule of thumb

Sampling Distribution

- The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population.
  - Example: Random sample of 100 individuals to run 10 miles. Average time to run 10 miles is:
    - 95.61 (take another random sample of 100 individuals)
    - 95.30 (another)
    - 93.43 (and another)
    - 94.16 (and we do this many times…)
    - We get a sampling distribution that represents the distribution of the point estimates based on the samples of the population

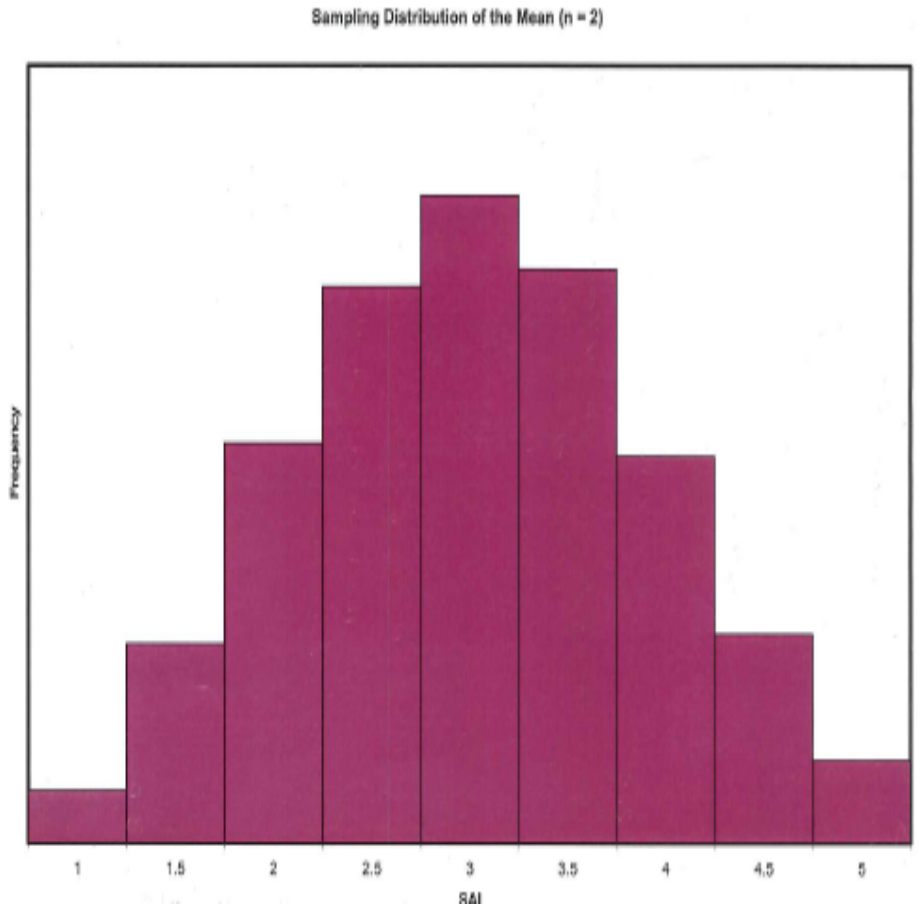


Variability in Estimates

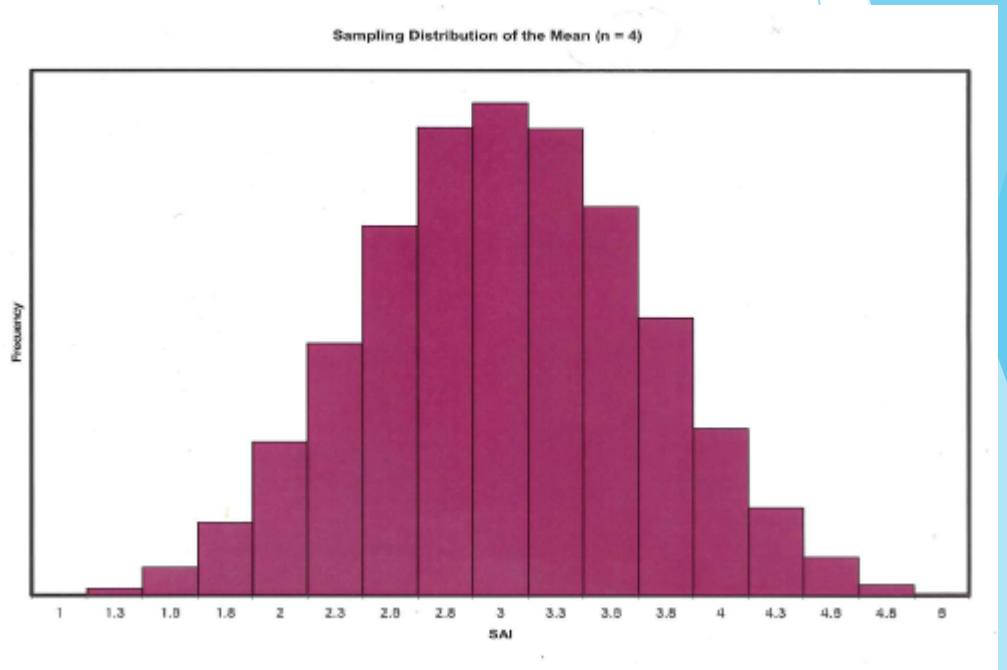
- The mean of this distribution is 94.152 . There is some variability around the mean. SD = 1.59 minutes. The standard deviation of the sample mean tells us how far the typical estimate is away from the actual population mean; 94.52 minutes.

  - It also describes the typical error of the point estimate, and for this reason we usually call this the Standard Error (SE).
  - Given n independent observations from a population with a standard deviation б, the standard error of the sample mean is equal to $SE = \frac{б}{\sqrt{n}}$

100 Samples – Calculate Means of each sample – SAI scores

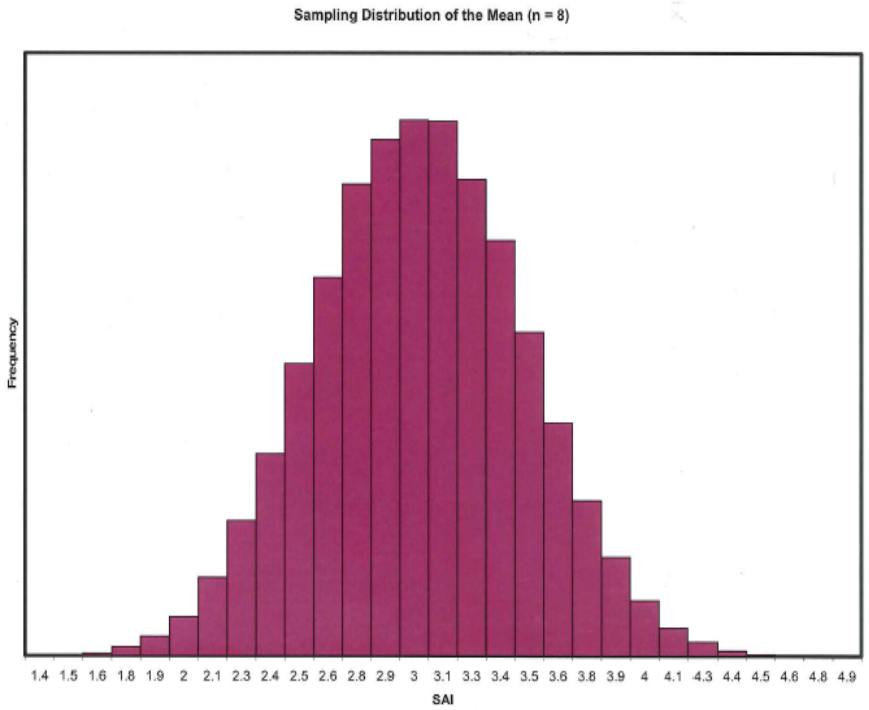


Each sample; n = 2

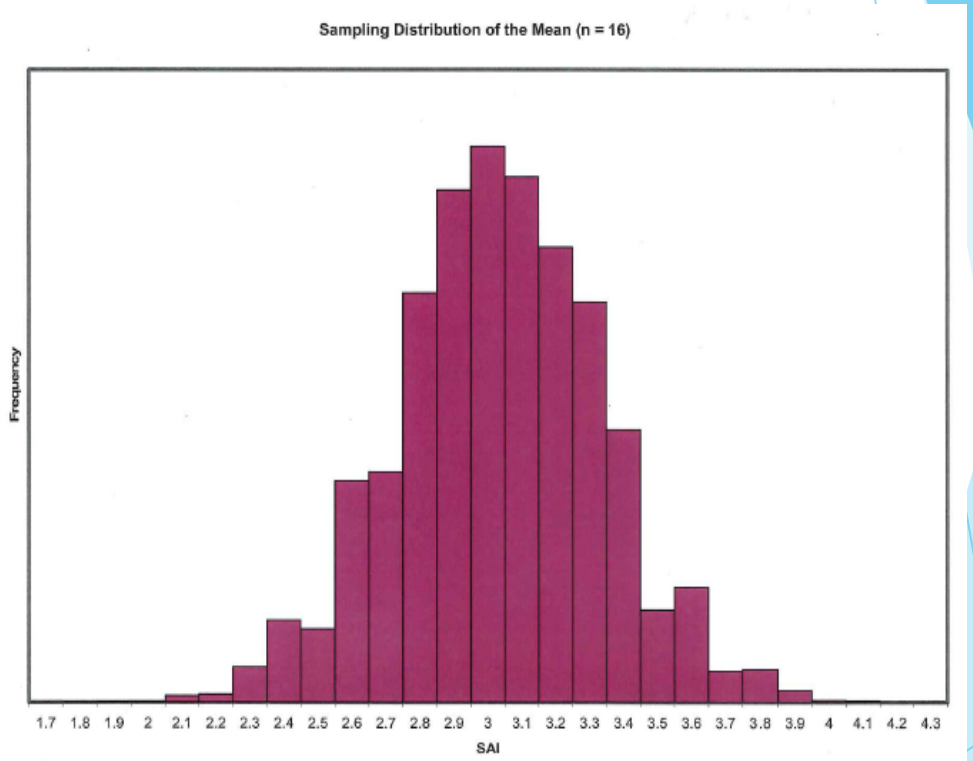


Each sample; n = 4

100 Samples – Calculate Means



Each sample; n = 8



Each sample; n = 16

100 Samples – Calculate Means

Sample size	$\bar{X}$ Typical Value	$\hat{\sigma}_{\bar{x}}$ Dispersion	Skewness Shape	Kurtosis
2	3.02	0.90	0.04	-0.53
4	3.02	0.64	0.02	-0.28
8	3.02	0.45	0.02	-0.15
16	3.02	0.32	0.01	-0.08
"Population"	3.02	1.28	0.06	-1.05

- SD; as we see the sample size increasing we see the SD getting smaller (definition of standard error of the mean)
- As the sample size increases, the precision increases
  - Bigger samples yield more precise inferences
  - Skewness and Kurtosis gets closer & closer to 0 – or a normal distribution

$$SE = \frac{\sigma}{\sqrt{n}}$$

Two Aspects of Inference that Depend on Sampling

- Accuracy: Representativeness (biased vs. unbiased)
  - How representative has nothing to do with the size of the sample
- Precision: Sample size (bigger samples yield more precise inferences)

If one has a large sample – Must be accurate (true or false)

- I am 95% sure that $75 < \mu < 125$ (less precise)
- I am 95% sure that $95 < \mu < 105$ (more precise)

What if your sample is biased – both accuracy and precision could be wrong…

Examples

Example

The standard deviation of student weights for a sample of 100 students is 5.65 kgs. What is the standard error if the sample mean is 65.5 kg.

Example

The standard deviation of student weights for a sample of 100 students is 5.65 kgs. What is the standard error if the sample mean is 65.5 kg.

$$SE = \frac{s}{\sqrt{n}} = \frac{5.65}{\sqrt{100}} = 0.565$$

Example

The distribution of the number of eggs laid by a certain species of hen during their breeding period is 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times and build a distribution of the sample means.

1. What is this distribution called?
2. Would you expect this distribution to be symmetric, right skewed or left skewed? Explain your answer.
3. Calculate the variability of this distribution and state the appropriate term used to refer to this value.
4. Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample of mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

Confidence Intervals

- A point estimate provides a single plausible value for a parameter.
  - It's rarely perfect (usually some error in this estimate).
- To account for this error we provide a plausible range of values for the parameter.
- Called a confidence interval (range likely to better hit the parameter rather than one estimate).
- Using just a sample statistic to estimate a parameter is like fishing in a murky lake with a spear. Using a confidence interval is like fishing with a net (much better chance of catching the fish).
- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Confidence Interval

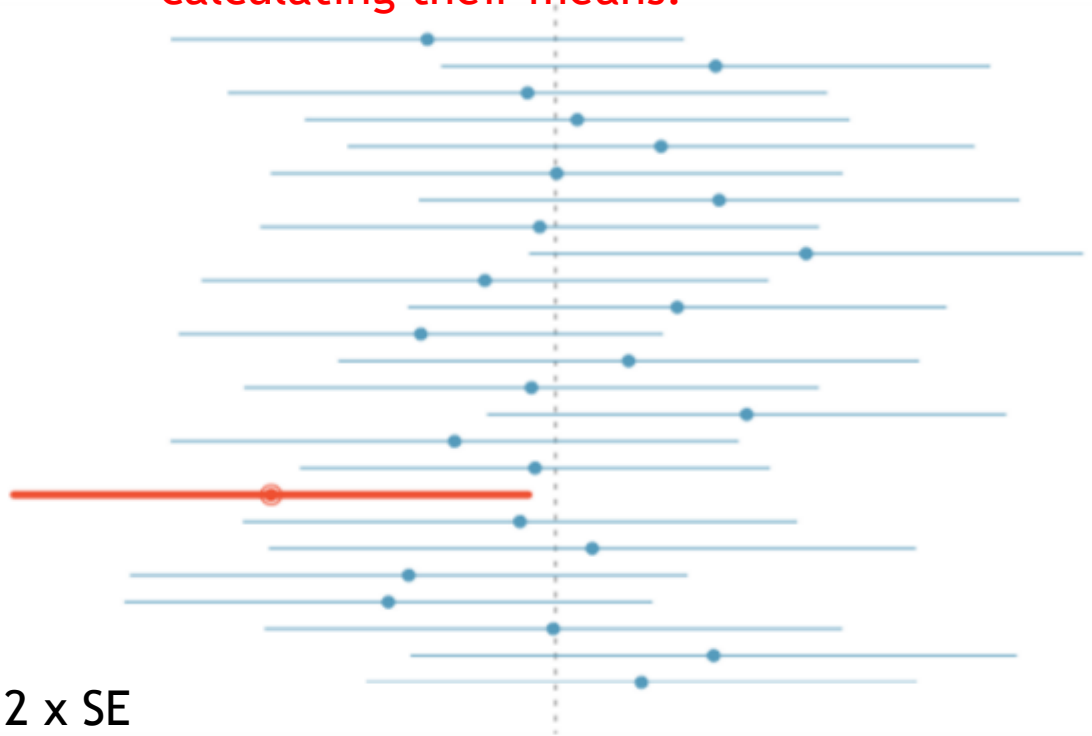
- Begin with 95% CI
- A plausible range of values for the population parameter is called a Confidence Interval.
  - The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.
  - The standard error (SE) represents the SD associated with the mean estimate.
  - If the interval spreads out 2 SD from the point estimate, we can be roughly 95% confident that we have captured the true parameter:
- Point estimate +/- 2 x SE

What does 95% Confidence Interval Mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* ± *2 x SE*.

Then about 95% of those intervals would contain the true population mean (μ).

Runners running a 10 mile stretch and calculating their means.



The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.

Or one is outside +- 2 x SE

Width of an Interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see the drawbacks of using a wider interval?

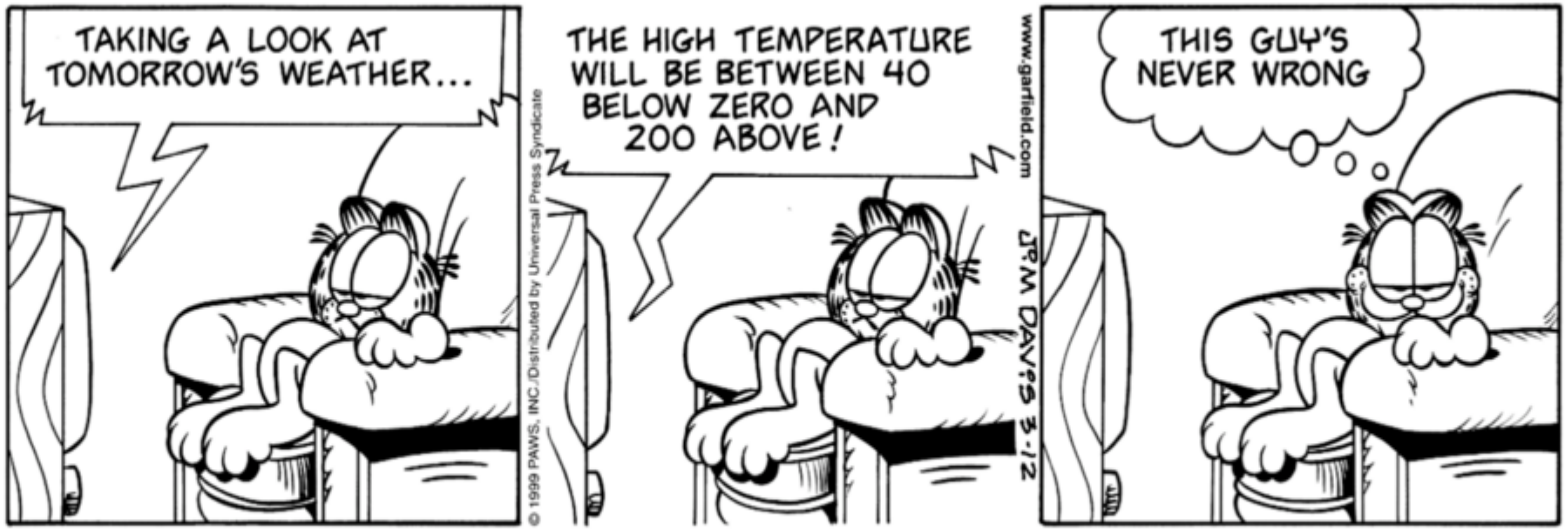


Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the Confidence Level (CI and CL are used interchangeably)

Point estimate +/- z* x SE (z* is called the z critical)

- In a CI, z* x SE is called the margin of error
  - The margin of error changes as the confidence level changes.
  - In order to change the confidence level we need to adjust the z* in the above formula.

- Commonly used confidence levels in practice are 90%, 95%, 98% and 99%.
- For a 95% confidence interval, z* = 1.96.
  - Where does this come from?

- Using the standard normal (z) distribution – can find the appropriate z* for any confidence level.

Practice

- Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

- A. Z = 2.05

- B. Z = 1.96

- C. Z = 2.33

- D. Z = -2.33

- E. Z = -1.65

98% Confidence Interval

Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

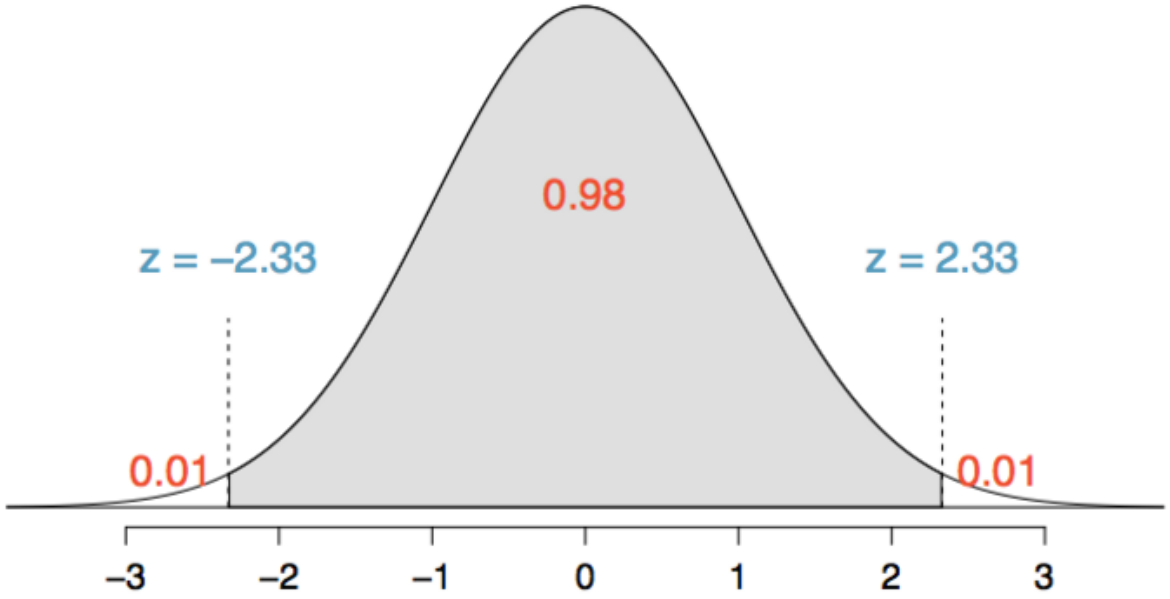
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

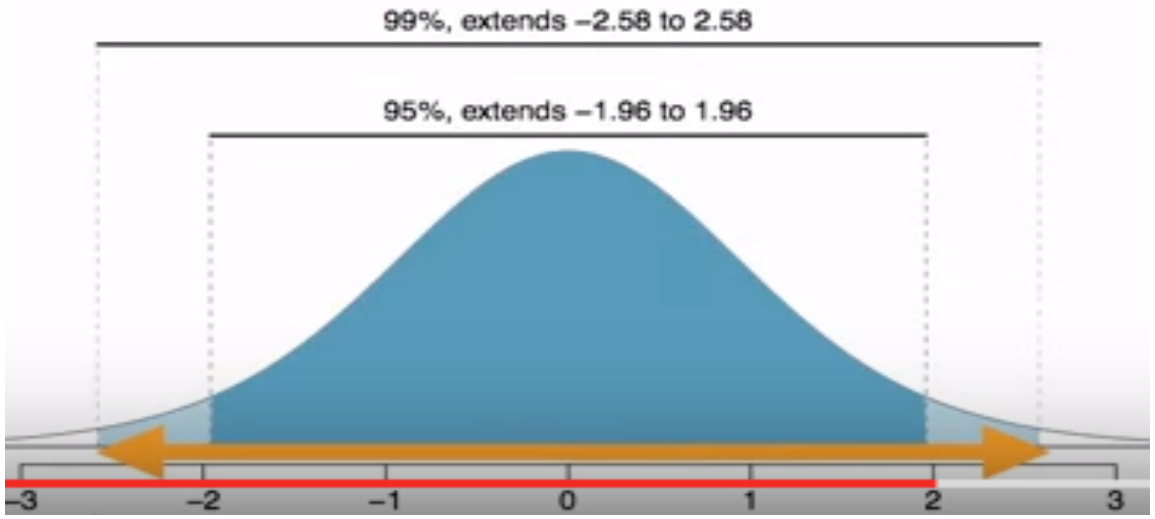
(d) Z = -2.33

(e) Z = -1.65



Changing the Confidence Level

- To create a 99% confidence level, we must widen the 95% interval.
- On the other hand, if we want an interval with lower confidence, such as 90%, we could make our original 95% interval slightly slimmer.



Important Conditions for Confidence Levels

- CL works when the sample distribution is nearly normal and the estimate of the standard error is sufficiently accurate:
  - This happens when…

- The sample observations are independent (random).
  - If the sample contains fewer than 10% of population
  - Simple random sample
- The sample is large: n > 30 is a good rule of thumb.
- The distribution of the sample observations is not strongly skewed.

Interpreting Confidence Intervals

- We are XX % confident that the population parameter is between the lower bound and upper bound of our confidence interval.

- Incorrect to say: “Confidence interval captures the population parameter with a certain probability.”
  - The CI only quantifies how plausible it is that the parameter value is in the interval.
- Another important consideration:
- Confidence intervals only try to capture the population parameter. Our intervals say nothing about the confidence of capturing:
  - individual observations
  - A proportion of the observations
  - Point estimates

Examples

- You are interested in measuring the number of hours students spend watching TV on weekdays. You collected 5 different samples and each sample has 40 students.

Sample	Mean (μ)	SD (σ)
1	5	1.2
2	6	1.1
3	4	0.8
4	3	0.5
5	8	1.6

- Calculate the estimate for the average time students watch TV
- You are not very sure about sample 5. Compute the 95% confidence interval for sample 5.

Review of Inference Foundations to This Point

- The standard deviation of student weights for a sample of 100 students is 5.65 kgs. Calculate a 60%, 75% and 95% confidence intervals for the average weight of students if the sample mean is 65.5 kg.
-

Review of Inference Foundations to This Point

- Point Estimate
  - Sample mean is a point estimate of the population mean
- Estimates generally vary from one sample to another. This sampling variation may be close to the population parameter, but it's not exact.
- The standard deviation associated with the estimate is called?
  - It describes the typical error or uncertainty associated with the estimate.

    Or
  - Point Estimate +/- 2 Standard Error

- So; mean = 3.2, SD = 1.74 n = 50

- SE =

Review of Inference Foundations to This Point

- So, with a SE of .25 we could construct a Confidence Interval
  - Mean +/- 2 SE 3.2 +/- 2 x .25 →
    - (2.7, 3.7)

- A more accurate interval uses the Z critical (Z score at what Confidence Level the researcher is seeking.

- Point Estimate +/- Z* x SE (SE is now called a Margin of Error)

- CI of 95%: Z* = 1.96 (not 2 SD) – more accurate

Review of Inference Foundations to This Point

- Confidence Interval; plausible range of values for the population parameter.
- 95% CI – roughly 95% of the time the estimate will be within 2 standard error (deviations) of the parameter.
  - Point estimate +/- Z* x SE
  - Draw it out
- Changing the CI: 99%
  - Margin of error: Point estimate +/- Z * SE
  - Draw it out
- Correct interpretation: we are XX% confident that the population parameter is between…