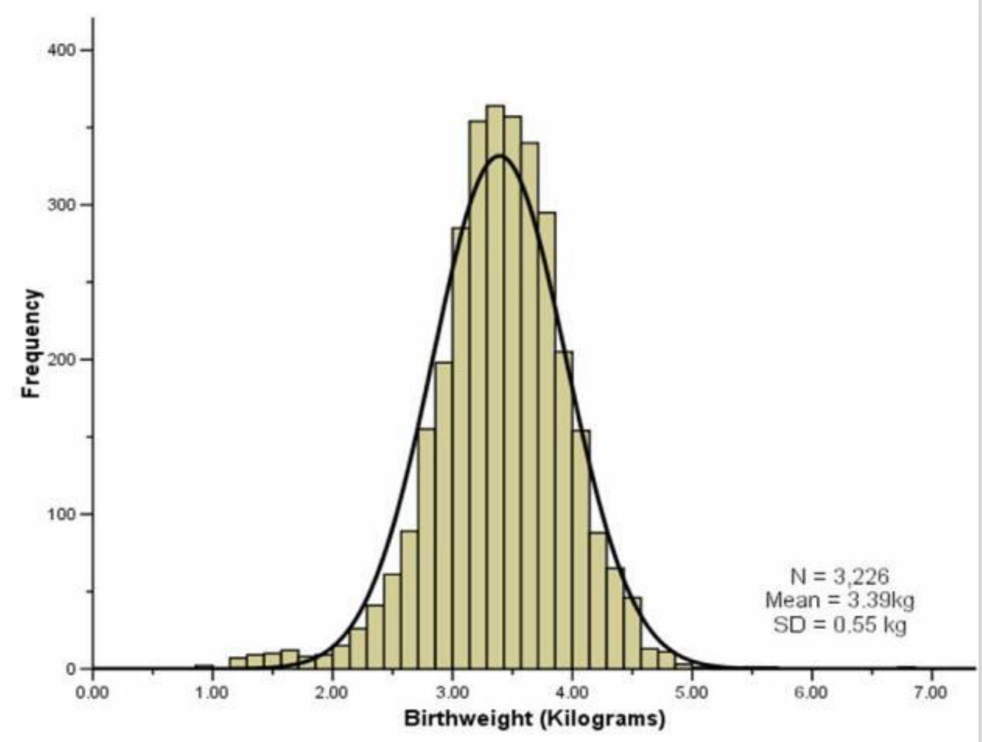
CONFIDENCE INTERVALS

MAT 152 – Statistical Methods I
Lecture 1
Instructor: Dustin Roten
Fall 2020

Reviewing the Central Limit Theorem



https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/statistical-distributions

- Let X = be the birth weight of a baby (in kilograms)

- Statistics for an individual:
  - Mean = μ_X
  - Standard Deviation = σ_X
  - $z = \frac{x-\mu}{\sigma}$
  - (Data on the left is NOT the population)

- Statistics for a sample:
  - Mean = $\mu \approx \mu_X$
  - Standard Deviation = $\frac{\sigma_X}{\sqrt{n}}$
  - (Data on the left IS a sample)

Example

- Suppose we want to determine the average GPA of community college students at Forsyth Tech. Three samples are collected:

	Sample 1	Sample 2	Sample 3
Sample Size (n)	50	37	42
Mean ($\bar{x}$)	3.21	3.23	3.11
Standard Dev. (s)	0.21	0.20	0.26

- Each sample mean can be used to estimate the "true" population mean: 3.21, 3.23, 3.11
- The "mean of the means" can be calculated to estimate the "true" population mean: 3.18
- Each of these means is called a **point estimate** of the "true" population mean.

True population mean: 3.2

Population standard deviation: 0.25

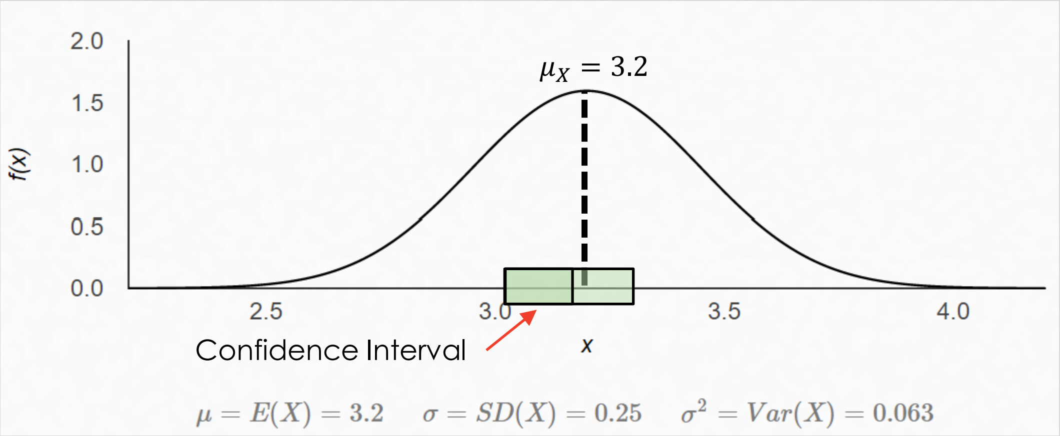
Confidence Intervals

- Using samples to estimate population parameters is called **inferential statistics**.
- **Point estimates** are “okay” at estimating population parameters but they are often not exact.

- Perhaps an **interval** can be constructed?
  - “The true population mean GPA lies between 3.1 and 3.3”

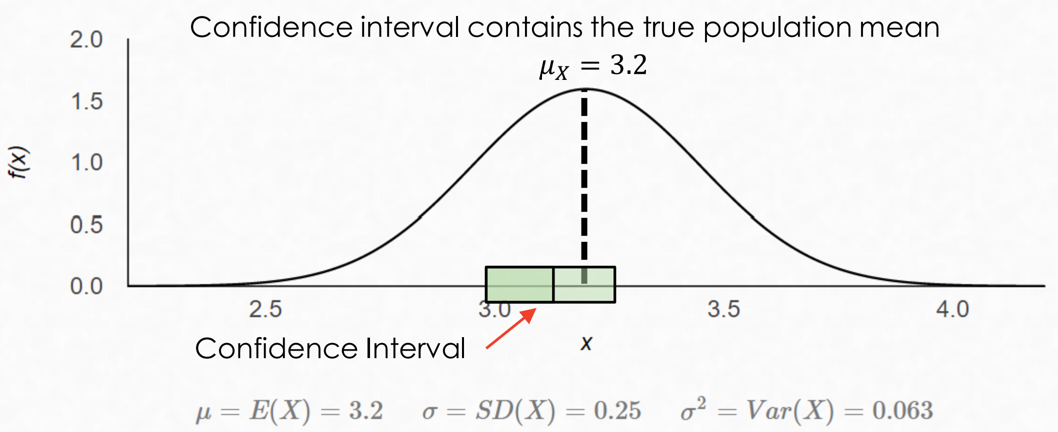
- How can an interval be constructed?

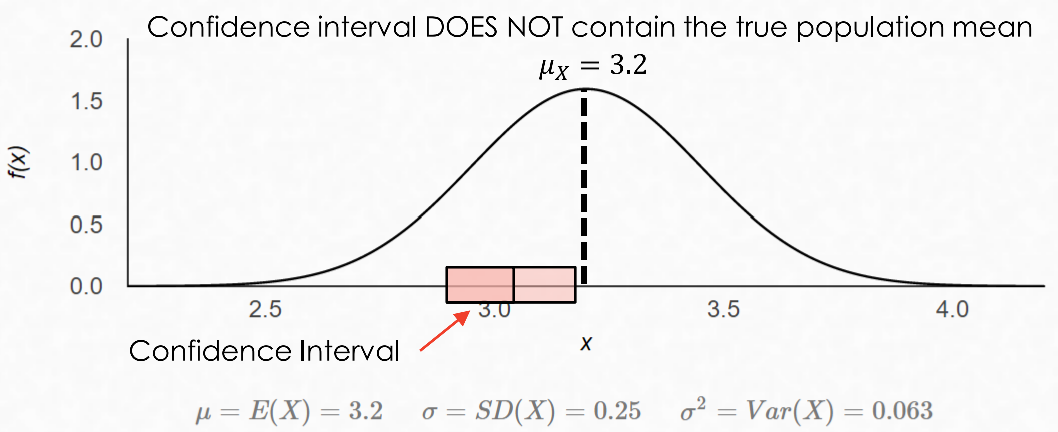
- How **confident** can we be in the interval?



Confidence Intervals

- A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers.
- It provides a range of reasonable values in which we expect the population parameter to fall.
- There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.



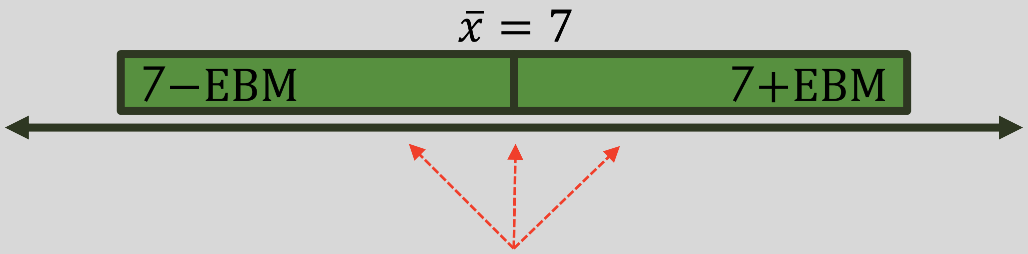


Calculating a Confidence Interval

- First, a confidence interval for a population mean will be calculated for a population with a **known** standard deviation.
- Normally, the population mean AND standard deviation are unknown. These are special cases.
- To estimate μ for a population, a sample of size n must be collected. The mean of the sample, $\bar{x}$, is a **point estimate** of μ.
- An **error bound for the population mean (EBM)** is required. This is the **margin of error**. (A measure of how close $\bar{x}$ is to μ)

Population
$\mu_X = ???$
$\sigma_X = 2$

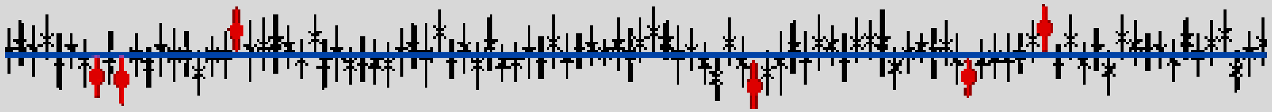
Sample
$\bar{x} = 7$ (point estimate)
$s = 2.1$ (point estimate)



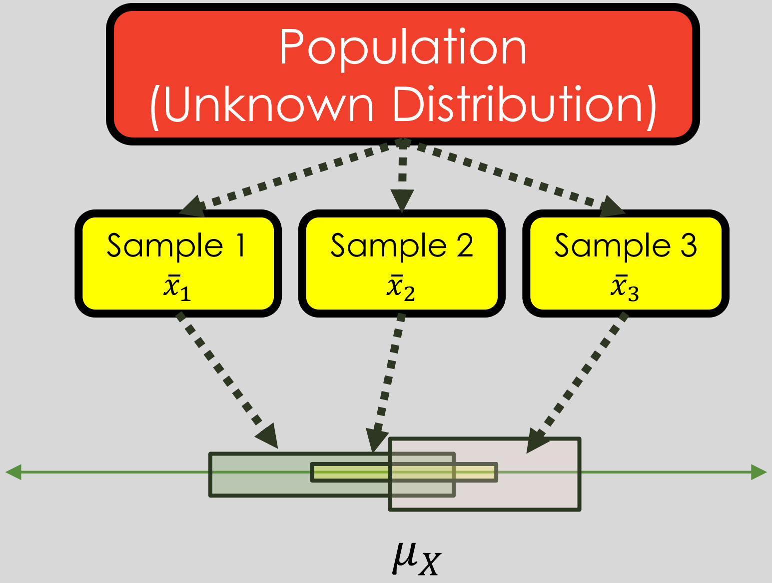
μ_X will likely fall in this range somewhere

Using the Central Limit Theorem

- The Central Limit Theorem tells us that the sample means roughly follow a normal distribution.
- Therefore, we can use the Empirical Rule to state the following:
    - 68% **Confidence Level** – 68% of confidence intervals contain the true population parameter.
    - 95% **Confidence Level** – 95% of confidence intervals contain the true population parameter.
    - 99% **Confidence Level** – 99% of confidence intervals contain the true population parameter.
    - XX% **Confidence Level** – XX% of confidence intervals contain the true population parameter.
        - (Any confidence level can be selected.)



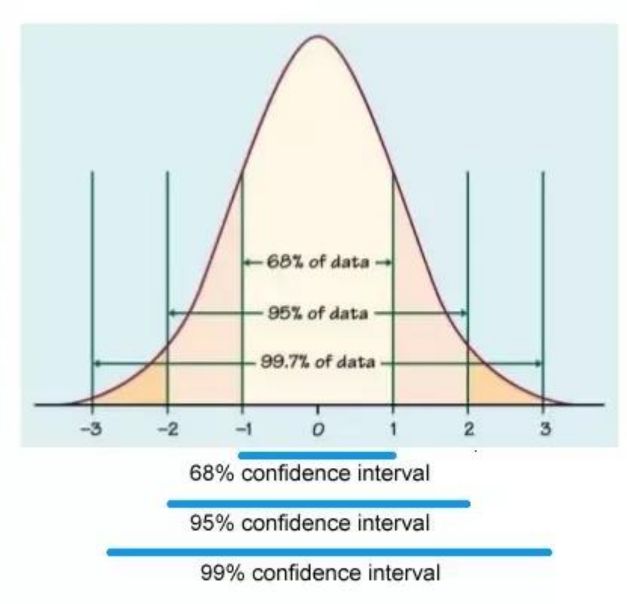
100 samples with 95% confidence intervals
(Population mean in blue)



Each sample has its own confidence interval. The population parameter will likely be within the “overlap” of these intervals.

Confidence Levels

- The Central Limit Theorem and the Empirical Rule also tell us that each **confidence level** has an associated z-score.
- Low confidence levels = small confidence intervals



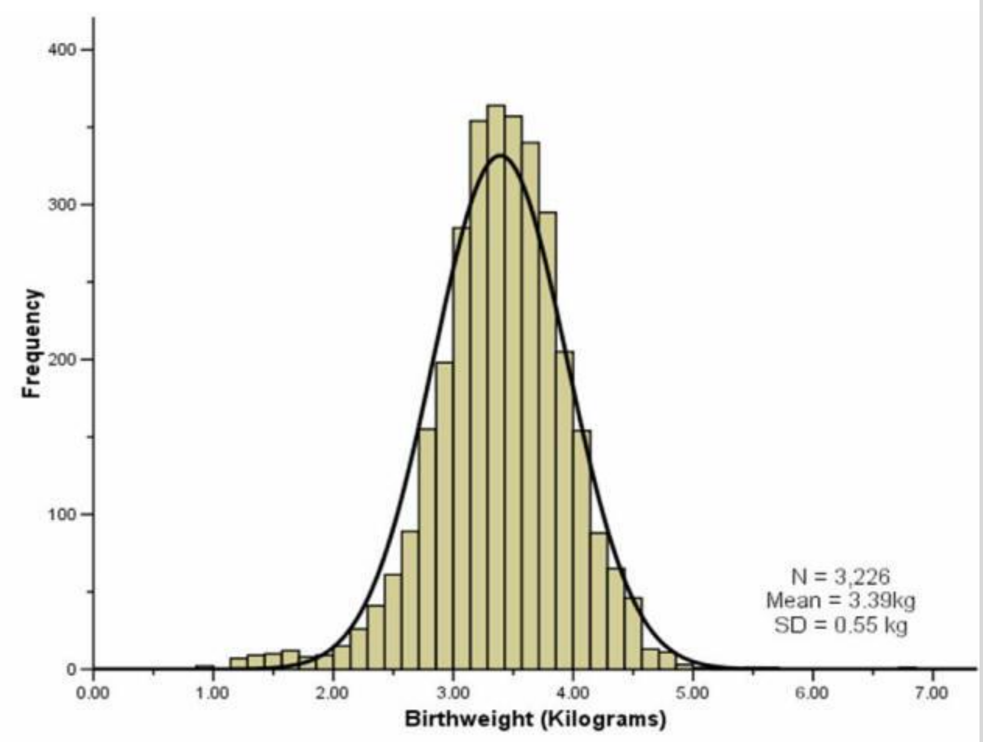
Confidence Level	Area in One Tail	Exact Z-score
50%	0.25	0.674
80%	0.1	1.282
90%	0.05	1.645
95%	0.025	1.96
98%	0.01	2.326
99%	0.005	2.576

Example

- Suppose the mean birthweight of babies born in London is a desired parameter. To estimate this values, a random sample of babies was constructed (right). Through previous study, the population standard deviation is known.
  - $\mu_X = ???$
  - $\sigma_X = 0.57\text{kg}$

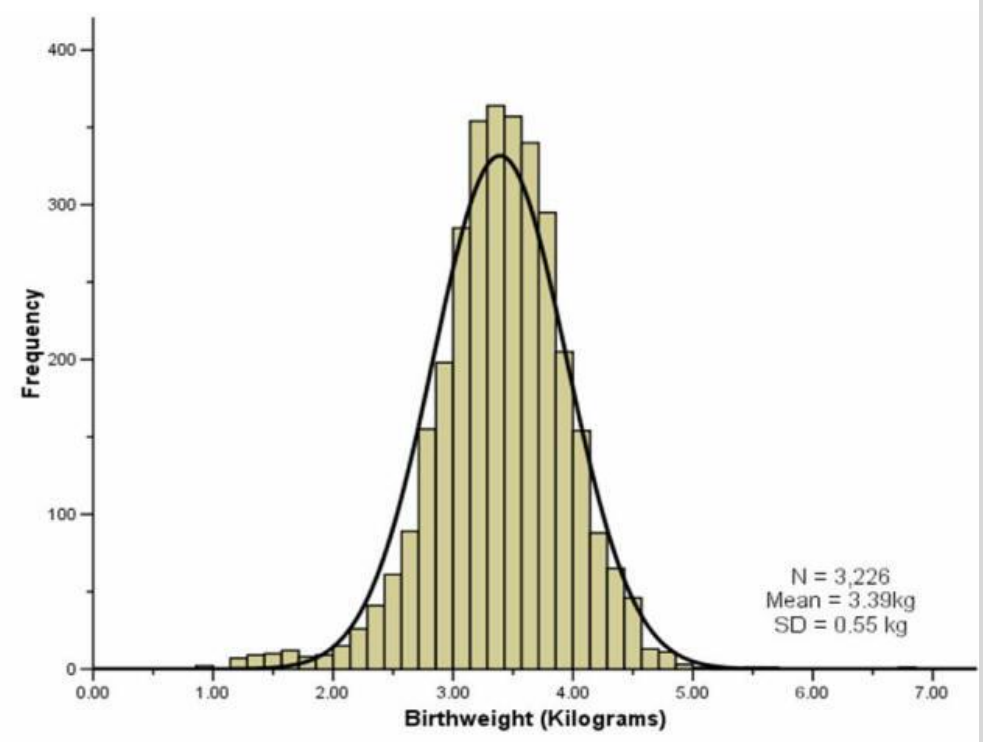
- Statistics from the sample:
  - $n = 3226$
  - $\bar{x} = 3.39\text{kg}$
  - $s = 0.55\text{kg}$

- How is the population mean estimated if a confidence level of 95% is desired?



Example (Cont.)

- 5-Step Process:
  1. Calculate the sample mean.
     (Already done! $\bar{x} = 3.39\text{kg}$)
  2. Find the z-score that corresponds to the confidence level.
  3. Calculate the error bound (EBM)
  4. Construct the confidence interval
  5. Write a sentence that interprets the estimate in the context of the situation in the problem.



Example (Step 2)

Find the z-score that corresponds to the confidence interval

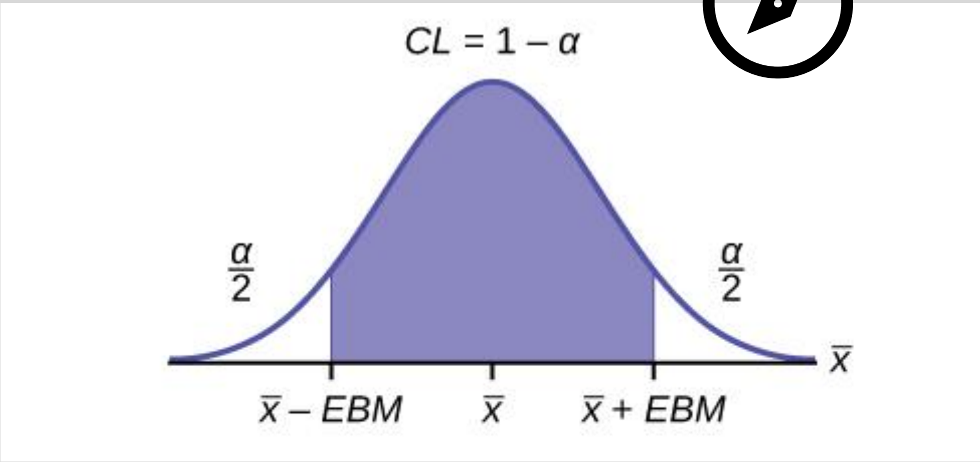
The confidence level is 95%

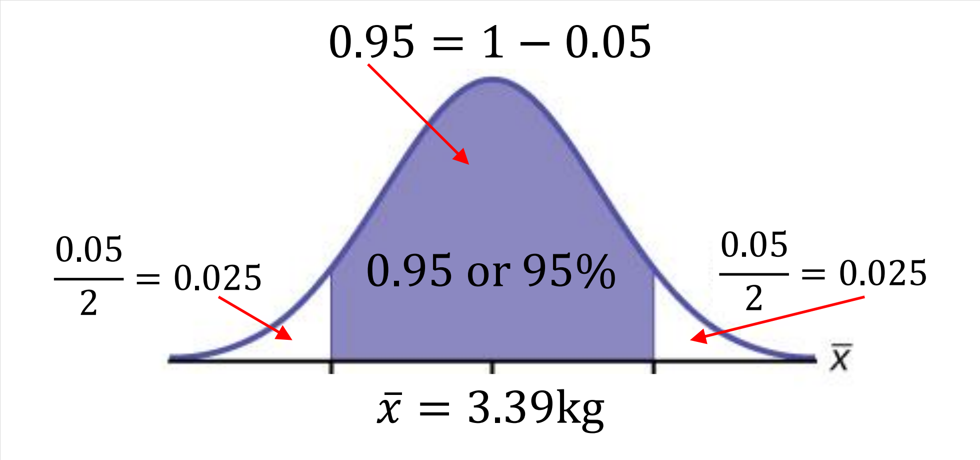
This means that the confidence interval must cover a range of 95% of the distribution.

There will be 5% left over (0.05). This is split between the two "empty" tails.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

Confidence Level	Area in One Tail	Exact Z-score
50%	0.25	0.674
80%	0.1	1.282
90%	0.05	1.645
95%	0.025	1.96
98%	0.01	2.326
99%	0.005	2.576





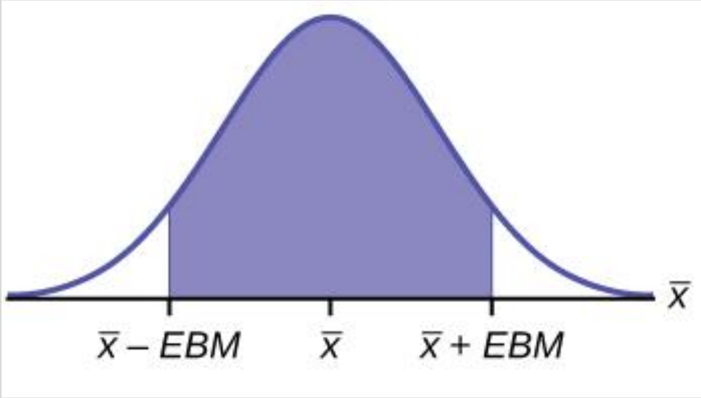
Example (Step 3)

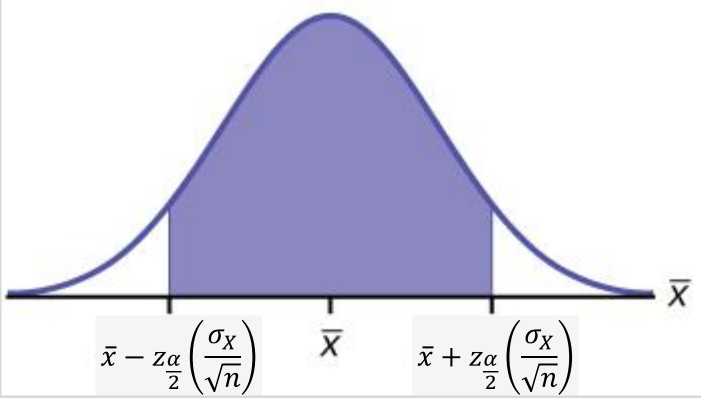
Calculate the error bound (EBM)

- Recall that the standard error is the standard deviation of a sample.
  - Population standard deviation: σ_X
  - Sample standard deviation (standard error): $\frac{\sigma_X}{\sqrt{n}}$

- The **confidence interval** that corresponds to the **95% confidence level** must contain 95% of the sampling distribution.
- The Empirical Rule states that this is $\pm 2\sigma$ of the mean. This is a z-score of 2. However, the EXACT z-score is 1.96.

$$\text{EBM} = z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 1.96\left(\frac{0.57}{\sqrt{3226}}\right) \approx 0.0197$$





Example (Step 4)

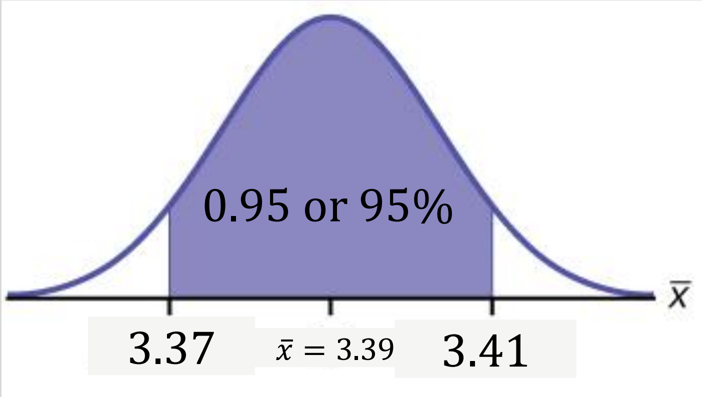
Construct the **confidence interval**

Lower bound:

$$\bar{x} - z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 3.39 - 1.96\left(\frac{0.57}{\sqrt{3226}}\right) \approx 3.37$$

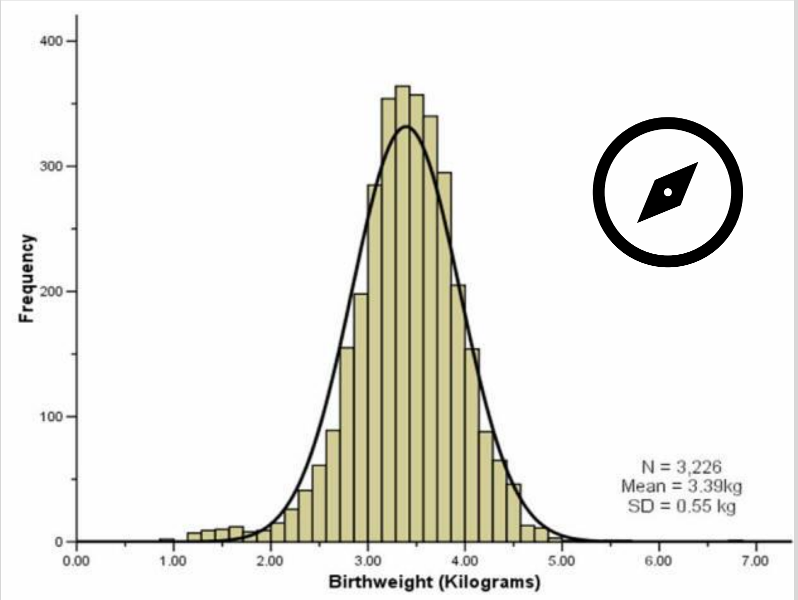
Upper bound:

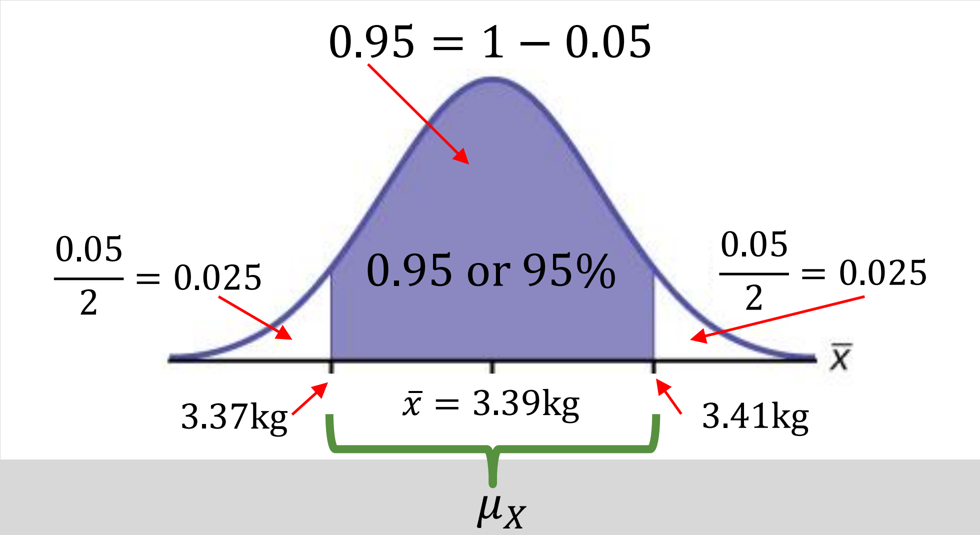
$$\bar{x} + z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 3.39 + 1.96\left(\frac{0.57}{\sqrt{3226}}\right) \approx 3.41$$



Example (Summary)

- 5-Step Process:
  1. Calculate the sample mean.
     (Already done! $\bar{x} = 3.39\text{kg}$)
  2. Find the z-score that corresponds to the confidence level.
     95% percent confidence level: $z_{\frac{\alpha}{2}} = 1.96$
  3. Calculate the error bound (EBM)
     $\text{EBM} = z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 1.96\left(\frac{0.57}{\sqrt{3226}}\right) \approx 0.0197$
  4. Construct the confidence interval
     $(3.37, 3.41)$
  5. Write a sentence that interprets the estimate in the context of the situation in the problem.
     We estimate with 95% confidence that the true population mean of birthweights is between 3.37 and 3.41 kilograms.





A Few Notes

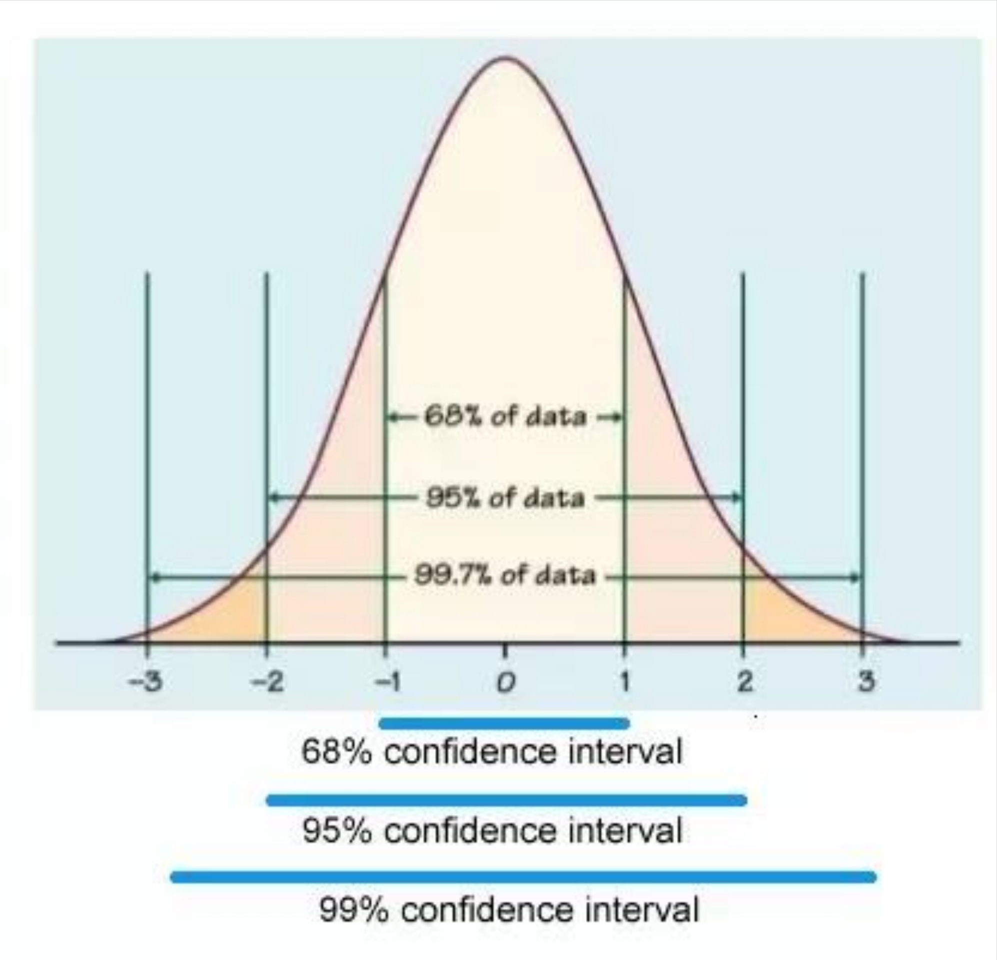
Generally, researchers want as small of confidence interval as possible to get an accurate approximation of the population mean. What if the confidence level and/or sample size changes?

The **confidence interval** can be narrowed by decreasing the **confidence level**; however, the confidence in the estimation decreases.

For example:

We can say that we are 100% confident that the mean GPA of college students in the United States is between 0 and 4.0. (This is not helpful!)

Narrowing the interval will decrease our confidence that we ACTUALLY include the population mean.

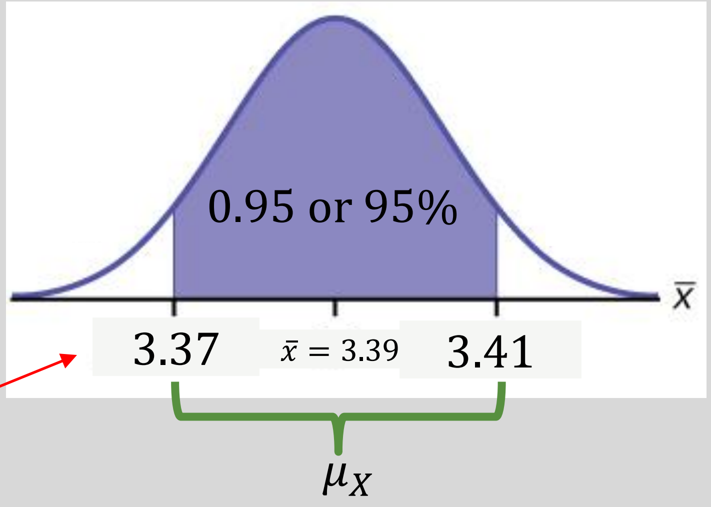


A Few Notes

The **confidence interval** can be narrowed by increasing the **sample size**.

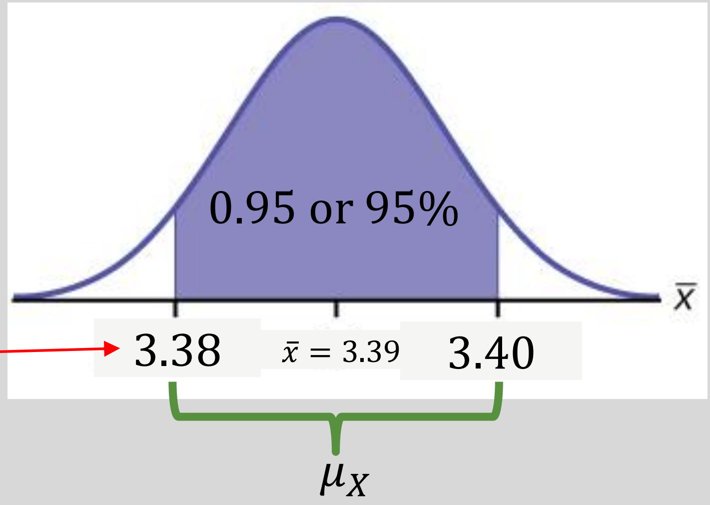
From the last example, n=3226

$$\text{EBM} = z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 1.96\left(\frac{0.57}{\sqrt{3226}}\right) \approx 0.0197$$



If the sample size is increased, the EBM decreases. Suppose n=30000

$$\text{EBM} = z_{\frac{\alpha}{2}}\left(\frac{\sigma_X}{\sqrt{n}}\right) = 1.96\left(\frac{0.57}{\sqrt{30000}}\right) \approx 0.0065$$



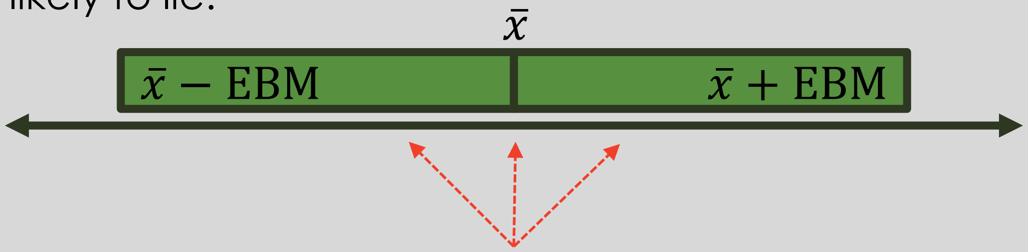
Summary

If a population has an UNKOWN mean and KNOWN standard deviation, the mean (μ_X) can be approximated by constructing a random sample.

The mean of the random sample is used as a **point estimate** of the population mean.

The desired **confidence level** and **standard error** dictate the **confidence interval**.

The **confidence interval** is where the TRUE population mean is likely to lie.



μ_X will likely fall in this range somewhere