

Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture 16

Confidence Intervals – Part I

What We've Done So Far

- ▶ Random Sampling: $X_1, \dots, X_n \sim \text{iid}$
- ▶ Use estimator $\hat{\theta}$ to learn about population parameter θ_0
- ▶ Estimator $\hat{\theta}$ is a random variable:
 - ▶ Distribution of $\hat{\theta}$ is called *sampling distribution*
 - ▶ Bias of an estimator
 - ▶ Variance of an estimator
 - ▶ Mean-squared Error (MSE) of an estimator
 - ▶ Consistency of an Estimator

Inference

Confidence Intervals

What values of θ_0 are consistent with the data we observed?

Hypothesis Testing

I think that $\theta_0 = 0$. Do the data we observed suggest that I should change my mind?

Am I Taller Than The Average American Male?



Source: Centers for Disease Control (pg. 16)

My height is 73 inches. Based on a sample of US males aged 20 and over, the Centers for Disease Control (CDC) reported a mean height of about 69 inches in a recent report.

Clearly I'm taller than the average American male!

Do you agree or disagree?

- (a) Agree
- (b) Disagree
- (c) Not Sure

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?
 - ▶ If the sample was very small there's a higher chance that it won't be representative of the population as a whole

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?
 - ▶ If the sample was very small there's a higher chance that it won't be representative of the population as a whole
 - ▶ Why? The variance of the sample mean is *decreasing with sample size* so bigger samples are less noisy.

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?
 - ▶ If the sample was very small there's a higher chance that it won't be representative of the population as a whole
 - ▶ Why? The variance of the sample mean is *decreasing with sample size* so bigger samples are less noisy.
- ▶ How much variability is there in height in the population?

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?
 - ▶ If the sample was very small there's a higher chance that it won't be representative of the population as a whole
 - ▶ Why? The variance of the sample mean is *decreasing with sample size* so bigger samples are less noisy.
- ▶ How much variability is there in height in the population?
 - ▶ If everyone is very similar in height, any sample we take will be representative of the population.

Remember: The Sample Mean is Random!

Just because the sample mean is 69 inches it doesn't follow that the population mean is 69 inches!

What Else Should We Consider?

- ▶ How big was the sample?
 - ▶ If the sample was very small there's a higher chance that it won't be representative of the population as a whole
 - ▶ Why? The variance of the sample mean is *decreasing with sample size* so bigger samples are less noisy.
- ▶ How much variability is there in height in the population?
 - ▶ If everyone is very similar in height, any sample we take will be representative of the population.
 - ▶ Remember: the variance of the sample mean is *increasing* with the population standard deviation.

Am I Taller Than The Average American Male?

Source: Centers for Disease Control (pg. 16)

Table : Height in inches for Males aged 20 and over (approximate)

Sample Mean	69 inches
Sample Std. Dev.	6 inches
Sample Size	5647
My Height	73 inches

We'll return to this example later.

For Now – Single Population, Normally Distributed

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Later we'll look at more than one population and talk about what happens if Normality doesn't hold.



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$?

- (a) $N(\mu, \sigma^2)$
- (b) $N(0, 1)$
- (c) $N(0, \sigma)$
- (d) $N(\mu, 1)$
- (e) Not enough information to determine.

Z-score!

Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. From above,

$$\begin{aligned}E[\bar{X}_n] &= \mu \\ \text{Var}(\bar{X}_n) &= \sigma^2/n \\ \Rightarrow SD(\bar{X}_n) &= \sigma/\sqrt{n}\end{aligned}$$

Thus,

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - E[\bar{X}_n]}{SD(\bar{X}_n)} \sim N(0, 1)$$

Remember that we call the standard deviation of a sampling distribution the **standard error**, written SE , so

$$\frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \sim N(0, 1)$$



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. What is the approximate value of the following?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right)$$



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. What is the approximate value of the following?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) \approx 0.95$$

What happens if I rearrange?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) = 0.95$$

What happens if I rearrange?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) = 0.95$$

$$P(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE) = 0.95$$

What happens if I rearrange?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) = 0.95$$

$$P(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE) = 0.95$$

$$P(-2 \cdot SE - \bar{X}_n \leq -\mu \leq 2 \cdot SE - \bar{X}_n) = 0.95$$

What happens if I rearrange?

$$P\left(-2 \leq \frac{\bar{X}_n - \mu}{SE(\bar{X}_n)} \leq 2\right) = 0.95$$

$$P(-2 \cdot SE \leq \bar{X}_n - \mu \leq 2 \cdot SE) = 0.95$$

$$P(-2 \cdot SE - \bar{X}_n \leq -\mu \leq 2 \cdot SE - \bar{X}_n) = 0.95$$

$$P(\bar{X}_n - 2 \cdot SE \leq \mu \leq \bar{X}_n + 2 \cdot SE) = 0.95$$

Confidence Intervals

Confidence Interval (CI)

A confidence interval is a range (A, B) constructed from the **sample data** that has a specified probability of containing a **population parameter**:

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Confidence Intervals

Confidence Interval (CI)

A confidence interval is a range (A, B) constructed from the **sample data** that has a specified probability of containing a **population parameter**:

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Confidence Level

The **specified probability**, typically denoted $1 - \alpha$, is called the confidence level. For example, if $\alpha = 0.05$ then the confidence level is 0.95 or 95%.

Confidence Interval for Mean of Normal Population

Population Variance Known

Confidence Interval for Mean of Normal Population

The interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ has approximately 95% probability of containing the population mean μ , provided that:

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Confidence Interval for Mean of Normal Population

Population Variance Known

Confidence Interval for Mean of Normal Population

The interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ has approximately 95% probability of containing the population mean μ , provided that:

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

But What Does This Mean?

Which quantities are random?



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. Which quantities are random variables?

- (a) μ only
- (b) σ and μ
- (c) σ only
- (d) σ, μ and \bar{X}_n
- (e) \bar{X}_n only

Which quantities are random?



Suppose $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. Which quantities are random variables?

- (a) μ only
- (b) σ and μ
- (c) σ only
- (d) σ, μ and \bar{X}_n
- (e) \bar{X}_n only

\bar{X}_n only.

Confidence Interval is a Random Variable!

1. X_1, \dots, X_n are RVs $\Rightarrow \bar{X}_n$ is a RV (repeated sampling)

Confidence Interval is a Random Variable!

1. X_1, \dots, X_n are RVs $\Rightarrow \bar{X}_n$ is a RV (repeated sampling)
2. μ , σ and n are constants

Confidence Interval is a Random Variable!

1. X_1, \dots, X_n are RVs $\Rightarrow \bar{X}_n$ is a RV (repeated sampling)
2. μ , σ and n are constants
3. Confidence Interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$ is also a RV!

Meaning of Confidence Interval

Meaning of Confidence Interval

If we sampled many times we'd get many different sample means, each leading to a **different** confidence interval. Approximately 95% of these intervals will contain μ .

Rough Intuition

What values of μ are consistent with the data?

CI for Population Mean: Repeated Sampling

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

CI for Population Mean: Repeated Sampling

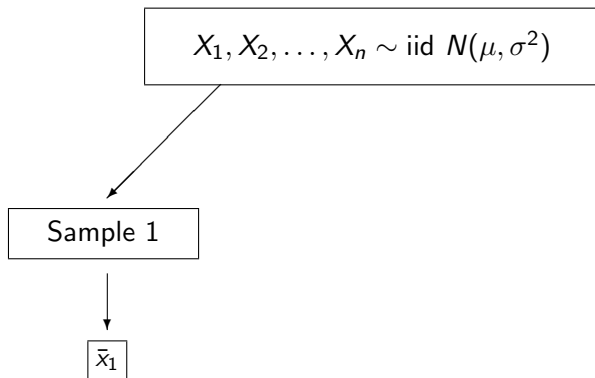
$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$



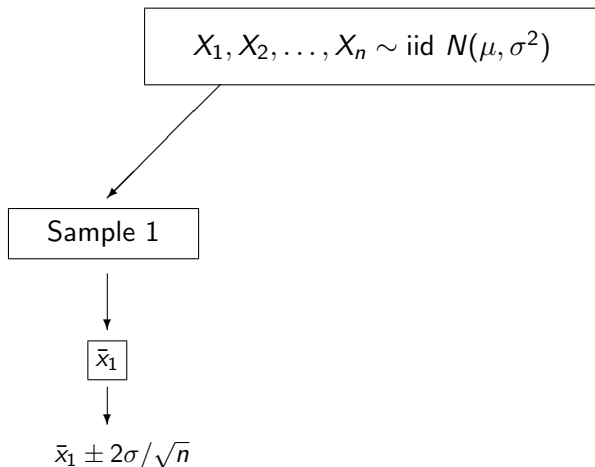
A diagram consisting of two rectangular boxes. The top box contains the mathematical expression $X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$. An arrow points from the bottom-left corner of this box down and to the left towards the top-left corner of a second box. The second box, located below and to the left of the first, contains the text "Sample 1".

Sample 1

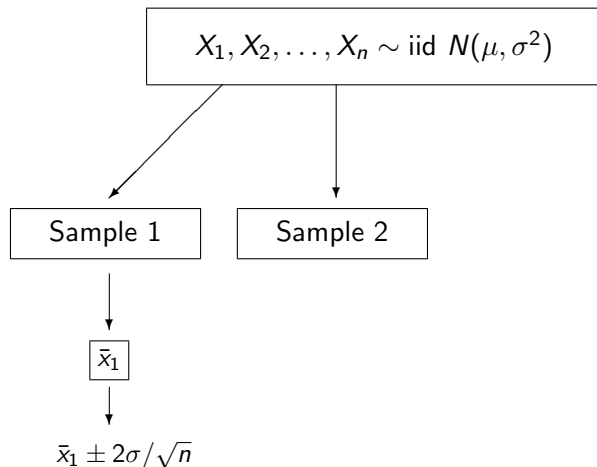
CI for Population Mean: Repeated Sampling



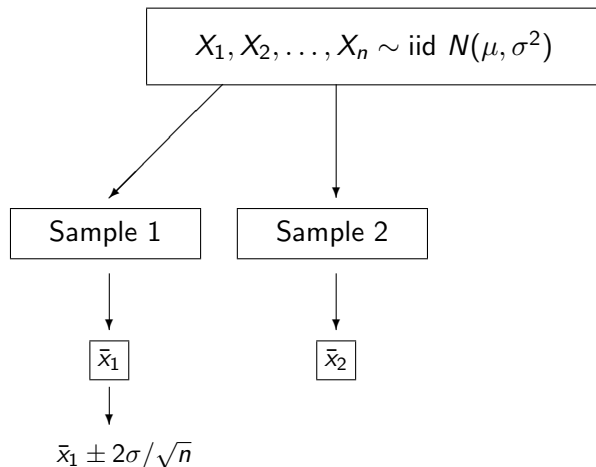
CI for Population Mean: Repeated Sampling



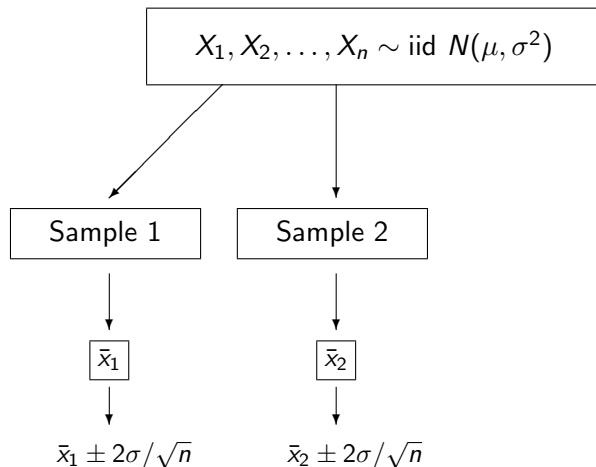
CI for Population Mean: Repeated Sampling



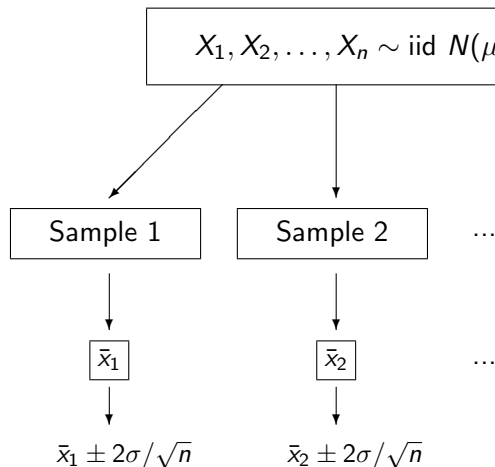
CI for Population Mean: Repeated Sampling



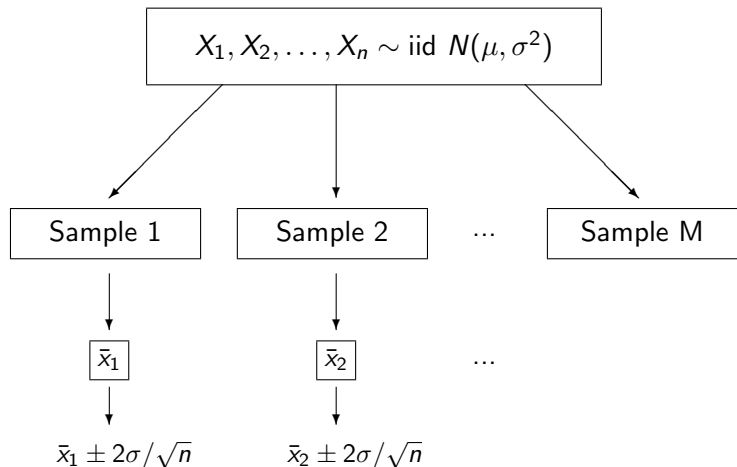
CI for Population Mean: Repeated Sampling



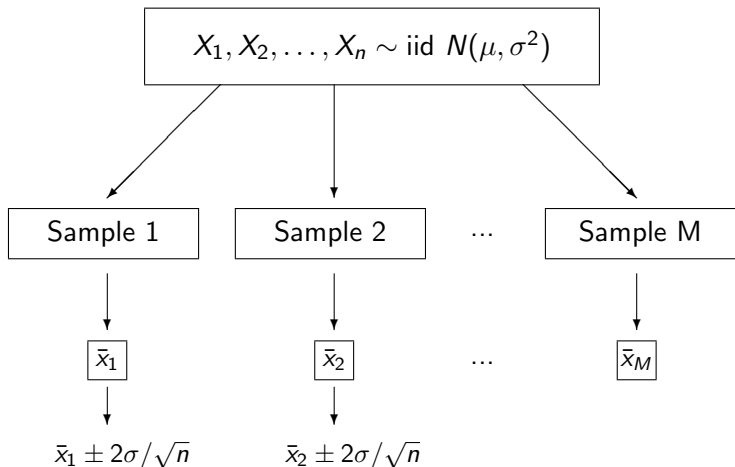
CI for Population Mean: Repeated Sampling



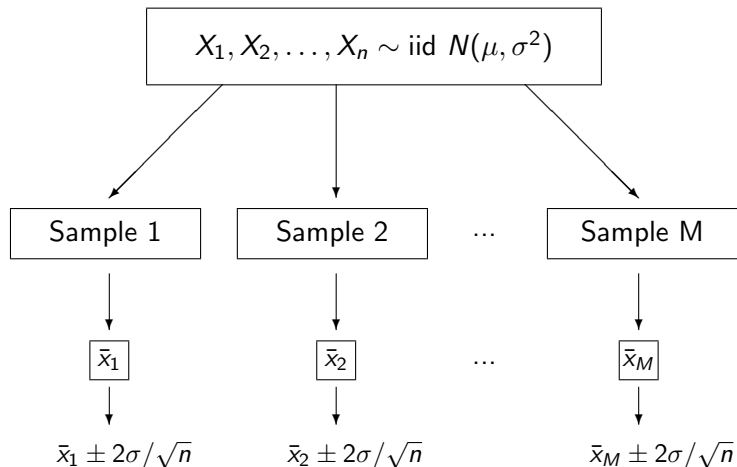
CI for Population Mean: Repeated Sampling



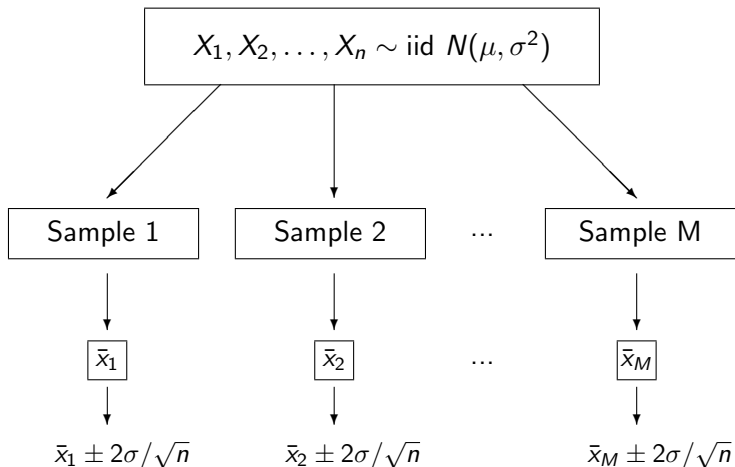
CI for Population Mean: Repeated Sampling



CI for Population Mean: Repeated Sampling

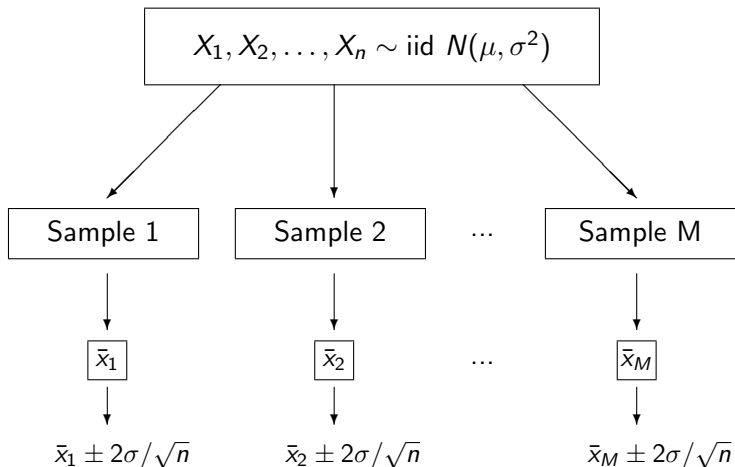


CI for Population Mean: Repeated Sampling



Repeat M times \rightarrow get M different intervals

CI for Population Mean: Repeated Sampling



Repeat M times \rightarrow get M different intervals

Large $M \Rightarrow$ Approx. 95% of these Intervals Contain μ

Simulation Example: $X_1, \dots, X_5 \sim \text{iid } N(0, 1)$, $M = 20$

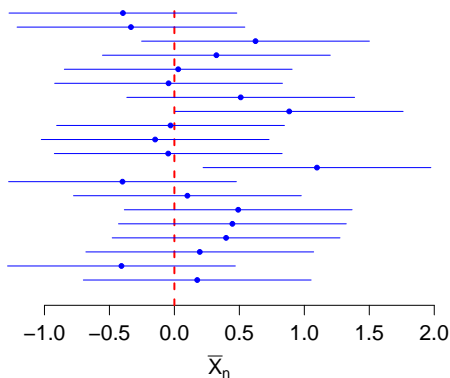


Figure : Twenty confidence intervals of the form $\bar{X}_n \pm 2\sigma/\sqrt{n}$ where $n = 5$, $\sigma^2 = 1$ and the true population mean is 0.

Meaning of Confidence Interval for θ_0

$$P(A \leq \theta_0 \leq B) = 1 - \alpha$$

Each time we sample we'll get a different confidence interval, corresponding to different realizations of the random variables A and B . If we sample many times, approximately $100 \times (1 - \alpha)\%$ of these intervals will contain the population parameter θ_0 .

True or False?



Suppose

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Then, there is about about a 95% chance that the population mean μ lies in the interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$.

- (a) True
- (b) False

True or False?



Suppose

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

Then, there is about about a 95% chance that the population mean μ lies in the interval $\bar{X}_n \pm 2\sigma/\sqrt{n}$.

- (a) True
- (b) False

FALSE! – μ is a constant!

Confidence Intervals: Some Terminology

Margin of Error

When a CI takes the form $\hat{\theta} \pm ME$, ME is the Margin of Error.

Confidence Intervals: Some Terminology

Margin of Error

When a CI takes the form $\hat{\theta} \pm ME$, ME is the Margin of Error.

Lower and Upper Confidence Limits

The lower endpoint of a CI is the **lower confidence limit (LCL)**, while the upper endpoint is the **upper confidence limit (UCL)**.

Confidence Intervals: Some Terminology

Margin of Error

When a CI takes the form $\hat{\theta} \pm ME$, ME is the Margin of Error.

Lower and Upper Confidence Limits

The lower endpoint of a CI is the **lower confidence limit (LCL)**, while the upper endpoint is the **upper confidence limit (UCL)**.

Width of a Confidence Interval

The distance $|UCL - LCL|$ is called the **width** of a CI. This means exactly what it says.

What is the Margin of Error



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the margin of error?

- (a) σ/\sqrt{n}
- (b) \bar{X}_n
- (c) σ
- (d) $2\sigma/\sqrt{n}$
- (e) $1/\sqrt{n}$

What is the Margin of Error



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the margin of error?

- (a) σ/\sqrt{n}
- (b) \bar{X}_n
- (c) σ
- (d) $2\sigma/\sqrt{n}$
- (e) $1/\sqrt{n}$

$2\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

What is the Width?



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the width of the interval?

- (a) σ/\sqrt{n}
- (b) $2\sigma/\sqrt{n}$
- (c) $3\sigma/\sqrt{n}$
- (d) $4\sigma/\sqrt{n}$
- (e) $5\sigma/\sqrt{n}$

What is the Width?



In the preceding example of a 95% confidence interval for the mean of a normal population when the population variance is known, which of these is the width of the interval?

- (a) σ/\sqrt{n}
- (b) $2\sigma/\sqrt{n}$
- (c) $3\sigma/\sqrt{n}$
- (d) $4\sigma/\sqrt{n}$
- (e) $5\sigma/\sqrt{n}$

$4\sigma/\sqrt{n}$, since the CI is $\bar{X}_n \pm 2\sigma/\sqrt{n}$

Example: Calculate the Margin of Error



$X_1, \dots, X_{100} \sim \text{iid } N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

What is the margin of error?

Example: Calculate the Margin of Error



$X_1, \dots, X_{100} \sim \text{iid } N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

What is the margin of error?

The confidence interval is $\bar{X}_n \pm 2\sigma/\sqrt{n}$ so

$$ME = 2\sigma/\sqrt{n} = 2 \cdot 1/\sqrt{100} = 2/10 = 0.2$$

Example: Calculate the Lower Confidence Limit



$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the lower confidence limit?

Example: Calculate the Lower Confidence Limit



$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the lower confidence limit?

$$LCL = \bar{x} - ME = 4.9 - 0.2 = 4.7$$

Example: Calculate the Upper Confidence Limit



$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the upper confidence limit?

Example: Calculate the Upper Confidence Limit



$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .
Want to create a 95% confidence interval for μ .

We found that $ME = 0.2$. The sample mean $\bar{x} = 4.9$. What is the upper confidence limit?

$$UCL = \bar{x} + ME = 4.9 + 0.2 = 5.1$$

Example: 95% CI for Normal Mean, Popn. Var. Known

$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .

95% CI for $\mu = [4.7, 5.1]$

What values of μ are plausible?

Example: 95% CI for Normal Mean, Popn. Var. Known

$X_1, \dots, X_{100} \sim N(\mu, 1)$ but we don't know μ .

95% CI for $\mu = [4.7, 5.1]$

What values of μ are plausible?

The data actually came from a $N(5, 1)$ Distribution.

Want to be more certain? Use higher confidence level.

What value of c should we use to get a $100 \times (1 - \alpha)\%$ CI for μ ?

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = 1 - \alpha$$

Want to be more certain? Use higher confidence level.

What value of c should we use to get a $100 \times (1 - \alpha)\%$ CI for μ ?

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + c\sigma/\sqrt{n}\right) = 1 - \alpha$$

Want to be more certain? Use higher confidence level.

What value of c should we use to get a $100 \times (1 - \alpha)\%$ CI for μ ?

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + c\sigma/\sqrt{n}\right) = 1 - \alpha$$

Take $c = \text{qnorm}(1 - \alpha/2)$

Want to be more certain? Use higher confidence level.

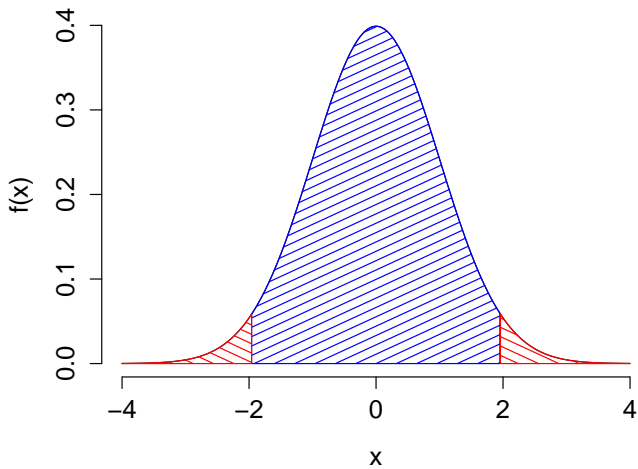
What value of c should we use to get a $100 \times (1 - \alpha)\%$ CI for μ ?

$$P\left(-c \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = 1 - \alpha$$

$$P\left(\bar{X}_n - c\sigma/\sqrt{n} \leq \mu \leq \bar{X}_n + c\sigma/\sqrt{n}\right) = 1 - \alpha$$

Take $c = \text{qnorm}(1 - \alpha/2)$

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma/\sqrt{n}$$



Confidence Interval for a Normal Mean, σ Known

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

What Affects the Margin of Error?

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

Sample Size n

ME decreases with n : bigger sample \implies tighter interval

Population Std. Dev. σ

ME increases with σ : more variable population \implies wider interval

Confidence Level $1 - \alpha$

ME increases with $1 - \alpha$: higher conf. level \implies wider interval

What Affects the Margin of Error?

$$\bar{X}_n \pm \text{qnorm}(1 - \alpha/2) \times \sigma / \sqrt{n}$$

Sample Size n

ME decreases with n : bigger sample \implies tighter interval

Population Std. Dev. σ

ME increases with σ : more variable population \implies wider interval

Confidence Level $1 - \alpha$

ME increases with $1 - \alpha$: higher conf. level \implies wider interval

Conf. Level	90%	95%	99%
α	0.1	0.05	0.01
$\text{qnorm}(1 - \alpha/2)$	1.64	1.96	2.56

But What if σ is Unknown?

- ▶ What we've done so far assumed that σ was known.
- ▶ In real applications this is typically not the case.

Why not try using the sample standard deviation s ?

This works, but requires a small change. Instead of basing the interval on quantiles of a normal distribution, we need to use a t distribution. We'll look at this next time.