

STAT 135

6. Confidence intervals

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)

Office hours: Tu 9:30-10:30 (in person), Th 1:30-2:30 (virtual)

Office: Evans 339

Email: rebeccabarter@berkeley.edu

Twitter: @rlbarter

GitHub: rlbarter

Some useful results from the CLT

General useful result for CLT

Corollary: if X_1, X_2, \dots, X_n is an IID sample from a population with mean μ and standard deviation σ , then if $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0,1), \quad \text{as } n \rightarrow \infty$$

Regardless of the form of the original distribution of the X_i .

General useful result for CLT

Corollary: if X_1, X_2, \dots, X_n is an IID sample from a population with mean μ and standard deviation σ , then if $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$:

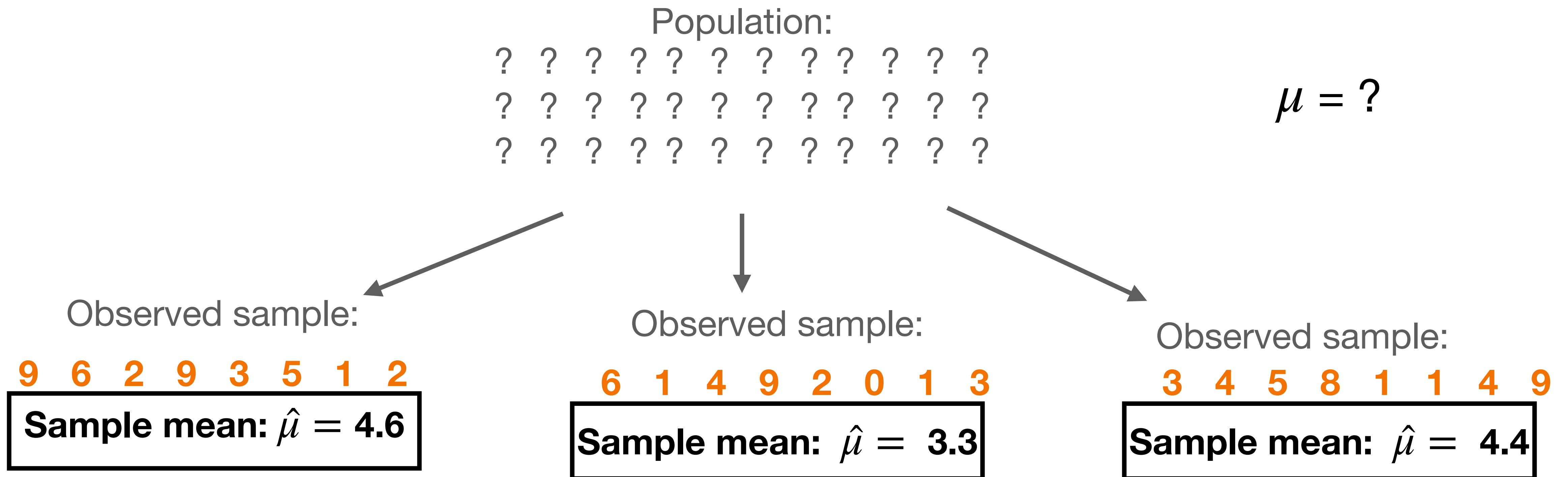
$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \rightarrow \Phi(z), \quad \text{as } n \rightarrow \infty$$

Where Φ is the CDF of the standard normal distribution

Confidence intervals

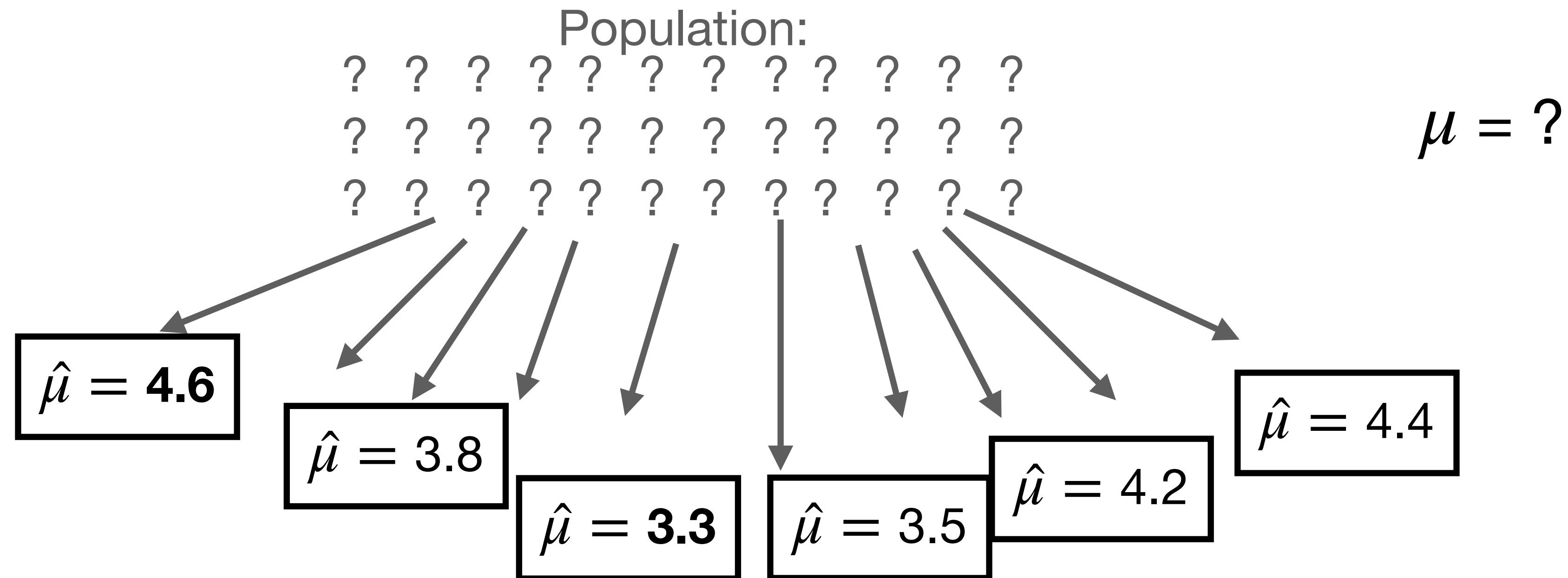
Confidence intervals

Recall that a parameter estimate that we obtain from a sample is just one possible version that we *could* have obtained.



The “true” parameter value is generally likely to lie within the range of parameter estimates that we *could* have obtained.

Confidence intervals



If we have enough different versions of these parameter estimates, we can try to make claims like “*the true parameter value probably **lies between 3.3 and 4.6***”

Confidence intervals

A **confidence interval** is an interval that is calculated in such a way that it contains the true population value of θ with some specified probability $(1 - \alpha)$.

$(1 - \alpha)$ is called the **coverage probability** or **confidence level**.

A common choice is $\alpha = 0.05$, which corresponds to a 95% confidence interval

A $(1 - \alpha)\%$ **confidence interval** $[L, U]$ for a parameter θ , is an interval calculated from a sample that contains θ with probability

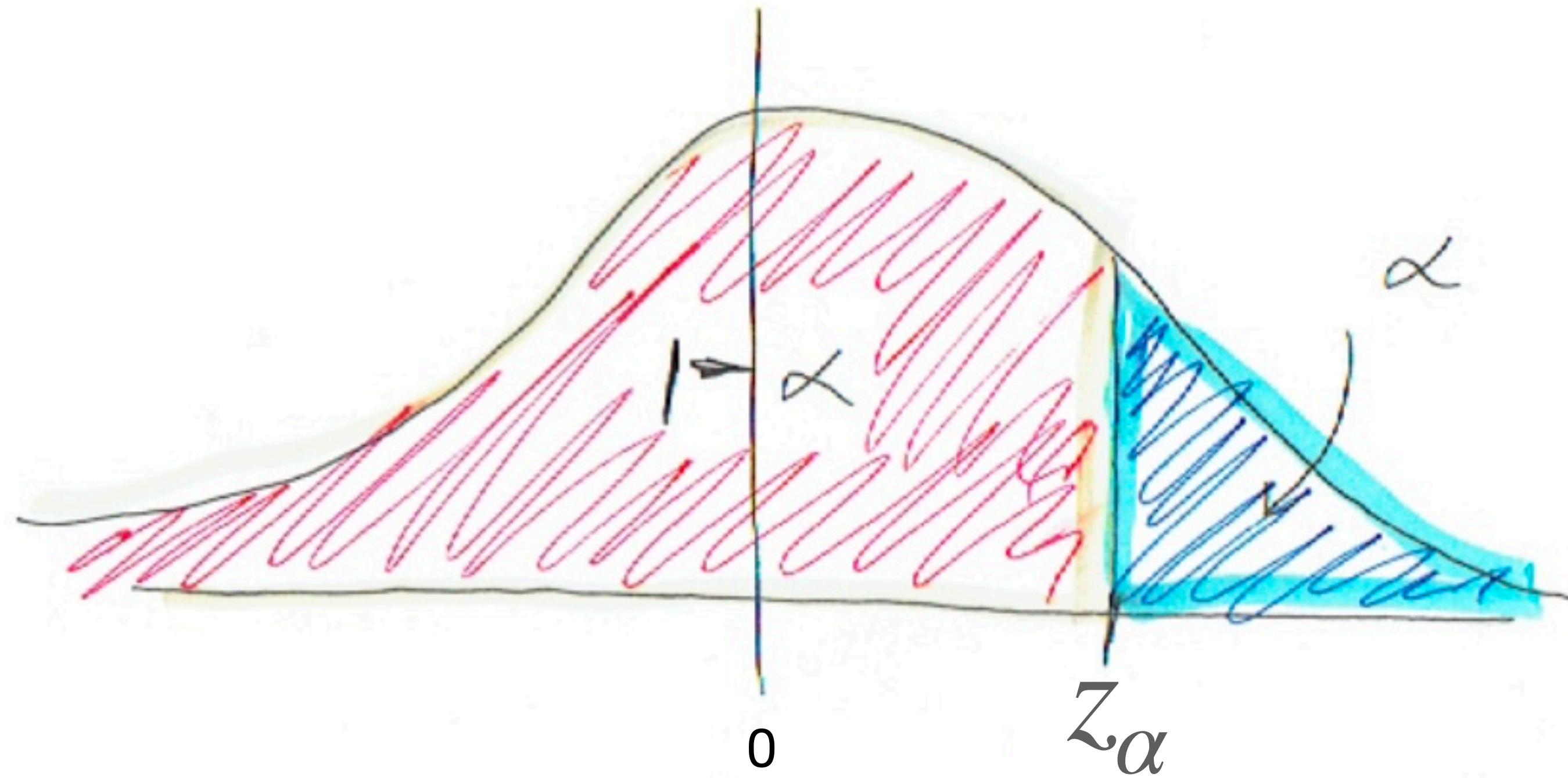
$$P(L \leq \theta \leq U) \geq 1 - \alpha$$

Confidence intervals

Let $Z \sim N(0,1)$

Define z_α to be the $(1 - \alpha)$ -quantile of the $N(0, 1)$ distribution, then:

$$P(Z < z_\alpha) = 1 - \alpha$$



$(1 - \alpha)$ quantile

Confidence intervals

If $Z \sim N(0,1)$ and z_α is the number such that

$$P(Z < z_\alpha) = 1 - \alpha$$

By the symmetry of the normal distribution, we also have

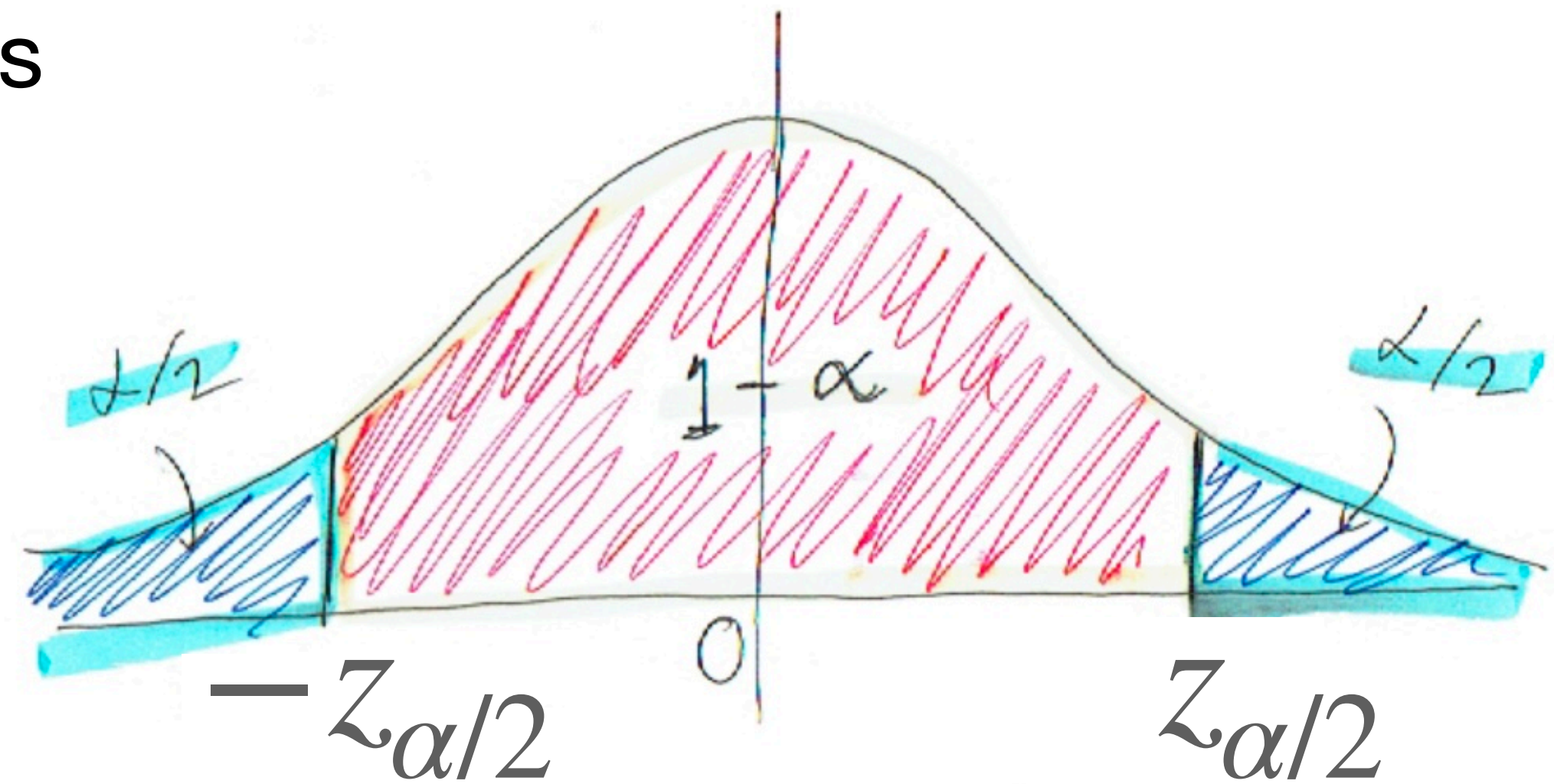
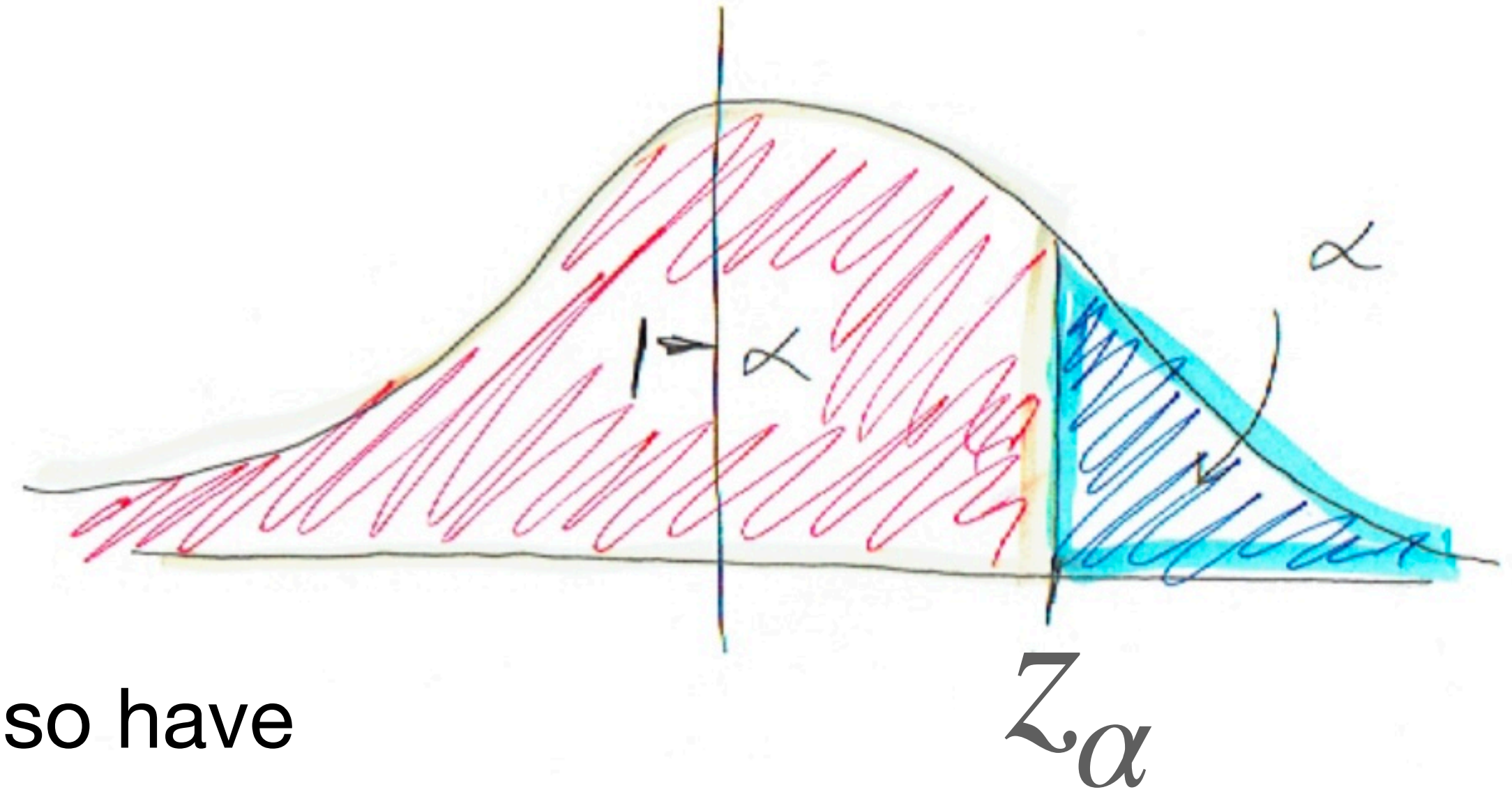
$$1 - \alpha = P(Z < z_\alpha) = P(-z_{\alpha/2} < Z < z_{\alpha/2})$$

A 95% CI for μ when the data is $N(0, 1)$ is thus $[-1.96, 1.96]$, since:

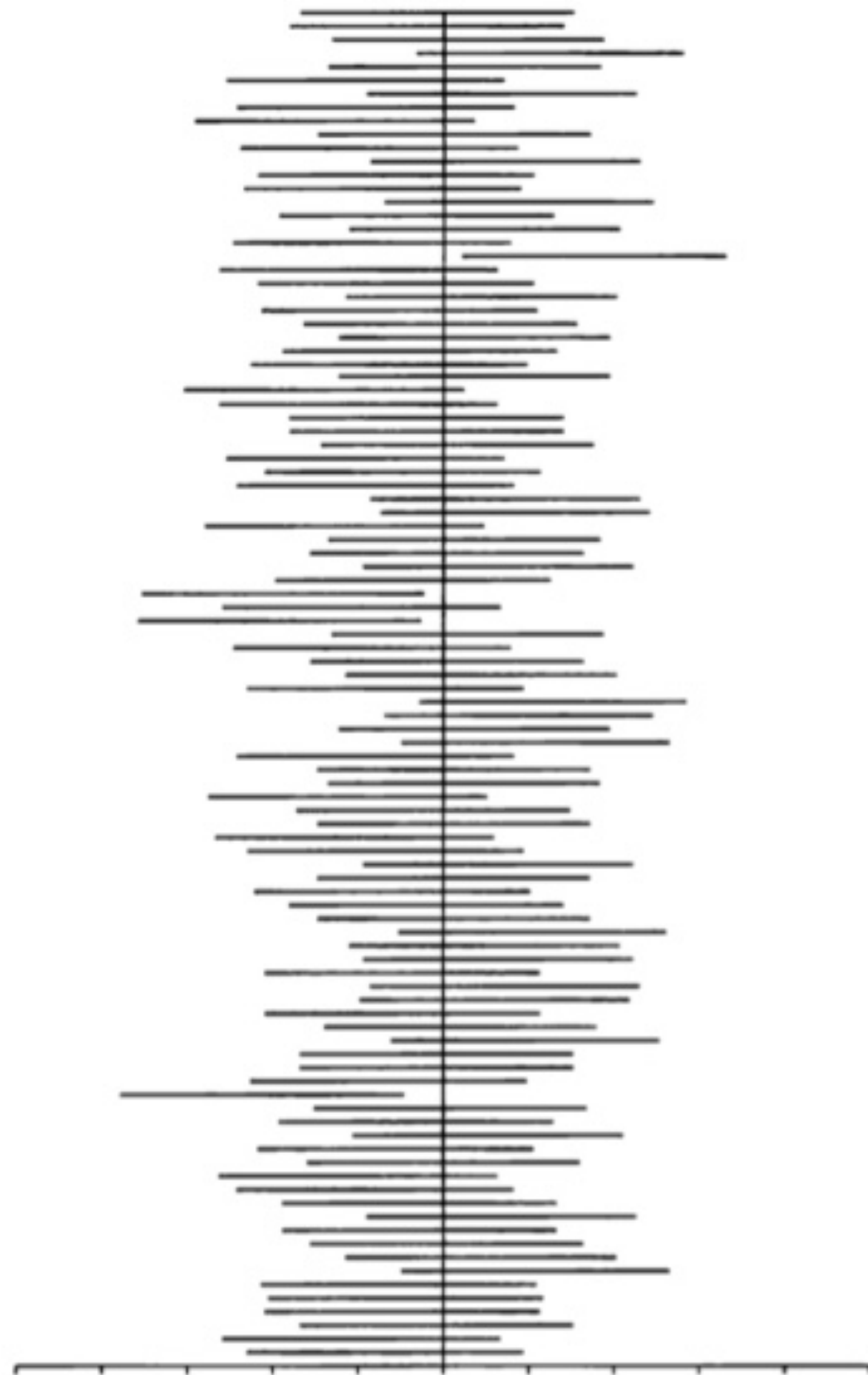
$$z_{0.05/2} = z_{0.025} = 1.96$$

Because

$$P(-1.96 < Z < 1.96) = 0.95$$



Confidence intervals



95% confidence intervals computed from IID random samples contain the true population parameter 95% of the time.

Confidence intervals

How do we generate confidence intervals for general parameter estimates?

If our estimator approximately satisfies (e.g., by the CLT, or MLE results)

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0,1)$$

$\sigma_{\hat{\theta}}^2 = \text{Var}(\hat{\theta})$ denotes the variance of $\hat{\theta}$

Then we have that (approximately)

Or more generally, that

$$P(-1.96 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 1.96) = 0.95$$

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}) = 1 - \alpha$$

Rearranging, gives:

$$P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$$

Confidence intervals

We have: $P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$

So a $(1 - \alpha)\%$ CI for θ (when $\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$ is approximately $N(0, 1)$), can be computed as:

$$[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} , \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$$

This interval contains the true θ with probability $1 - \alpha$

Note that the interval is centered at the sample estimate, $\hat{\theta}$.

Confidence intervals

Some generally useful N(0, 1) quantile results:

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}) = 1 - \alpha$$

90% CI, $\alpha = 0.1$

$$z_{\alpha/2} = z_{0.05} = 1.645$$

$$P(-1.64 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 1.64) = 0.90$$

```
> round(qnorm(0.05), 2)
```

```
[1] -1.64
```

```
> round(qnorm(0.95), 2)
```

```
[1] 1.64
```

95% CI, $\alpha = 0.05$

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

$$P(-1.96 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 1.96) = 0.95$$

```
> round(qnorm(0.025), 2)
```

```
[1] -1.96
```

```
> round(qnorm(0.975), 2)
```

```
[1] 1.96
```

99% CI, $\alpha = 0.01$

$$z_{\alpha/2} = z_{0.005} = 2.5$$

$$P(-2.58 < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < 2.58) = 0.99$$

```
> round(qnorm(0.005), 2)
```

```
[1] -2.58
```

```
> round(qnorm(0.995), 2)
```

```
[1] 2.58
```

Notice that the confidence intervals become wider as the confidence increases: the wider the interval, the more confident we are that it contains the true parameter

Confidence intervals for the sample mean

If X_1, \dots, X_n (where n is fairly large) are IID from any distribution that has population mean μ and population SD, σ . Let $\bar{X} = \sum_{i=1}^n X_i$. Let's calculate a $(1 - \alpha)\%$ CI for μ .

By the CLT, we know that $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \underset{\text{approx}}{\sim} N(0,1)$

This means that

And re-arranging gives

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$
$$P\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

So a $(1 - \alpha)\%$ confidence interval for μ is:

$$\left[\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]$$

Confidence intervals for the MLE

If X_1, \dots, X_n (where n is fairly large) are IID from any distribution with parameter θ and $\hat{\theta}_{MLE}$ is the MLE estimate of θ , a $(1 - \alpha)\%$ CI for θ is:

By the asymptotic normality of the MLE, we know that


$$\frac{\hat{\theta}_{MLE} - \theta}{\sigma_{\hat{\theta}_{MLE}}} = \frac{\hat{\theta}_{MLE} - \theta}{\sqrt{1/nI(\theta)}} \sim N(0,1)$$

Note that we substituted $\hat{\theta}_{MLE}$ into the fisher information

So a $(1 - \alpha)\%$ confidence interval for $\hat{\theta}_{MLE}$

$$\left[\hat{\theta}_{MLE} - z_{\alpha/2} \sigma_{\hat{\theta}_{MLE}}, \hat{\theta}_{MLE} + z_{\alpha/2} \sigma_{\hat{\theta}_{MLE}} \right] = \left[\hat{\theta}_{MLE} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{MLE})}}, \hat{\theta}_{MLE} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{MLE})}} \right]$$

But there's a problem!

$$[\hat{\theta} - z_{\alpha/2} \sigma_{\hat{\theta}} , \hat{\theta} + z_{\alpha/2} \sigma_{\hat{\theta}}]$$


This is the “true” SD *of the parameter estimate*... which we typically don't know

For the mean: $\sigma_{\bar{X}} = \sqrt{\sigma/n}$

But we don't know the value of σ !

But we can estimate σ from the data using:

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

For a general estimator, $\hat{\theta}$,
you can estimate $\sigma_{\hat{\theta}}$ using the
bootstrap

CI for the mean: unknown pop variance

If the X_i are IID and **Normally distributed** with unknown population variance σ^2 , then an unbiased estimate is

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

And it turns out that:

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

Where t_{n-1} is the student's t distribution with $n - 1$ degrees of freedom

(It turns out that even if the data are not normal, for large sample sizes, this still approximately holds)

So compute the following $(1 - \alpha)\%$ CI.

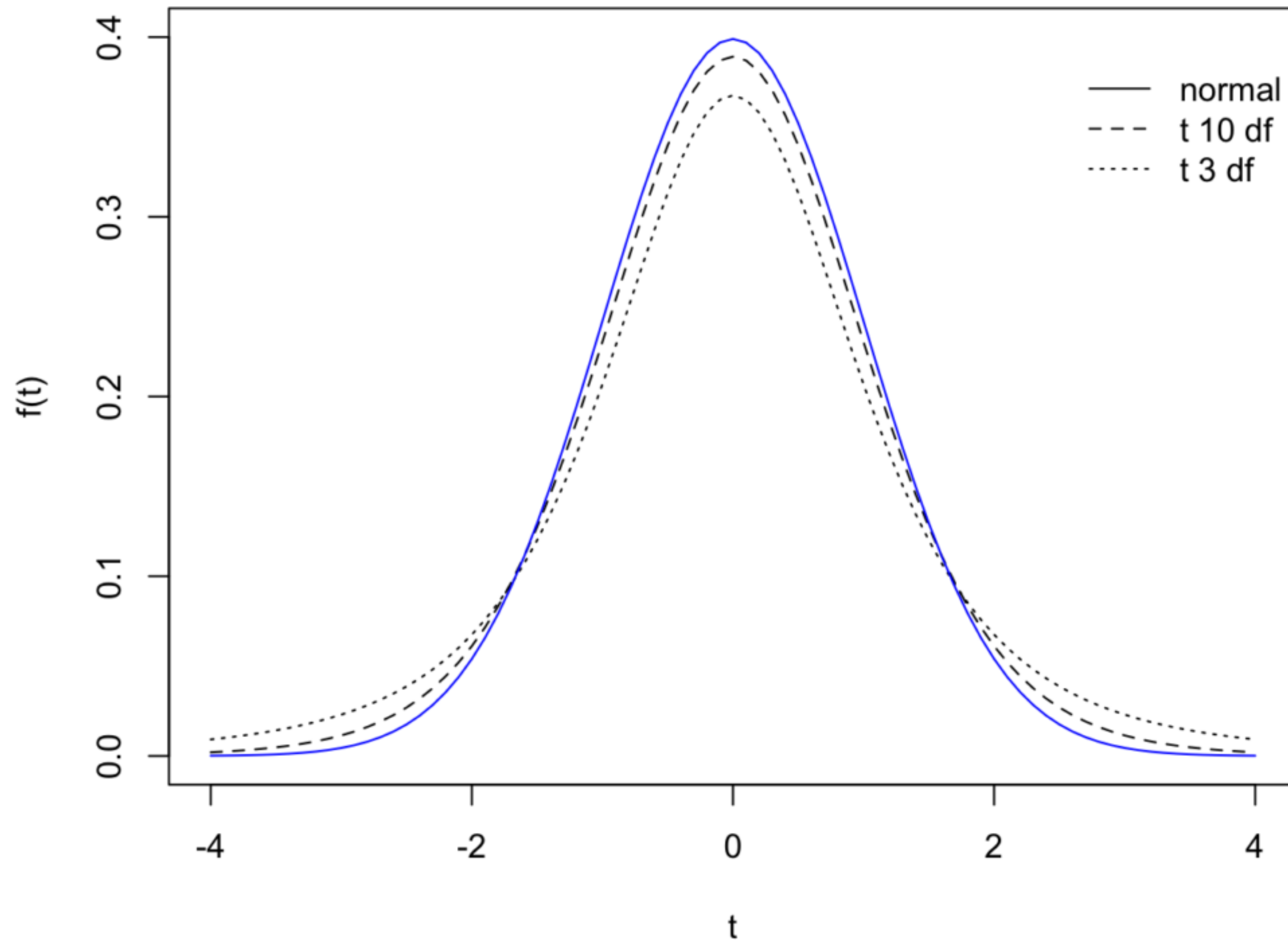
$$\left[\bar{X} - \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}}, \bar{X} + \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}} \right]$$

Where $t_{n-1, \alpha}$ is the value such that $P(T \leq t_{n-1, \alpha}) = 1 - \alpha$

```
> qt(0.05, 10)
[1] -1.812461
> qt(0.05, 100)
[1] -1.660234
> qt(0.05, 1000)
[1] -1.646379
```

```
> qt(0.025, 10)
[1] -2.228139
> qt(0.025, 100)
[1] -1.983972
> qt(0.025, 1000)
[1] -1.962339
```

Student's t-distribution t_{n-1}



A n gets bigger, the t-dist looks more and more like a Normal distribution

... When n is bigger (e.g. > 30), you can use the normal distribution

Confidence interval example

A random sample of 500 shoppers at Walgreens spend an average of \$14.80 per visit.



What else do we need to know to compute a 95% CI?

The spending distribution is Normal and from our sample of 500 shoppers, the sample standard deviation is:

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \$5.5. \quad \bar{x} = 14.8; \quad \hat{\sigma} = 5.5$$
$$n = 500; \quad t_{n-1, \alpha/2} = t_{499, 0.975} = 1.96$$

Then a 95% CI for the population mean is

$$\left[\bar{X} - \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}}, \bar{X} + \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}} \right] = \left[14.8 - \frac{1.96 \times 5.5}{\sqrt{500}}, 14.8 + \frac{1.96 \times 5.5}{\sqrt{500}} \right] = [14.3, 15.3]$$

Confidence interval example

How would our CI change if we **only had 50 people in our sample** (assuming the sample mean and SD remain the same)?



Then a 95% CI is

$$\left[\bar{X} - \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}}, \bar{X} + \frac{t_{n-1, \alpha/2} \hat{\sigma}}{\sqrt{n}} \right] = \left[14.8 - \frac{1.96 \times 5.5}{\sqrt{50}}, 14.8 + \frac{1.96 \times 5.5}{\sqrt{50}} \right] = [13.3, 16.3]$$

(The original CI was [14.3, 15.3])

The interval got *wider*,
because our estimates
are more variable

Confidence interval example

A $(1 - \alpha)\%$ CI for the mean of IID RVs where σ is known is $\left[\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right]$

Does increasing the confidence of the interval lead to a wider or narrower interval?

Wider.

Theoretical explanation: $z_{\alpha/2}$ will be larger

Intuitive explanation: Keeping all else the same, we are more confident that a wider interval will contain the true parameter

Does increasing the sample size lead to a wider or narrower interval?

Narrower.

Theoretical explanation: n will be larger

Intuitive explanation: The more data we have, the more we are able to narrow down our interval on the true parameter

Coverage

Coverage

The **coverage** of $(1 - \alpha) \%$ confidence interval is the (expected) proportion of the intervals that *actually* contain (“cover”) the true parameter.

What is the *intended* coverage of a 90% confidence interval?

... 90%!

Is the *actual* coverage of a 90% confidence interval 90%?

Yes.... But only if all of the assumptions you made about the independence and distributions of the data are true.

Example

See ci.R

A shiny app

Let's play around with Bruce Dudek's (University at Albany) confidence interval simulation shiny app!

<https://shiny.rit.albany.edu/stat/confidence/>

Confidence Interval for a Sample Mean: A simulation

This app randomly samples N data points from a Normal Distribution. The user specifies N. Sample means are computed for each simulated sample.

Choose which Graphs

☒ Confidence Interval Graph Only

☐ Confidence Interval Graph Plus Sampling Distribution of the Mean

Display Choice for CI Graph

☒ Means Plus CI

☐ Means Only

Sample Size

4

Number of Simulated Samples

Five

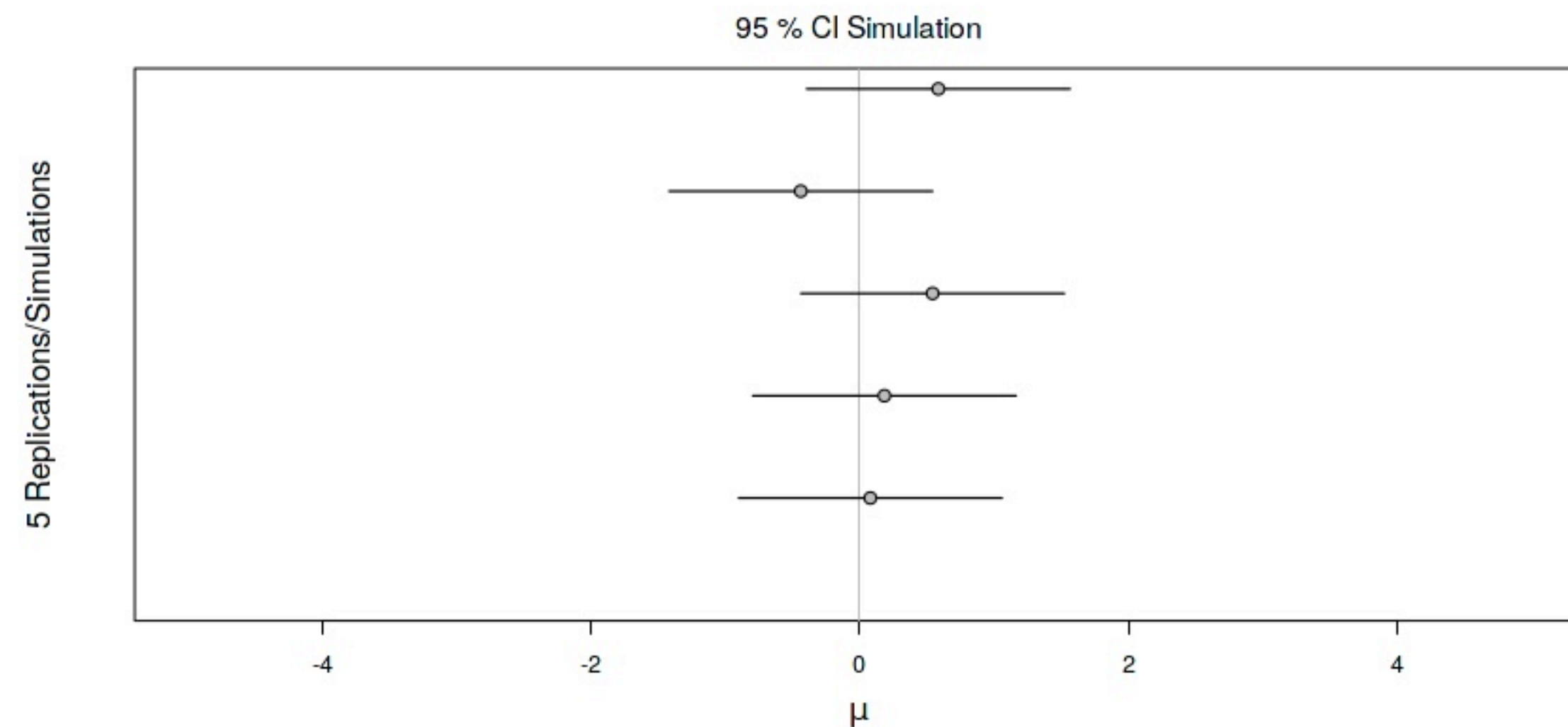
Confidence Level

95

Plots

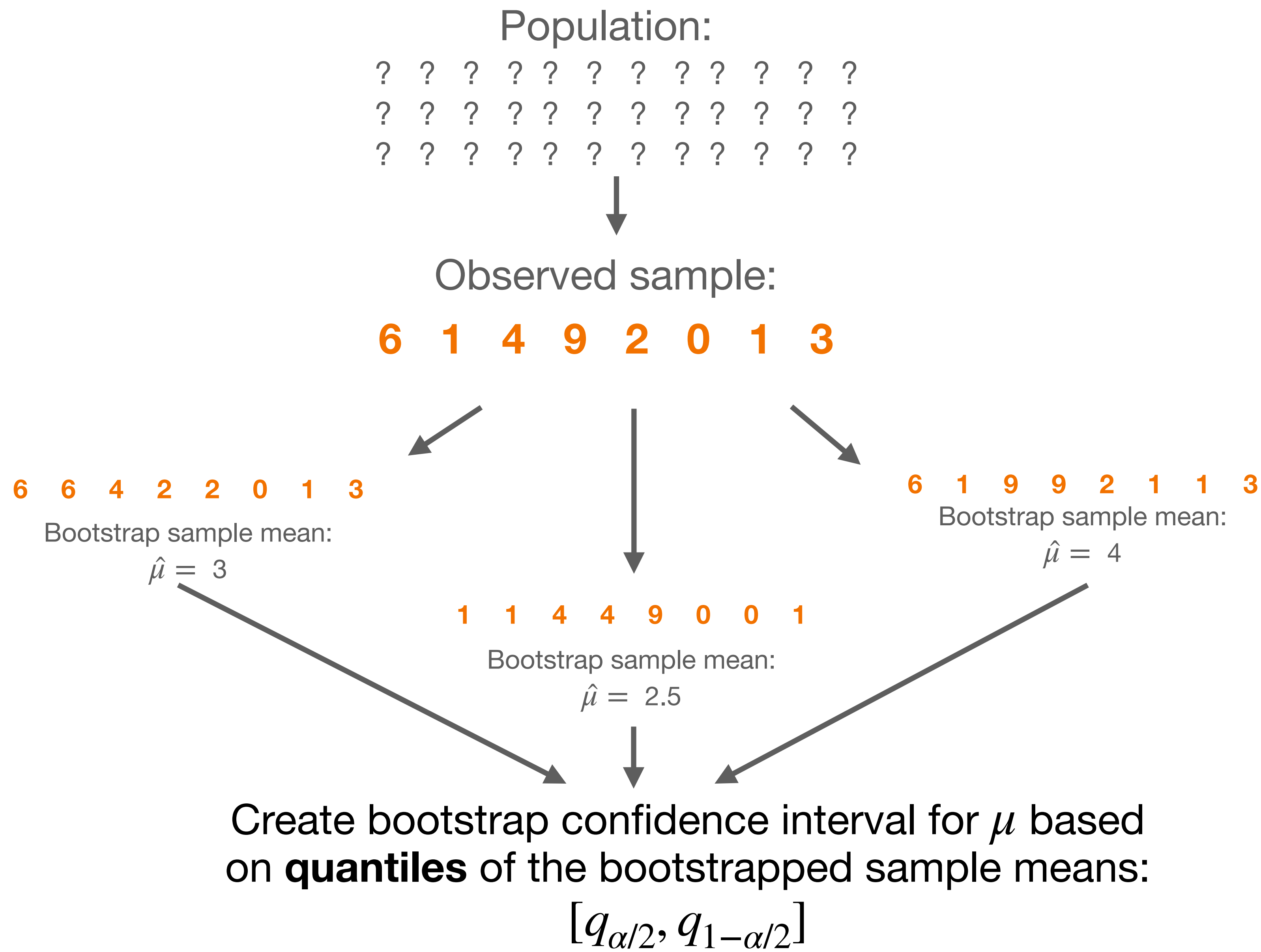
About

Click your mouse on the plot to see an explanation of the graph.



Confidence intervals via non-parametric bootstrap

Confidence intervals via non-parametric bootstrap



Example

See bootstrap_ci.R

Confidence interval for a proportion

Confidence interval example for proportions

You will show in your homework that an approximate confidence interval for a population proportion (when the data is IID Bernoulli(p)) is

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

This is based on the Normal approximation of \hat{p}

Confidence interval example for proportions

A random sample of 50 Berkeley residents found that 17 of them had used a food delivery app in the past 7 days.

Let's calculate a 90% confidence interval for the proportion of all Berkeley residents that have used a food delivery app in the past 7 days

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

The relevant information is: $z_{\alpha/2} = 1.64$, $\hat{p} = 17/50 = 0.34$, $n = 50$

So the 90% CI is: $\left[0.34 - 1.64 \sqrt{\frac{0.34(1 - 0.34)}{50}}, 0.34 + 1.64 \sqrt{\frac{0.34(1 - 0.34)}{50}} \right] = [0.33, 0.35]$