Confidence Intervals for a Mean

David Gerard

2017-10-25

Learning Objectives

- Inference for a population mean.
- Confidence intervals for a population mean.
- Interpreting confidence intervals.
- Sections 4.1 and 4.2 of DBC.

Review: Statistics vs Probability

- **Statistics (Inference):**
  - Just observe a sample. What can we conclude (probabilistically) about the population?
  - Sample $\longrightarrow$ Population?
  - Messy and more of an art.
  - No correct answers. Lots of wrong answers. Some "good enough" answers.
- **Probability (from the viewpoint of Statisticians):**
  - Logically self-contained, a subset of Mathematics.
  - One correct answer.
  - We know the population. What is the probability of the sample?
  - Population $\longrightarrow$ Sample?

Speed of Light

In 1879, Albert Michaelson ran an experiment to estimate the speed of light. Let's use his data. (Different from the famous Michaelson-Morley experiment.)

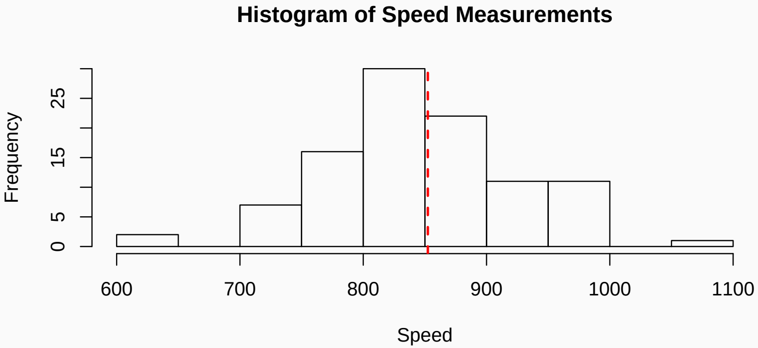
```
library(tidyverse)
data("morley")
glimpse(morley)


Observations: 100
Variables: 3
$ Expt  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ Run   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
$ Speed <int> 850, 740, 900, 1070, 930, 850, 950, 980, ...
```

Speed is in units km/s with 299,000 subtracted.

A histogram

```
hist(morley$Speed, xlab = "Speed",
     main = "Histogram of Speed Measurements", xlim = c(600
abline(v = mean(morley$Speed), col = 2,
       lty = 2, lwd = 2)
```



What can we say

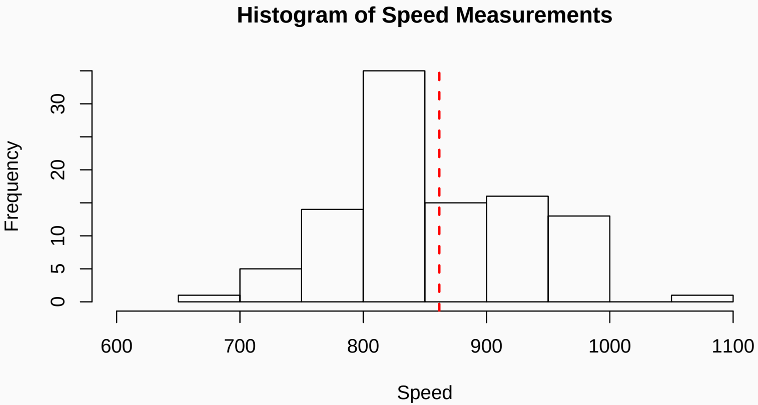
If this experiment were done with no bias, then:

- $E[\bar{X}] = \mu$
- $SD(\bar{X}) = \sigma/\sqrt{n}$
- $\bar{X} \underset{n\to\infty}{\longrightarrow} \mu$ (Law of Large Numbers)
- $\bar{X} \sim N(\mu, \sigma^2/n)$, approximately (Central Limit Theorem).

Point Estimate

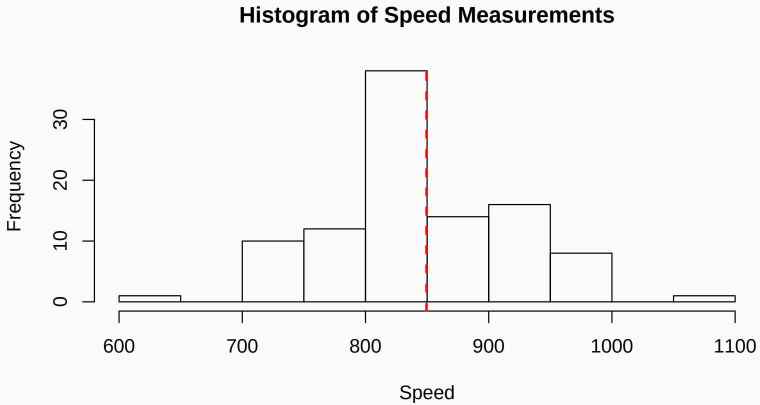
- Right now, our best guess for the value of μ is $\bar{X} = 852.4$.
- However, point estimates are not exact.

A different sample



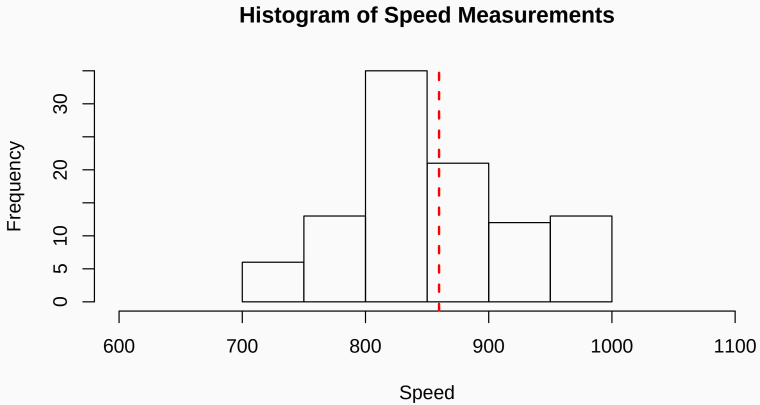
$\bar{X} = 861.7$

A different sample



$\bar{X} = 849.5$

A different sample



$\bar{X} = 859.7$

A different sample

$\bar{X} = 849.2$

- Unfortunately, we never actually observe other values of $\bar{X}$.
- Luckily, we have theory that says that for most random variables, we know the distribution of $\bar{X}$.
- $\bar{X} \sim N(\mu, \sigma^2/n)$.
- So we know on average how far away $\bar{X}$ will be from μ on average.

Recall: 68-95-99.7 rule

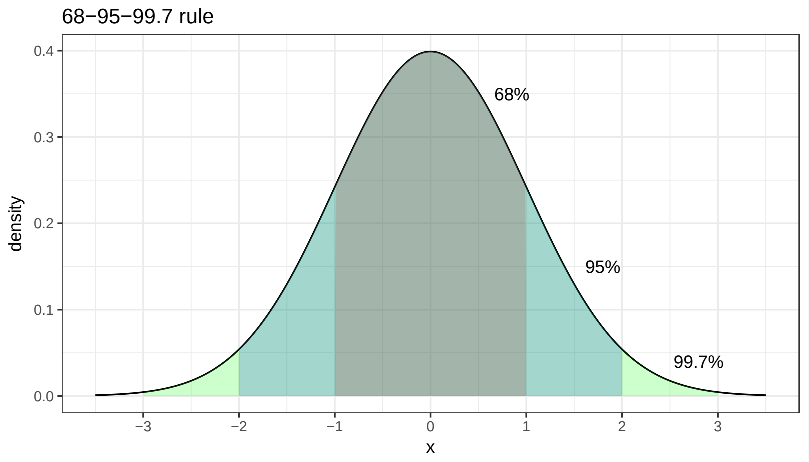
68-95-99.7 rule

In the Normal distribution with mean μ and standard deviation σ

- Approximately 68% of the observations fall within σ of μ
- Approximately 95% of the observations fall within 2σ of μ
- Approximately 99.7% of the observations fall within 3σ of μ

This rule does not depend on the values of μ and σ.

Recall: 68-95-99.7 rule



A random interval

Applying this rule to $\bar{X}$

$$P\left(\mu - 2\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 2\sigma/\sqrt{n}\right) = 0.95$$

Rearranging terms we get

$$P\left(\bar{X} - 2\sigma/\sqrt{n}) \leq \mu \leq \bar{X} + 2\sigma/\sqrt{n}\right) = 0.95.$$

That is, the *random interval* $(\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n})$ covers the mean μ in 95% of all samples.

What about σ?

- σ is a population parameter, that we generally don't know.
- Recall that we use s, the sample standard deviation, as a point estimate of σ.
- For large n, using s instead of σ doesn't matter.
- For small n (e.g. $n \leq 30$), intervals are too small (more on this later).

Calculating 95% Confidence Intervals for Mean

1. Take a random sample of size n calculate the sample mean $\bar{X}$
2. If n is large enough, then can assume $\bar{X} \sim N(\mu, \sigma^2/n)$
3. The 95% confidence interval is given by

$$\left(\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}\right)$$

1.96 is slightly more accurate than 2. In practice this doesn't matter too much.

Intuition of CI

What if we repeat the following over and over again:

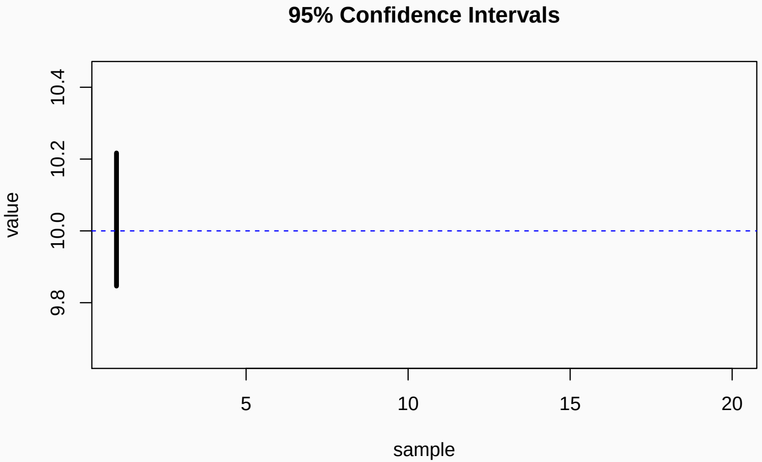
1. Draw a sample of size n.
2. Calculate a 95% confidence interval.

Then 95% of these intervals will cover the true parameter.

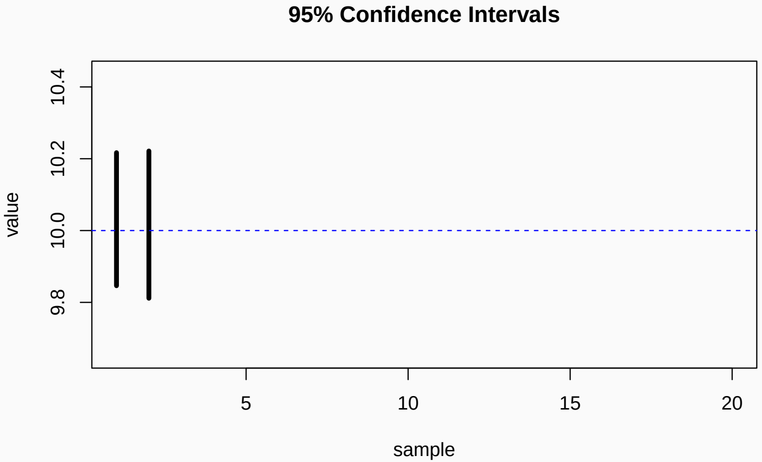
Visual

```r
mu        <- 10
sigma     <- 1
n         <- 100
simout    <- replicate(20, rnorm(n = n, mean = mu,
                                 sd = sigma))
xbar_vec  <- colMeans(simout)
s_vec     <- apply(simout, 2, sd)
lower_vec <- xbar_vec - 1.96 * s_vec / sqrt(n)
upper_vec <- xbar_vec + 1.96 * s_vec / sqrt(n)
```

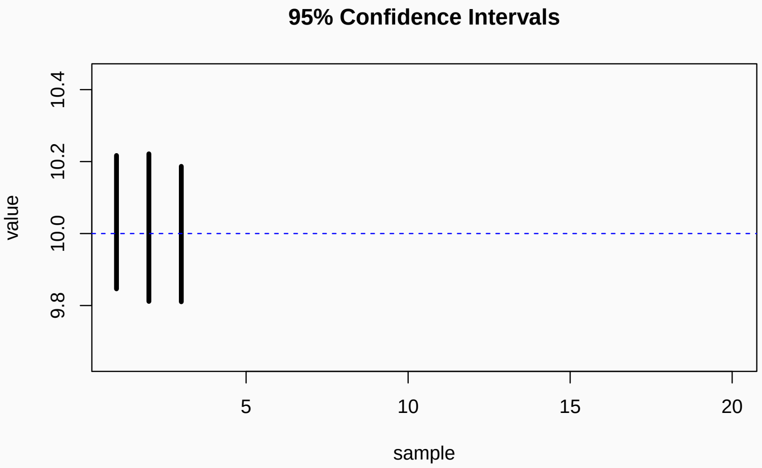
Covering True Mean i



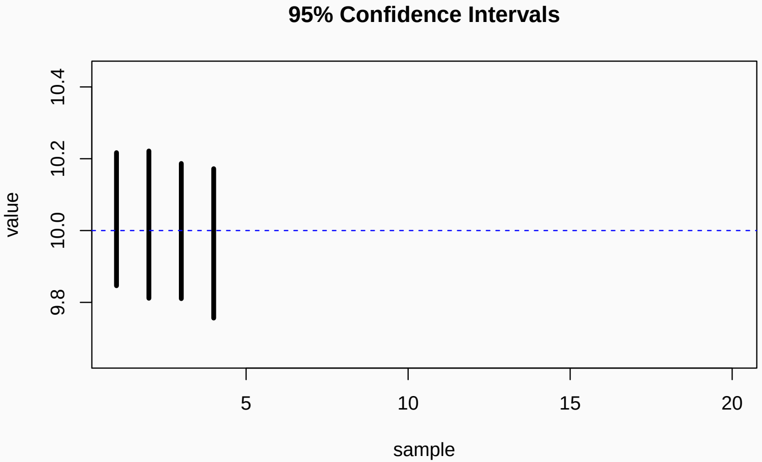
Covering True Mean ii



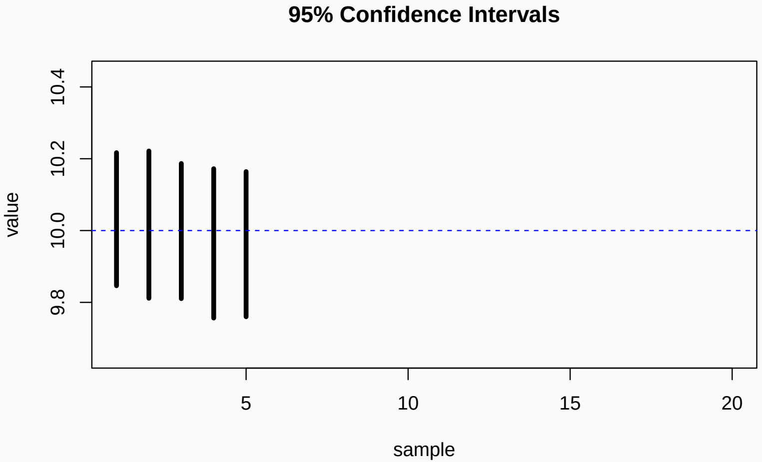
Covering True Mean iii



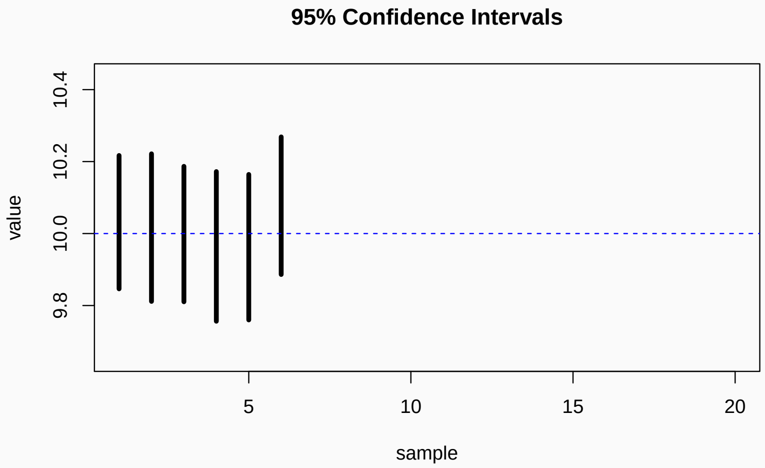
Covering True Mean iv



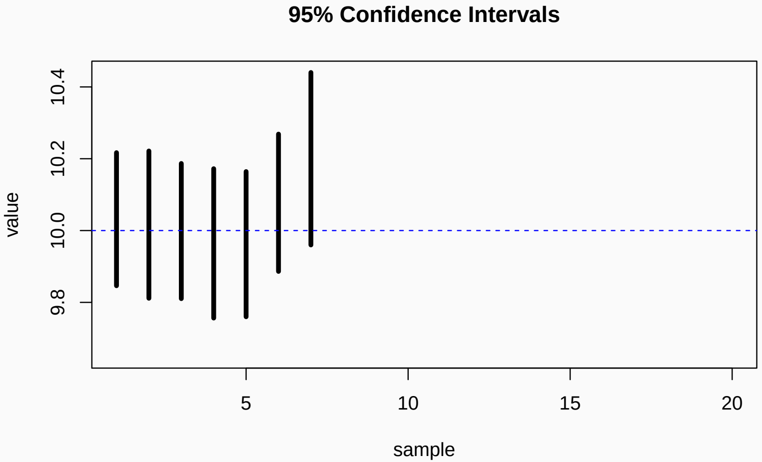
Covering True Mean v



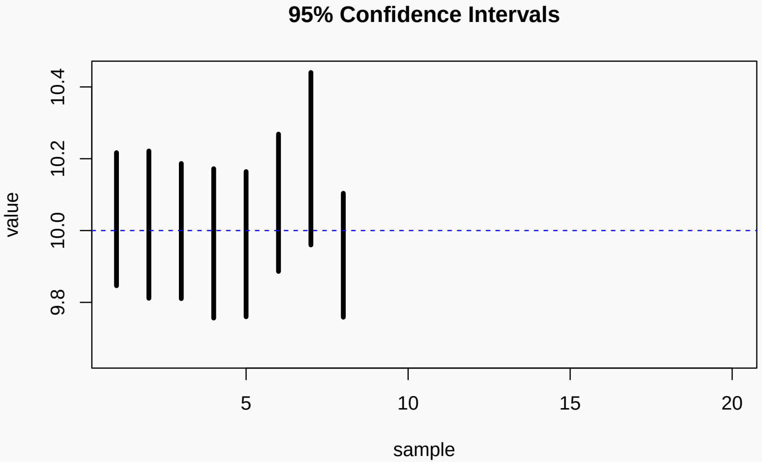
Covering True Mean vi



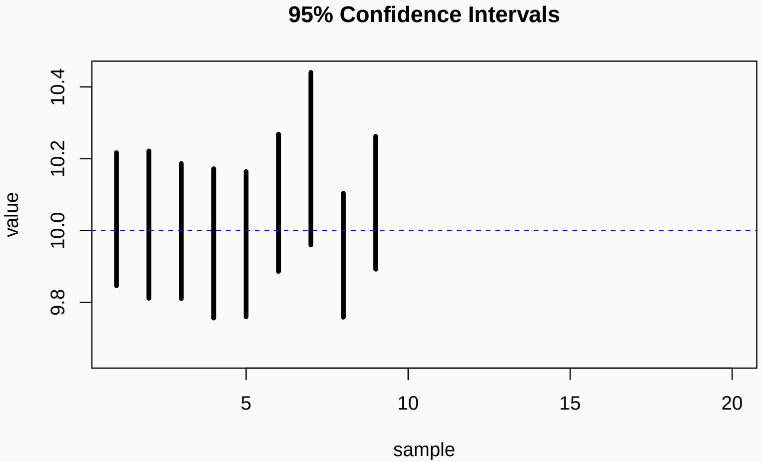
Covering True Mean vii



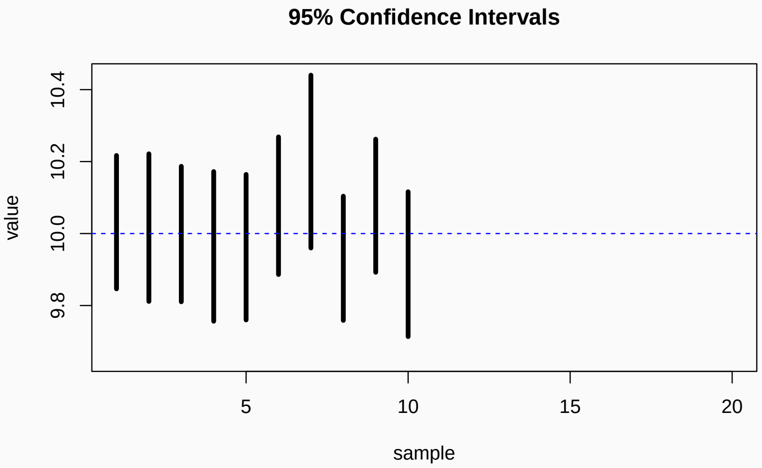
Covering True Mean viii



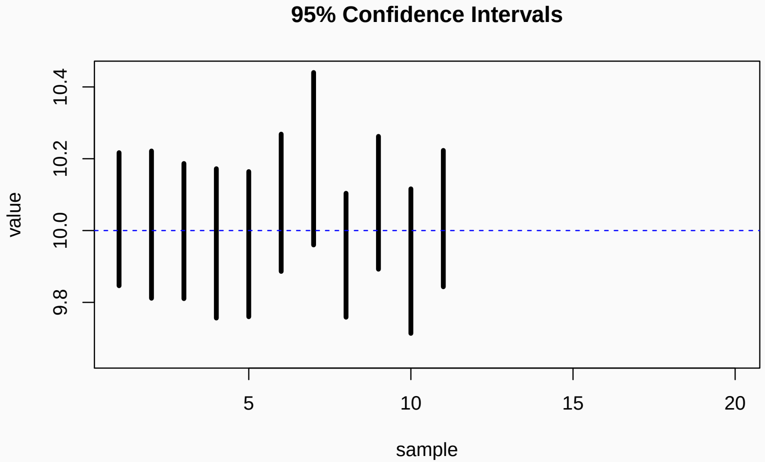
Covering True Mean ix

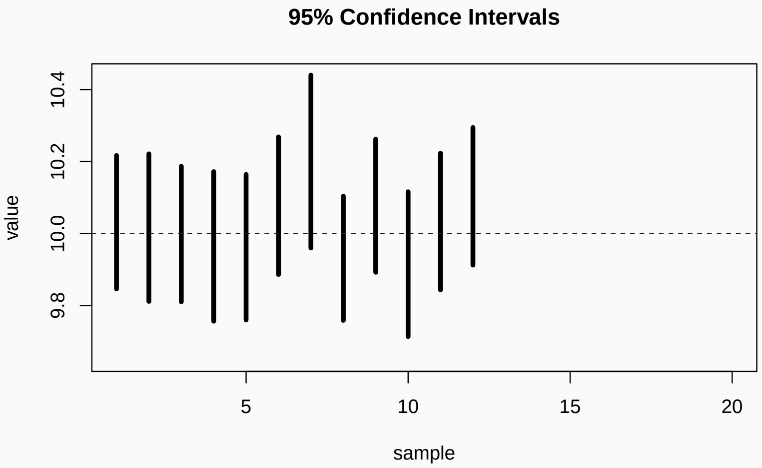


Covering True Mean x

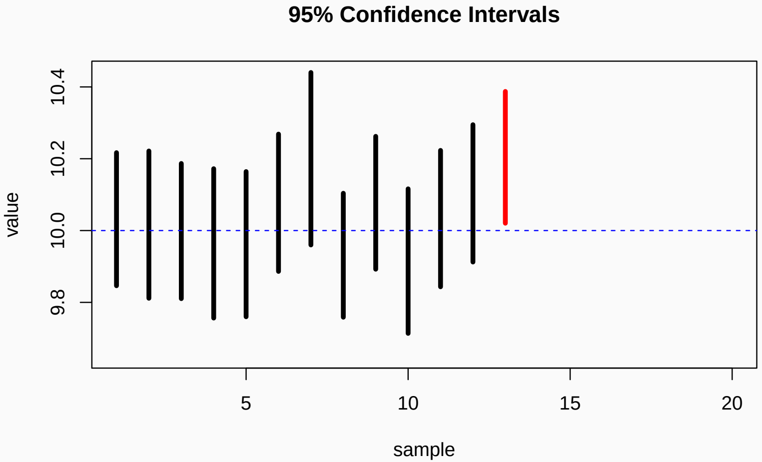


Covering True Mean xi

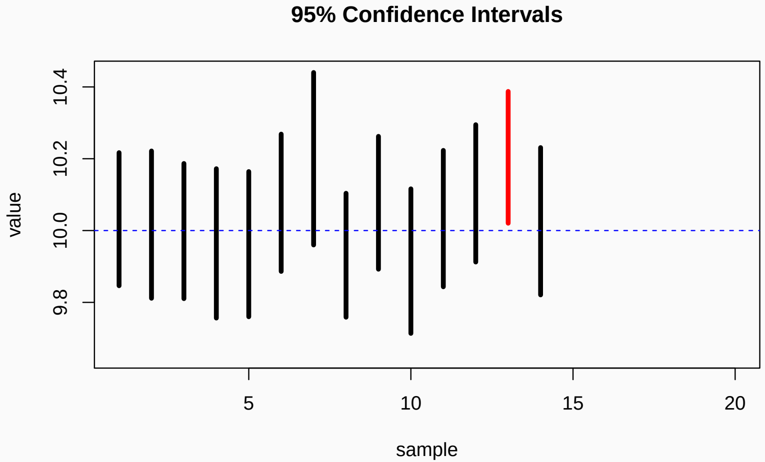



95% Confidence Intervals
value
10.4
10.2
10.0
9.8
5
10
15
20
sample

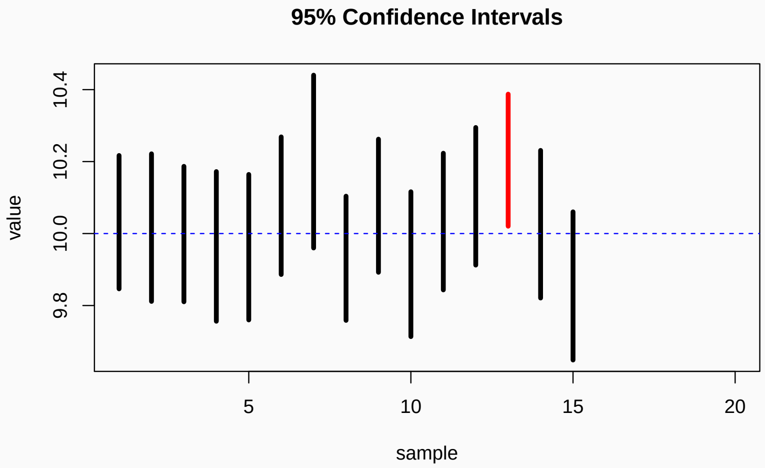
Covering True Mean xiii



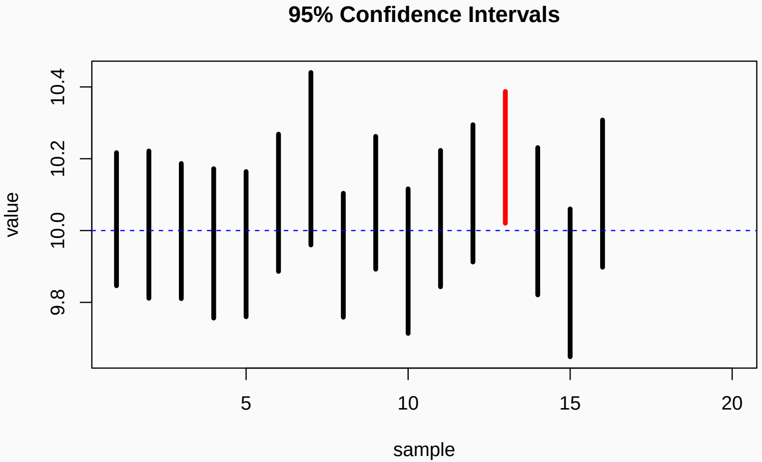
Covering True Mean xiv



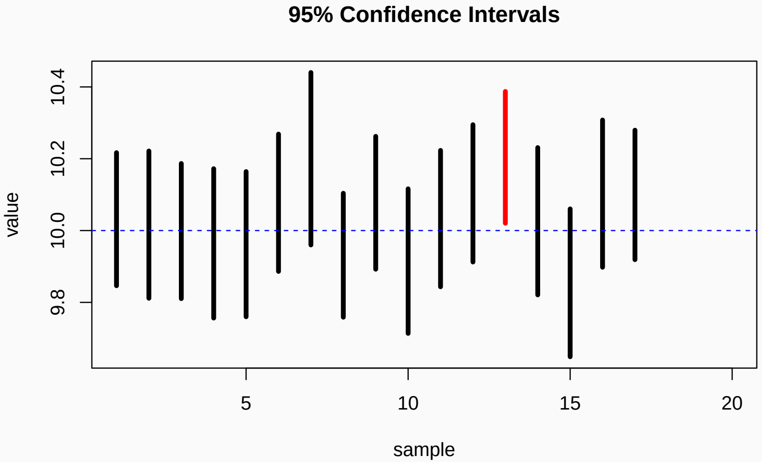
Covering True Mean xv



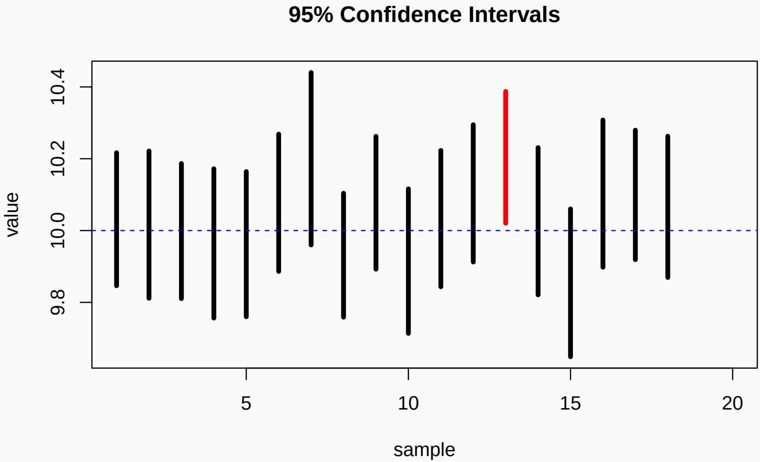
Covering True Mean xvi

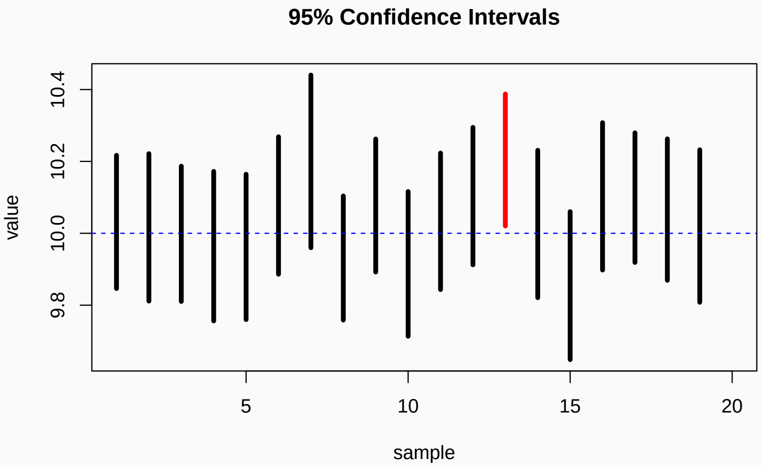


Covering True Mean xvii

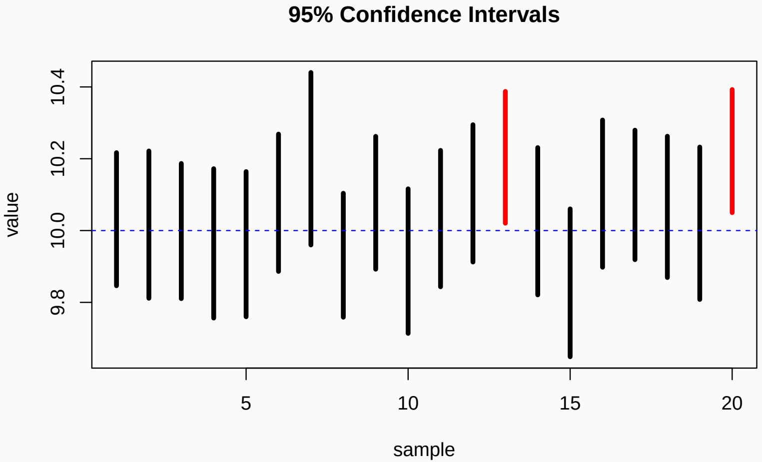


Covering True Mean xviii




95% Confidence Intervals
value
10.4
10.2
10.0
9.8
5
10
15
20
sample

Covering True Mean xx



Michaelson Experiment

- Using this procedure, a 95% confidence interval for the speed of light is $(299837, 299868)$ km/s.
- The actual speed of light is 299,792 km/s.
- Is this one of the 5% of times or is it due to bias?

Michaelson Experiment

- Using this procedure, a 95% confidence interval for the speed of light is $(299837, 299868)$ km/s.
- The actual speed of light is 299,792 km/s.
- Is this one of the 5% of times or is it due to bias?
- Probably bias since this our observed $\bar{X} = 852.4$ correponds to the 99.999999999999th percentile of a $N(792, s^2)$ distribution.
- But pretty close for 1879!

Correct/Incorrect Descriptions of CI

Let l and u be the lower and upper bounds, respectively, of a 95% confidence interval.

What does "With 95% Confidence, μ is between (l, u)" mean? Which interpretations are correct/incorrect?

1. The probability of μ being between l and u is 95%.
2. Prior to sampling, the probability of μ being between l and u is 95%.
3. 95% of the population's distribution is between l and u.
4. If we were to draw another sample, the new $\bar{X}$ would be between l and u with 95% probability.
5. 95% of new $\bar{X}$'s would lie between l and u.
6. We used a procedure that captures the true μ 95% of the time in repeated samples.

1 is wrong

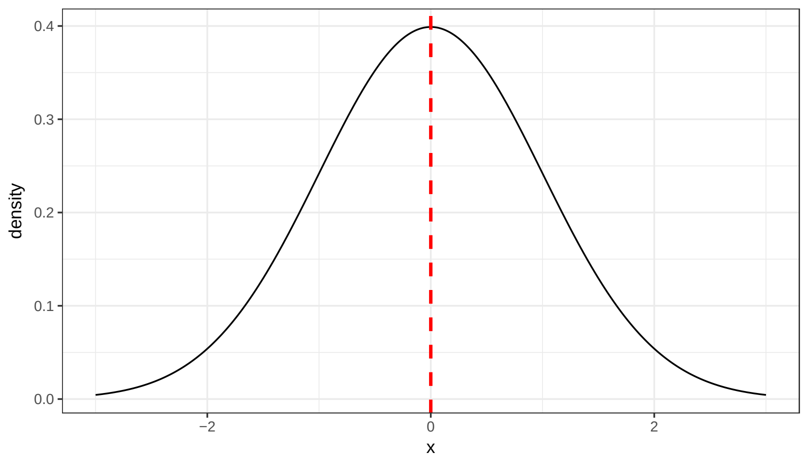
Given that we observed an interval, μ is either in the interval or it's not in the interval. Thus, the probability of μ being between l and u is either 0 or 1, but we don't know which.

2 is correct

“Prior to sampling” makes the statement correct because we haven’t yet made our interval and it is the interval that is random.

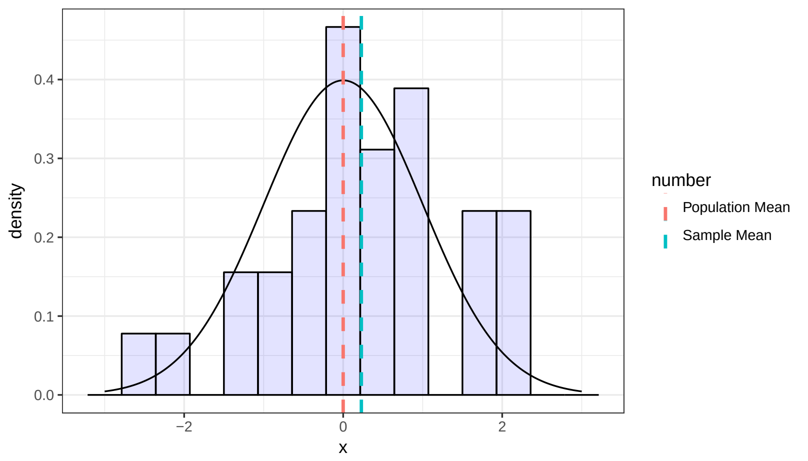
3 is wrong

Distribution of population:



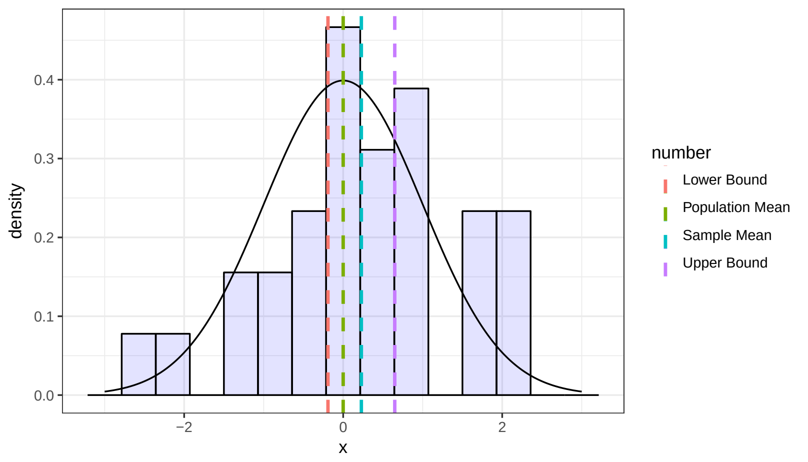
3 is wrong

Obtain a sample



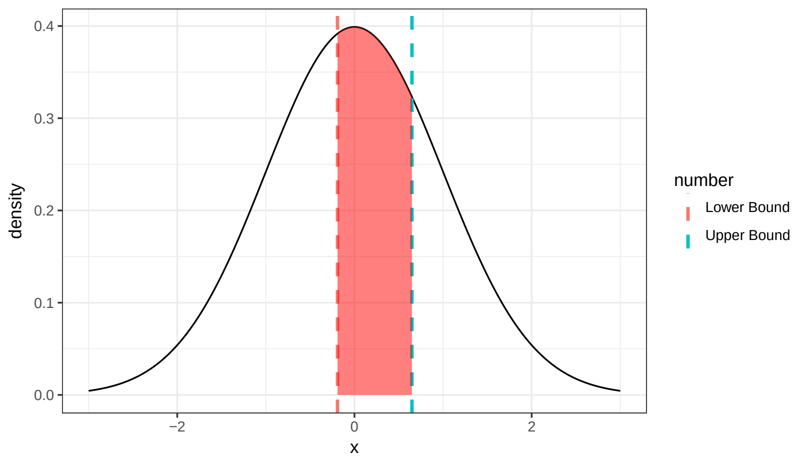
3 is wrong

Calculate confidence interval



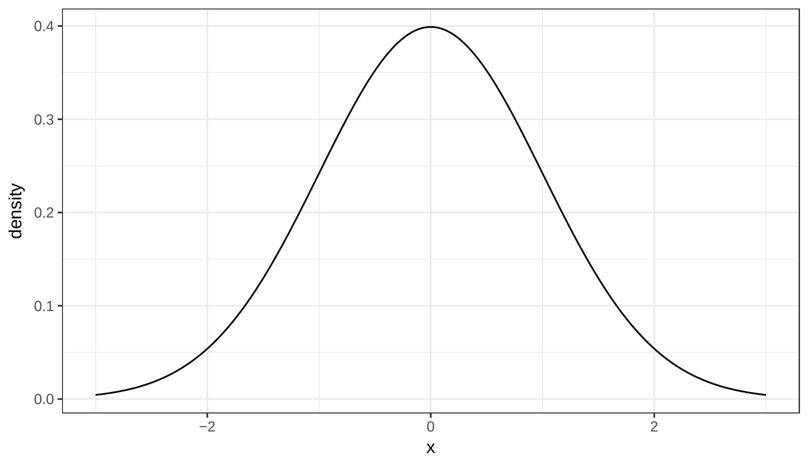
3 is wrong

95% of population is NOT within the bounds of the CI.



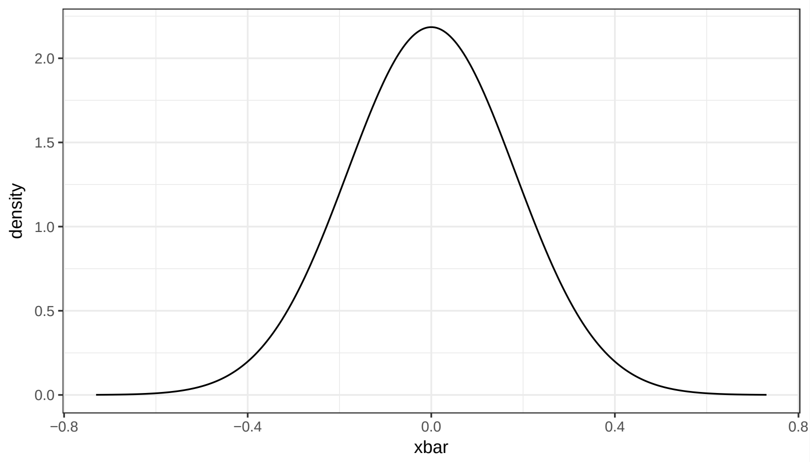
4 and 5 are wrong

Distribution of Population



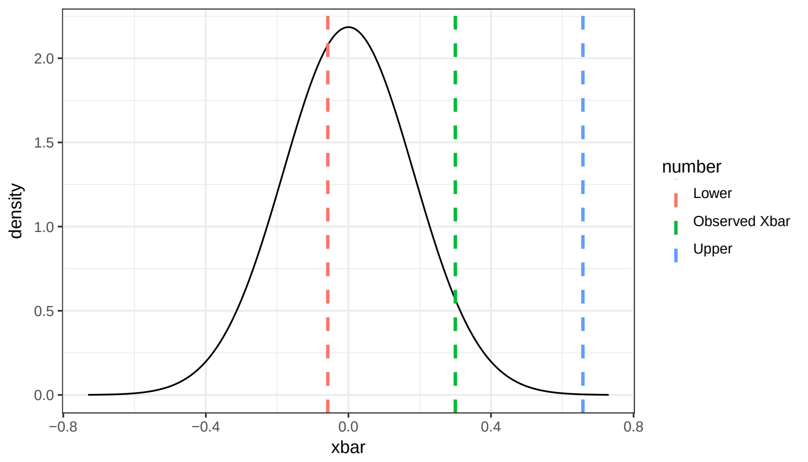
4 and 5 are wrong

Distribution of $\bar{X}$ when $n = 30$



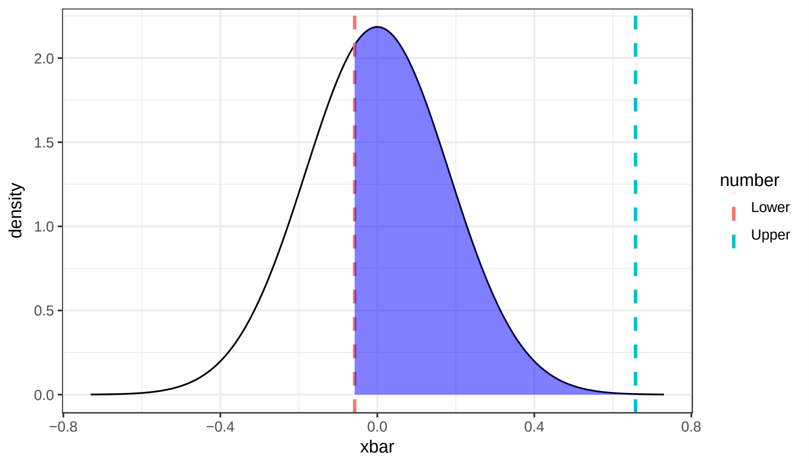
4 and 5 are wrong

What if we observed this $\bar{x}$



4 and 5 are wrong

Then 95% of future $\bar{x}$'s are not within CI bounds.



6 is correct

If we used this procedure over and over again, then 95% of the resulting CI's would capture μ.

General form of a confidence interval

In general, a CI for a parameter has the form

$$\text{estimate} \pm \text{margin of error}$$

where the margin of error is determined by the confidence level $(1-\alpha)$, the population SD σ, and the sample size n.

A $(1-\alpha)$ confidence interval for a parameter θ is an interval computed from a SRS by a method with probability $(1-\alpha)$ of containing the true θ.

For a random sample of size n drawn from a population of unknown mean μ and known SD σ, a $(1-\alpha)$ CI for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

General form of a confidence interval

Here z^* is the **critical value**, selected so that a standard Normal density has area $(1 - \alpha)$ between $-z^*$ and z^*.

The quantity $z^*\sigma/\sqrt{n}$, then, is the **margin error**.

If the population distribution is normal, the interval is *exact*. Otherwise, it is *approximately correct for large n*.

Intuition

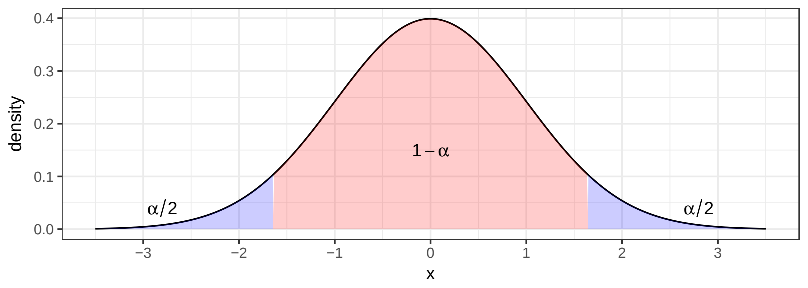
- We knew from normal theory that about 95% of $\bar{x}$'s would be within 2 standard deviations of μ.
- Suppose we want to capture μ more often (99%) or are willing to capture it less often (90%). Then we need to find how many standard deviations make it so that $\bar{x}$ is away from μ 99% of the time or 90% of the time.
- In general, we need to find the number of standard deviations so that $\bar{x}$ is away from μ about $1 - \alpha$ of the time.

General form of a confidence interval

Finding z^*

For a given confidence level $(1-\alpha)$, how do we find z^*?

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



$$P(-z^* \le Z \le z^*) = (1-\alpha) \iff P(Z < -z^*) = \frac{\alpha}{2}$$

General form of a confidence interval

Thus, for a given confidence level $(1-\alpha)$, we can look up the corresponding z^* value on the Normal table.

Common z^* values:

Confidence Level	90%	95%	99%
z^*	1.645	1.96	2.576

Some cautions on using the formula

- Any formula for inference is correct only in specific circumstances.
- The data must be a SRS from the population.
- Because $\bar{x}$ is not resistant, outliers can have a large effect on the confidence interval.
- If the sample size is small and the population is not Normal, the true confidence level will be different.
- You need to know the standard deviation σ of the population (or have a large enough sample where $s \approx \sigma$).