Stat 23400 Lecture 12: Point Estimation and Confidence Intervals

Outline for Today

- Topics:
  - The Estimation Problem (5.1 in OIS, 7.1 and 7.2 in MMSA).
  - Introduction to Confidence Interval (8.1, 8.3 in MMSA).
  - t-based Confidence Interval (8.3 in MMSA)
- Lab07: Confidence Intervals

The Estimation Problem

Estimation: Estimating the mean

- The estimation problem: once we decide the distribution to be used to model the data, there will be certain parameters that remains unknown, and thus has to be estimated.

  E.x. Life length of a part $X \sim \text{Exp}(\beta)$. With the random sample $X_1, \cdots, X_n$, we want to determine the value of β. We put a hat, $\hat{\beta}$ to denote the estimate.
- **Point estimate**: estimate the parameter with a single number.
- **Point estimator** for the mean: we can use the sample mean $\bar{X}$ to estimate the mean, $\bar{X}$ is called the estimator and $\bar{x}$ is the point estimate.

Distribution of the Sample Mean

As a funciton of n random variables, the point estimator $\bar{X}$ is a random variable. Therefore, we are interested in its distribution.

Recall that if $X_1, \cdots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, then

- $E(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
- $\hat{\mu} = \bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$.

In general (without normality assumption on X), by CLT

$$\bar{X} \dot{\sim} \mathrm{N}(\mu, \sigma^2/n)$$

Example: Estimating the Mean

It is claimed that salaries for workers in area B is the same as for those in area A.
For area A, it is known that $\mu = 31,100$ and $\sigma = 2,500$.
A random sample of $n = 30$ workers from area B are obtained, with $\bar{x} = 29,900$.

- What is $P(\bar{X} = 29,900)$?
- What is $P(\bar{X} \leq 29,900)$?
- Is the claim likely to be true?

Estimation: Estimating the variance

- Point estimator:

$$\boxed{\hat{\sigma}^2(X_1, \ldots, X_n) = S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}}.$$

- Point estimate for the variance:

$$\boxed{\hat{\sigma}^2(x_1, \ldots, x_n) = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}}.$$

- Why use (n-1) instead of n? We will show that the sample variance is an **unbiased estimator** for the population variance σ^2, i.e.,

$$\mathrm{E}(S^2) = \sigma^2 .$$

Shortcut formulas for Sample Variance

We first show the shortcut formula for sample variance, as follows.

$$\boxed{S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{(n-1)} = \frac{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2}{(n-1)}}.$$

Proof:

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)\,, \\
&= \sum_{i=1}^{n}(X_i^2) - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\,, \\
&= \sum_{i=1}^{n}(X_i^2) - 2n\bar{X}^2 + n\bar{X}^2\,, \\
&= \sum_{i=1}^{n}(X_i^2) - n\bar{X}^2\,.
\end{aligned}$$

Proof: the Sample Variance is Unbiased Estimator

Definition

An estimator $\hat{\theta}$ is an **unbiased estimator** of θ if

$$\mathrm{E}(\hat{\theta}) = \theta\,.$$

Proof of sample variance: We only need to show $\mathrm{E}(S^2) = \sigma^2$.
First, note that

$$\mathrm{E}(X_i^2) = \sigma^2 + \mu^2\,,$$

$$\mathrm{E}(\bar{X}^2) = \mathrm{Var}(\bar{X}) + \big(\mathrm{E}(\bar{X})\big)^2 = \frac{\sigma^2}{n} + \mu^2\,.$$

Plugging these in the shortcut formula for sample variance, we have

$$(n-1)\mathrm{E}(S^2) = n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) = (n-1)\sigma^2\,.$$

Confidence Intervals

Confidence intervals: Introduction

Suppose we go out for fishing. Which way do you have a better chance to catch a fish, using a spear or a net?

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

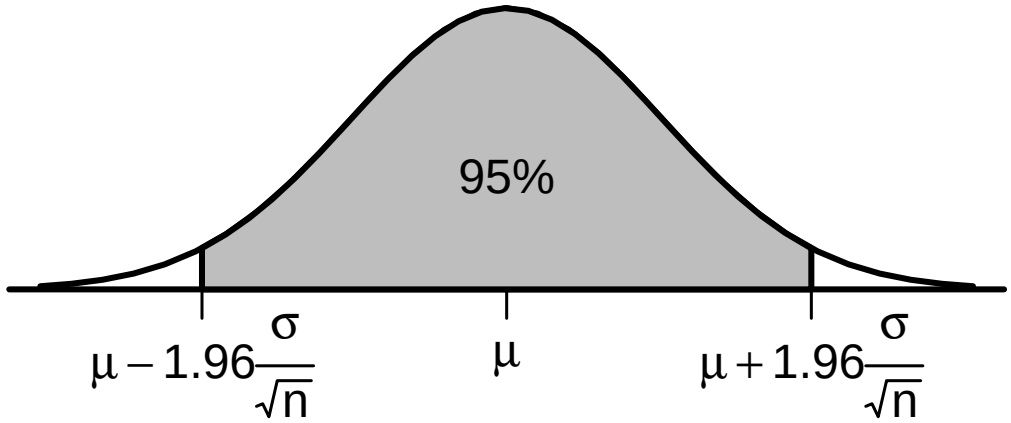
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear. If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter. A plausible range of values for the population parameter is called a **confidence interval** (C.I.).
- Using a confidence interval is like fishing with a net.

C.I. for the Population Mean

- Recall CLT: for large n,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- For a normal curve, 95% of its area is within 1.96 SDs from the center. That means, **for 95% of the time, $\bar{X}$ will be within $1.96\frac{\sigma}{\sqrt{n}}$ from μ.**



- We will show later that **for 95% of the time, the interval**

$$(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

will cover the true population mean μ.

- Hence, we call this interval

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

a **95% confidence interval for** μ.

C.I. Example: 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.)

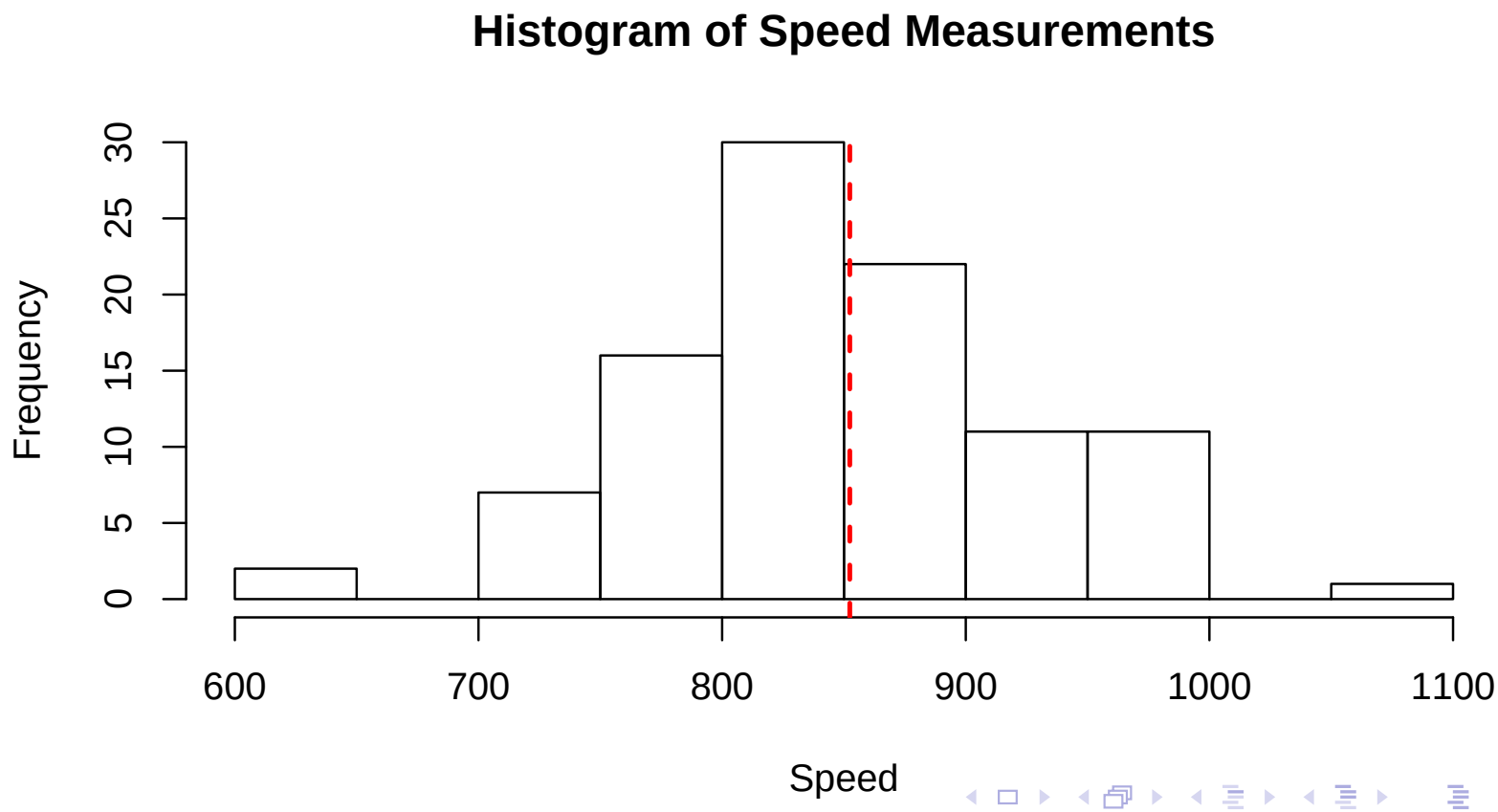
```
library(tidyverse)
data("morley")
glimpse(morley)
```

```
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)
```

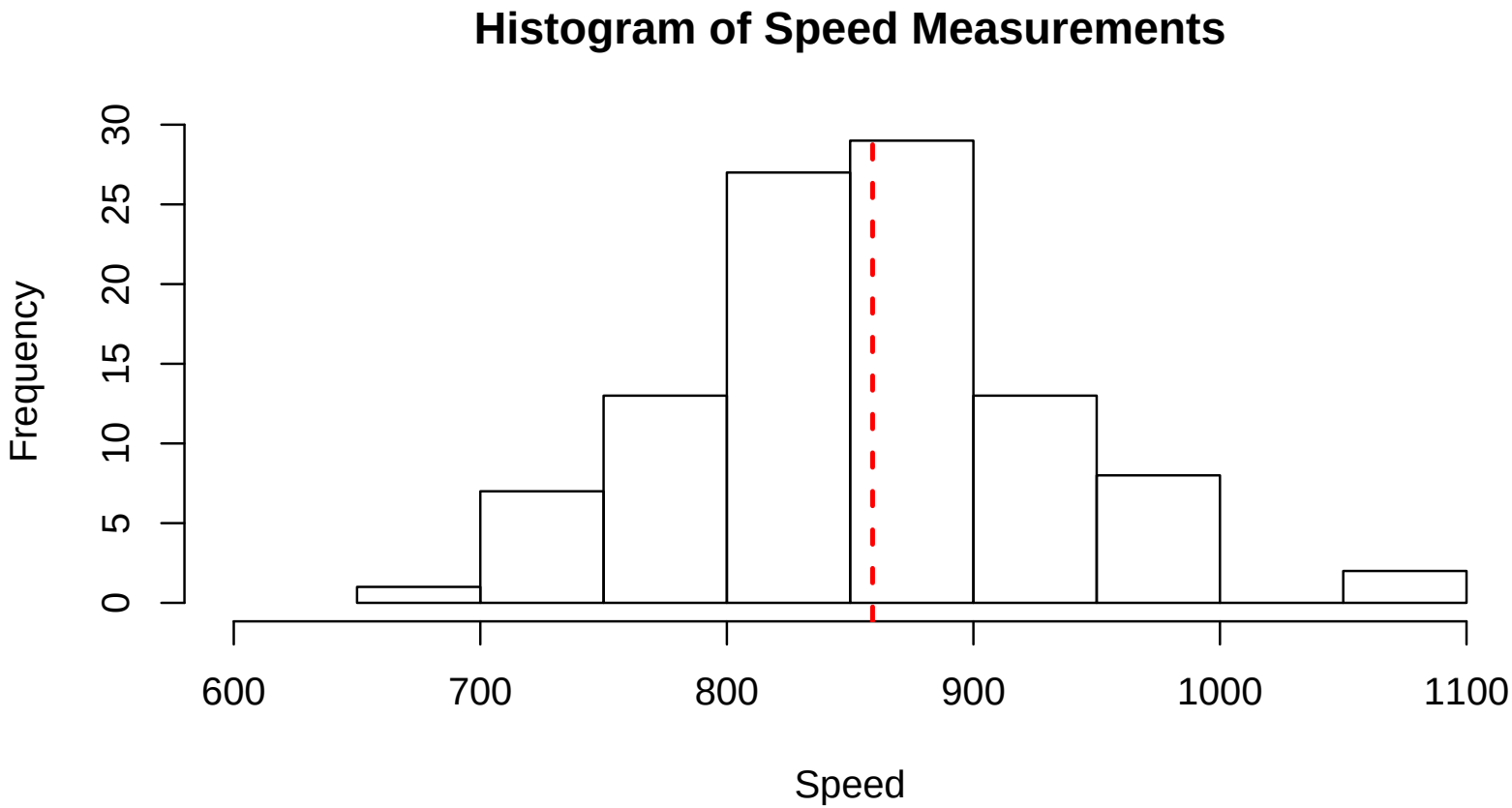


What can we say

If this experiment were done with no bias, then:

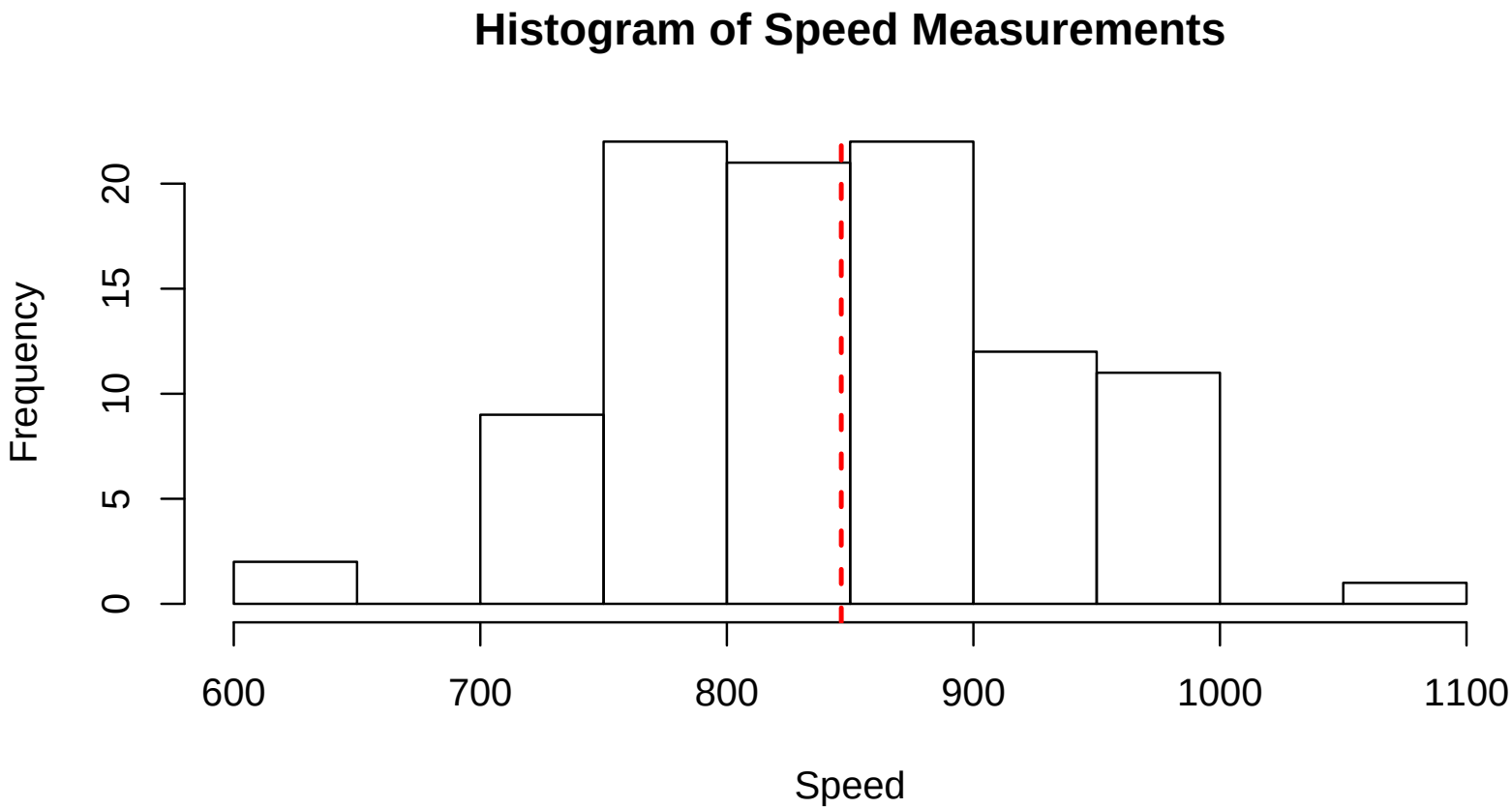
- $E[\bar{X}] = \mu$
- $SD(\bar{X}) = \sigma/\sqrt{n}$
- $\bar{X} \underset{n\to\infty}{\longrightarrow} \mu$ (Law of Large Numbers)
- $\bar{X} \sim N(\mu, \sigma^2/n)$, approximately (Central Limit Theorem).
- Right now, our point estimate for the value of μ is $\bar{X} = 852.4$.

A different sample



$\bar{x} = 861.7$

A different sample



$\bar{x} = 849.5$

Example: C.I. for the mean

- 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 $\bar{X}$.

$$\bar{X} \sim N(\mu, \sigma^2/n)\,.$$

- We thus have

$$P\left(\mu - 1.96\sigma/\sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma/\sqrt{n}\right) = 0.95$$

Rearranging terms we get

$$P\left(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}\right) = 0.95.$$

- That is, the *random interval* $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ 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).
- For now, let's assume σ is known.

Calculating 95% Confidence Intervals for Mean

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

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

Intuition/Interpretation of CI

What if we repeat the following over and over again:

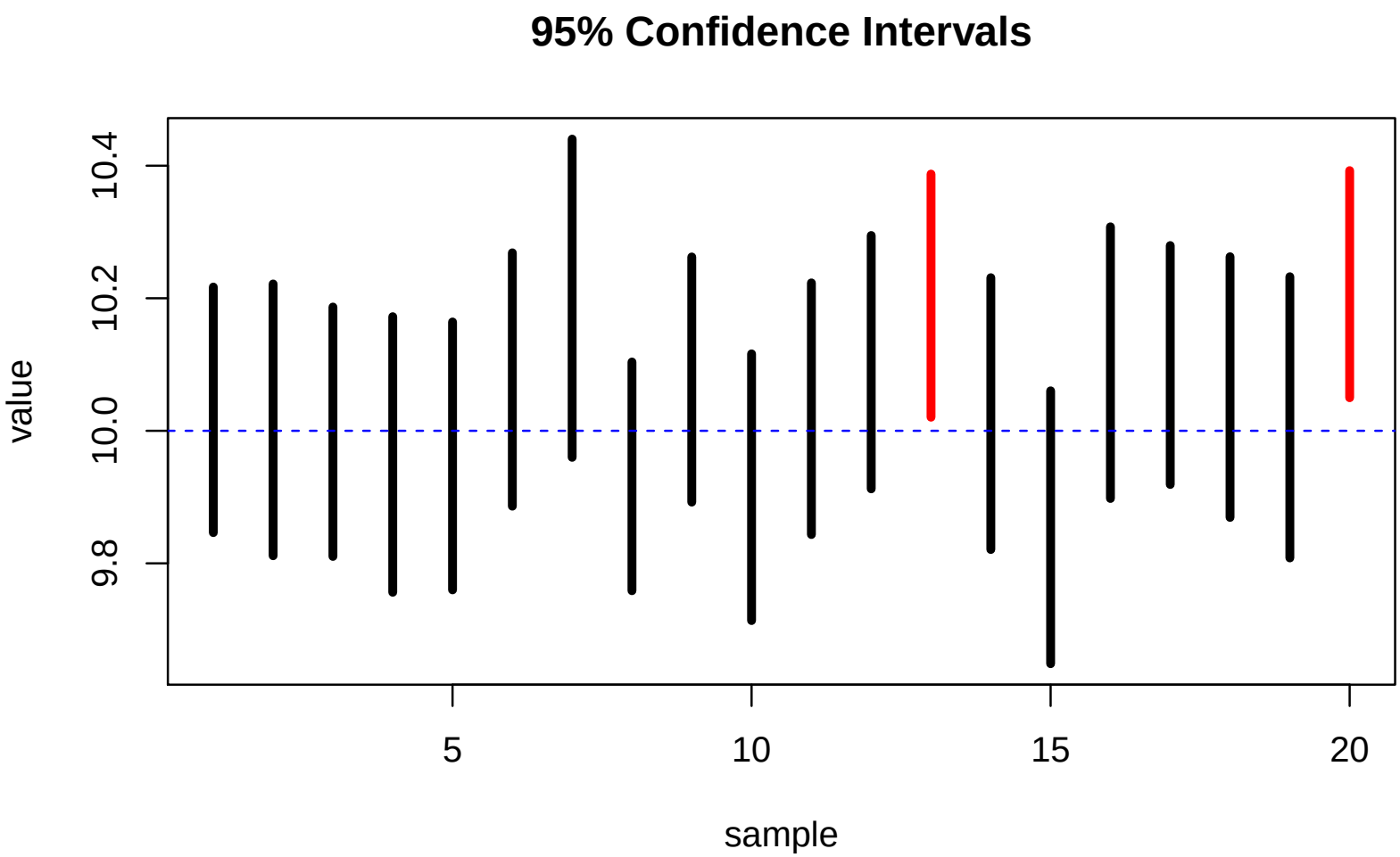
- Draw a sample of size n.
- Calculate a 95% confidence interval.

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

Visualization

```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



What does "95% confidence" mean?

What is the thing that has a 95% chance to happen?

- It is the **procedure to construct the 95% interval**.
- About 95% of the intervals constructed following the procedure (taking a SRS and then calculating $\bar{X} \pm 1.96\, s/\sqrt{n}$) will cover the true population mean μ.
- After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn't. We don't know.
- Just like lottery, before you pick the numbers and buy a lottery ticket, you have some chance to win the prize. After you get the ticket, you either win or lose.

Michaelson Experiment

- Using this procedure, a 95% confidence interval for the speed of light is $(299837, 299868)$ km/s.
- The actual speed of light we know as of today is 299,792 km/s.
- Is this one of the 5% of times or is it due to bias?
- Probably bias since this observed $\bar{x} = 852.4$ corresponds to the 99.999999999999th percentile of a $N(792, s^2)$ distribution.

Correct/Incorrect Descriptions of CI

Let l and u be the lower and upper bounds, respectively, of a 95% confidence interval for a sample.
Which interpretations are correct/incorrect?

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

General form of a confidence interval

- In general, a CI for a parameter has the form

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

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

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

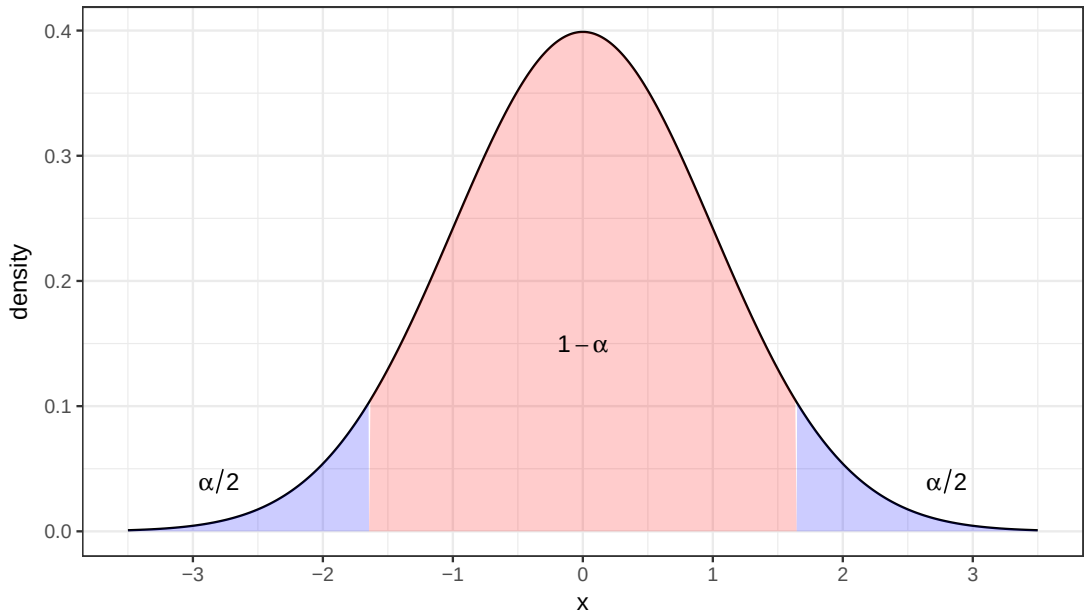
- z^* is called the **critical value**, and $z^* \frac{\sigma}{\sqrt{n}}$ is the **margin of error**.
- If the population distribution is normal, the interval is *exact*. Otherwise, it is *approximately correct for large n*.

Finding Critical Value z^*

For a given confidence level $(1-\alpha)$, how do we find z^*? Let $Z \sim N(0,1)$.

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

$$z^* = z_{\frac{\alpha}{2}}, \text{ where } P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$



Finding Critical Values

For a given confidence level $(1-\alpha)$, we can look up the corresponding $z_{\frac{\alpha}{2}}$ value on the Normal table.

Common $z_{\frac{\alpha}{2}}$ values:

α	0.1	0.05	0.01
$\frac{\alpha}{2}$	0.05	0.025	0.005
Confidence Level	90%	95%	99%
$z_{\frac{\alpha}{2}}$	1.645	1.96	2.576

R code to find z_α:

- `qnorm(1-`α`)`
- `qnorm(`α`, lower.tail = FALSE)`
- `abs(qnorm(`α`))`

Some cautions on using the formula

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- The data must be a simple random sampling from the population.
- Because $\bar{x}$ is not robust, 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 $s \approx \sigma$).

t-Confidence Interval: Variance is unknown

Confidence Interval: Variance Unknown

$$\boxed{\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

- This CI is valid only if the variance σ^2 is known.
- Most of the time, σ^2 is not known.
- If n is large enough, we can replace σ with s and the CI is still approximately correct. Mainly because of the Law of the Large Numbers

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \underset{n \to \infty}{\longrightarrow} \sigma^2$$

Example: Average number of exclusive relationships

E.x. A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

Problem

However, for small n (rule of thumb $n \leq 30$), this approximation is not accurate! Not even when the $X_1, X_2, \ldots, X_n$ are exactly $N(\mu, \sigma^2)$!

Note:

To perform inference with small n, we will require that the X_i's are well approximated by a normal distribution.

Recall that for $X_1, X_2, \ldots, X_n$, independent with $X_i \sim N(\mu, \sigma^2)$, we have exactly

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

But we want the distribution of

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

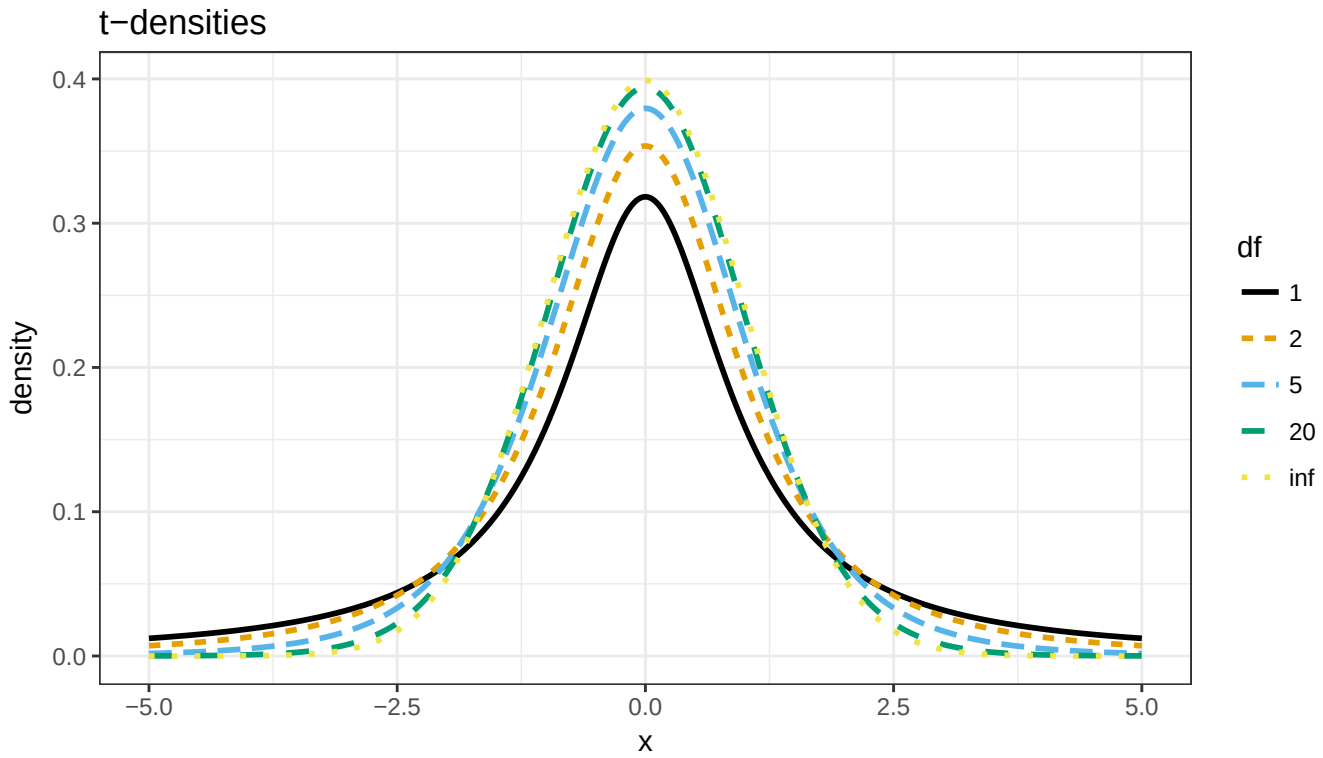
Theorem

$X_1, X_2, \ldots, X_n$, independent with $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where t_ν represents the t-distribution with ν degrees of freedom.

Properties of t



- Only one parameter - the degrees of freedom ($\nu > 0$)
- Symmetric about 0
- Bell-shaped - similar to normal distribution
- More spread out than normal - heavier tails
- As ν increases, converges to the Normal distribution.

t Probability Table (p.430-431 in OIS)

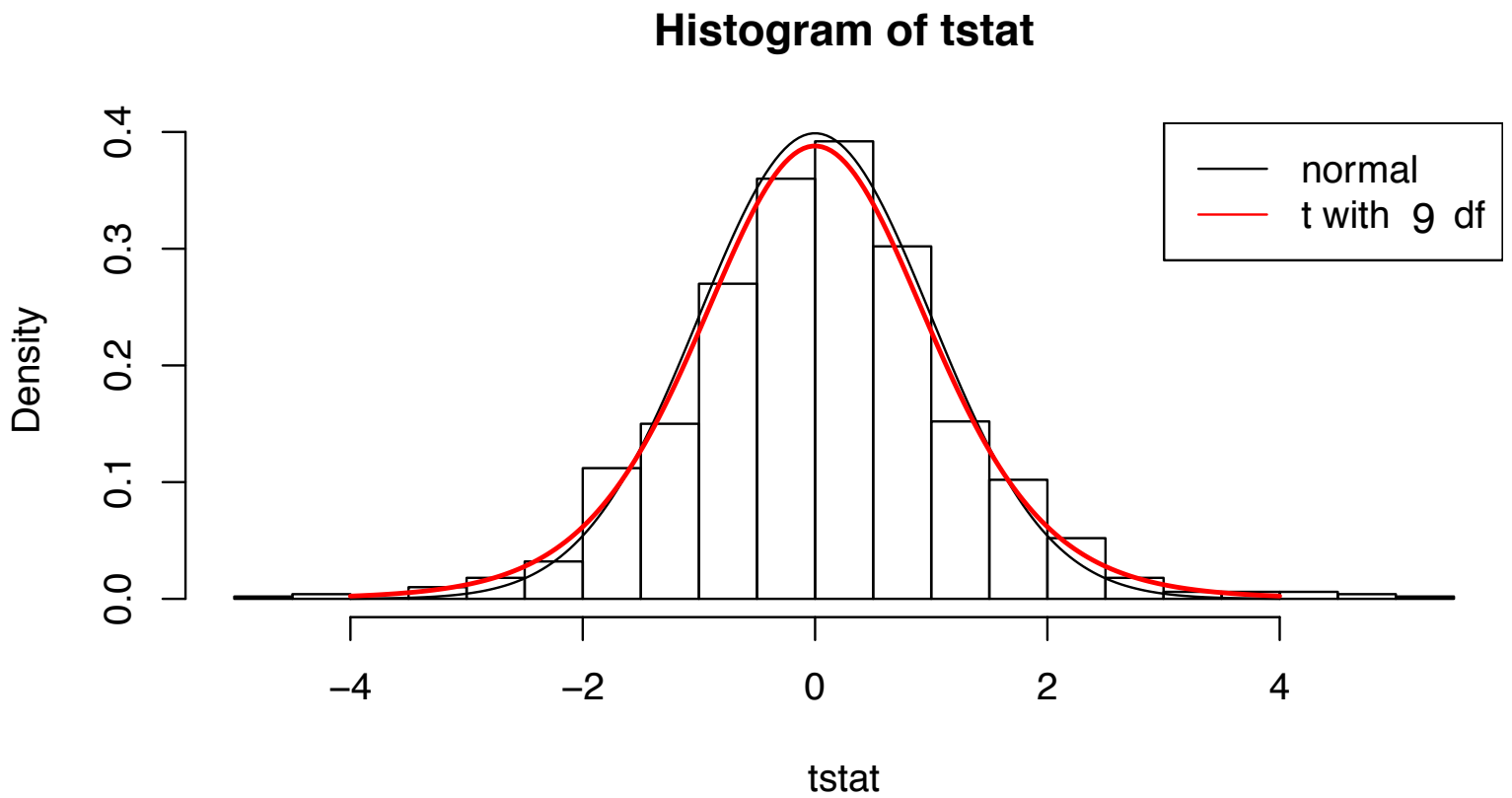
one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

Empirical Example

```
x_matrix <- replicate(1000, rnorm(10))
xbar     <- colMeans(x_matrix)
s        <- apply(x_matrix, 2, sd)
tstat    <- xbar / (s / sqrt(10))
```

Histogram of t-statistics

t-distribution fits better in the tails



Confidence Intervals with Unknown σ

- Now let us derive the confidence intervals with unknown σ, and possibly small n.
- The goal is to find a confidence interval for μ when σ is unknown.
- That is, we want a random interval that captures μ in $(1 - \alpha)$ of repeated samples.

Confidence Intervals with Unknown σ

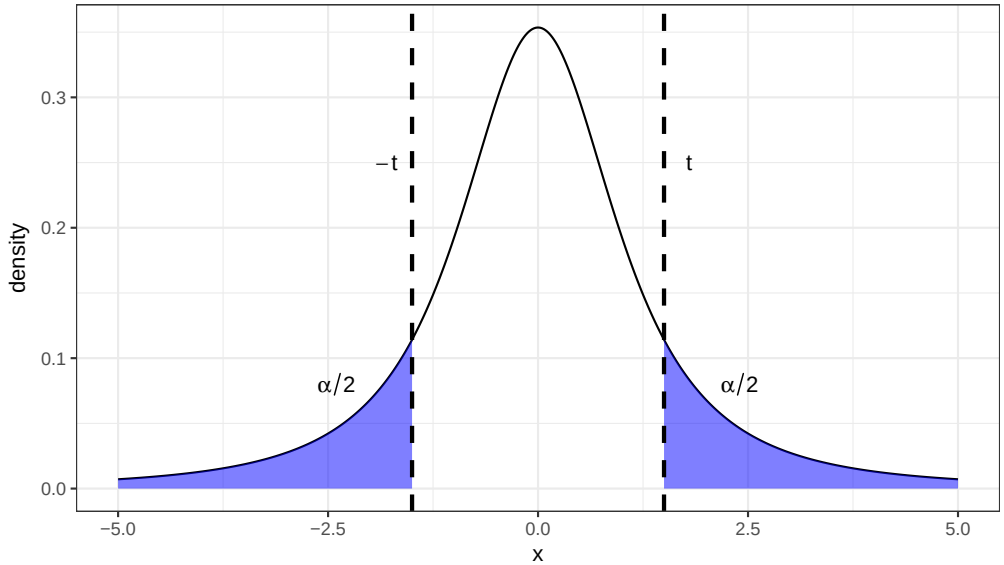
When σ is unknown, we obtained a $(1-\alpha)$ confidence interval for the mean as

$$\boxed{(\bar{X} - t^* \frac{S}{\sqrt{n}}, \bar{X} + t^* \frac{S}{\sqrt{n}})}$$

- This works for any sample size n, not just when n is small.
- For large n, it will approximately equal the normal-based CI.
- These confidence intervals are again random. In addition to having a random center $\bar{X}$, they have a random width $t^* \frac{S}{\sqrt{n}}$.
- The t intervals are wider than the normal intervals because the t distribution has larger tails. This corrects for the uncertainty in estimating σ.

How do you get t^*?

The critical value, $t^* = t_{n-1,\alpha/2}$ is chosen such that $100(1-\alpha)\%$ of the area under the t_{n-1} density lies between $-t^*$ and t^*.



You can use the R function qt to find $t_{n-1,\alpha/2}$.

- `qt(`$1-\alpha/2$`, df = n-1)`
- `qt(`$\alpha/2$`, df = n-1, lower.tail = FALSE)`
- `abs(qt(`$\alpha/2$`, df = n-1))`

Some notes on Approximation

1. If the underlying population is Normally distributed, the interval is exact. (i.e. exact if $X_1, X_2, \ldots, X_n$ are $N(\mu, \sigma^2)$).
2. Otherwise, the interval is approximately correct if n is not too small (say, $n \geq 15$), the data are not strongly skewed (and there are no outliers).
3. With n sufficiently large (say $n \geq 30$), the approximation is correct even if the data are clearly skewed.

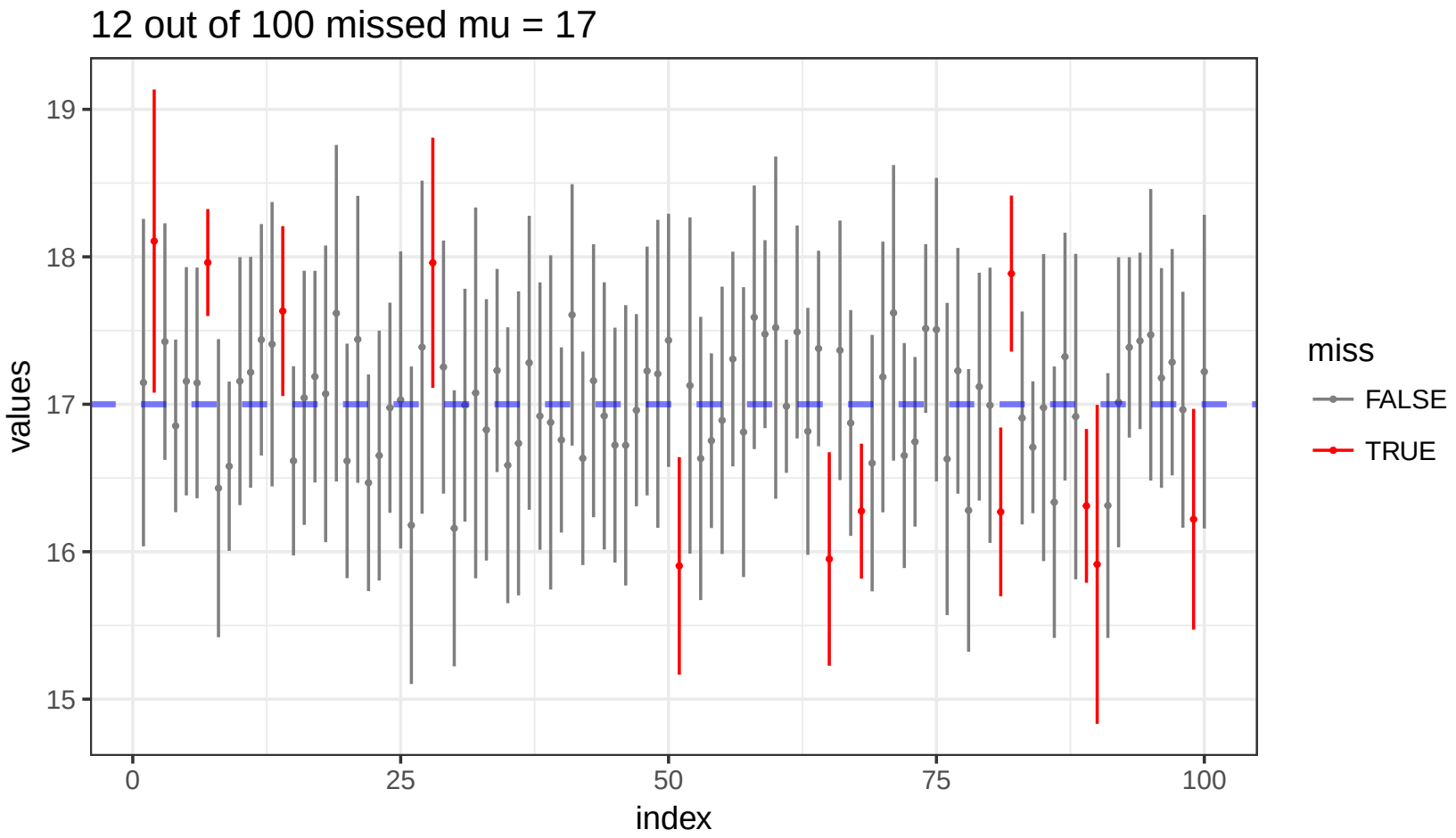
Does t-based CI really matter?

Simulate 100 samples and calculate their corresponding normal and t 95% confidence intervals:

```
mu       <- 17
sigma2   <- 2
n        <- 10
alpha    <- 0.05
x_matrix <- replicate(100, rnorm(n, mu, sqrt(sigma2)))
xbar     <- colMeans(x_matrix)
s        <- apply(x_matrix, 2, sd)
z_alpha  <- abs(qnorm(alpha / 2))
t_alpha  <- abs(qt(alpha / 2, df = n - 1))
lower_z  <- xbar - z_alpha * s / sqrt(n)
upper_z  <- xbar + z_alpha * s / sqrt(n)
lower_t  <- xbar - t_alpha * s / sqrt(n)
upper_t  <- xbar + t_alpha * s / sqrt(n)
```

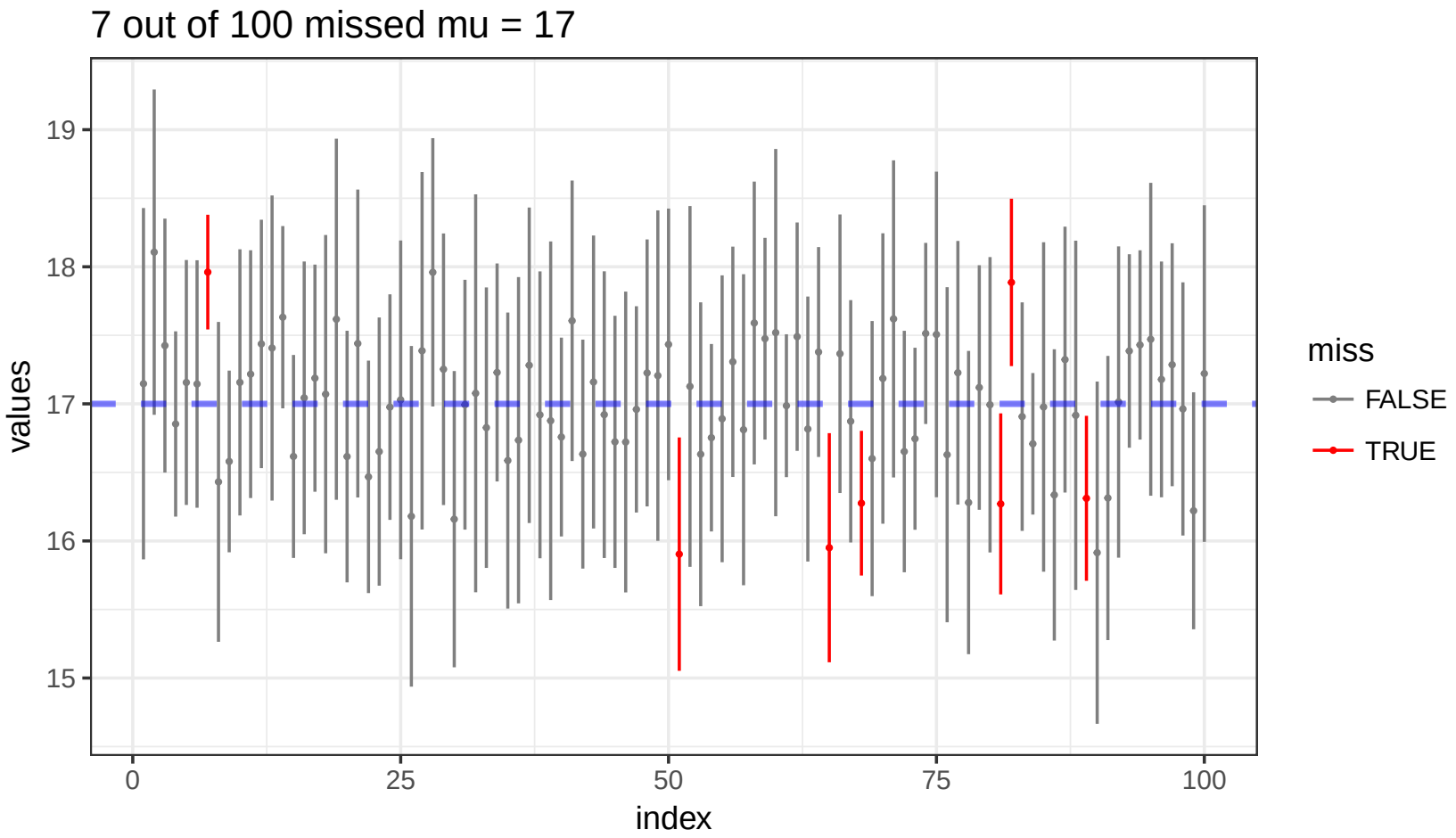
Does this really matter?

Normal based intervals



Does this really matter?

t based intervals



Example: Thermal Conductivity of Glass

- Thermal Conductivity is measured in terms of watts of heat power transmitted per square meter of surface per degree Celsius of temperature difference on the two sides of the material.
- In these units, glass has conductivity about 1.
- The National Institute of Standards and Technology provides exact data on properties of materials. Here are measurements of the thermal conductivity of 11 randomly selected pieces of a particular type of glass:

  `1.11, 1.07, 1.11, 1.07, 1.12, 1.08,`
  `1.08, 1.18, 1.18, 1.18, 1.12`

- Find a 95% CI for the mean conductivity of this type of glass.

Example: Thermal Conductivity of Glass — t-Statistic

The sample mean and sample SD are

$$\bar{x} \approx 1.1182, \qquad s \approx 0.04378$$

```
> conduct = c(1.11,1.07,1.11,1.07,1.12,1.08,1.08,1.18,1.18,1.18,1.12)
> mean(conduct)
[1] 1.118182
> sd(conduct)
[1] 0.04377629
```

95% t-based C.I.

$n = 11$, $df = 11 - 1 = 10$

The critical value $t_{df,\alpha/2}$ is at the intersection of row $df = 10$ and two tail probability 0.05.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

95% CI for the mean conductivity of this type of glass is

$$\begin{aligned}\bar{x} \pm t_{10,0.05/2}\frac{s}{\sqrt{n}} &= 1.1182 \pm 2.23 \times \frac{0.04378}{\sqrt{11}} \\ &= 1.1182 \pm 0.0294 \\ &= (1.0888, 1.1476)\,.\end{aligned}$$

Summary: Z-Interval for mean, μ

- If σ^2 is known, then for normal random variables, $(1-\alpha)$ C.I. is

$$\boxed{\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).}$$

- If σ^2 is known and sample size is big, by CLT, $(1-\alpha)$ C.I. is

$$\boxed{\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).}$$

- If σ^2 is unknown and sample size is large, $(1-\alpha)$ C.I. is

$$\boxed{\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right).}$$

S is the sample SD.

Review: Confidence Interval for mean μ (σ unknown)

If σ^2 is unknown and n is small, t-based $(1-\alpha)$ C.I. for μ is

$$\boxed{\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right).}$$

- S is the sample SD.
- If the underlying population is normally distributed, the t-based CI is exact.
- Otherwise, the interval is approximately correct if n is not too small (e.g., $n \geq 15$), the data are not strongly skewed (and there are no outliers).
- With n sufficiently large (e.g., $n \geq 30$), the t-based C.I. is approximately accurate even if the data are clearly skewed.