Lecture 8: Estimation and Confidence Intervals

STAT 324

Ralph Trane

University of Wisconsin–Madison

Spring 2020

Estimation

The art of coming up with our "best guess" for the truth. This is called *an estimate*.

An *estimator* is a function that takes values from a sample and provides an estimate ("best guess").

Estimation

In order to come up with a good estimator, it is important to know how the sample was gathered. Three important definitions:

- a sample is called a **simple random sample** (SRS) if every possible element is equally likely to be sampled
    - unless otherwise stated, **all samples in this class are SRS**
- a sample is drawn **with replacement** if an element is replaced to the population before the next element is drawn. Otherwise, we say it is drawn **without replacement**.
    - without replacement = each element can only be sampled once.
- a collection of RVs $X_1, X_2, \ldots, X_n$ are said to be **independent and identically distributed** (iid) if
    1. they are all independent of each other
    2. they all follow the same distribution.

Technically, SRS **ONLY** if done with replacement. However, if population is "big enough", sample without repleacement is approximately the same as with replacement. (Recall last weeks discussion.)

Estimation: Population Mean

- A car manufacturer uses an automatic device to apply paint to engine blocks.
- engine blocks get very hot, so the paint must be heat-resistant,
- important that the amount applied is of a minimum thickness
- warehouse contains thousands of blocks painted by the automatic device
- he manufacturer wants to know the average amount of paint applied by the device

16 blocks will be selected at random, and the paint thickness measured in mm. Let $X_1, \ldots, X_{16}$ be RVs indicating the thickness of the 16 blocks.

Let's assume these RVs are iid -- i.e. independent and identically distributed. There exists some true expected value of these: $E(X_i) = \mu$. There also exists some true variance, $\mathrm{Var}(X_i) = \sigma^2$.

Estimation: Population Mean

Now we actually observe 16 realizations of these RVs:

```
paint_thickness <- data.frame(
  thickness = c(1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83,
                1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49)
)
```

Using these data, what would be your "best guess" for the true mean μ?

I would use the sample average:

```
paint_thickness %>%
  summarize(Mean = mean(thickness))
```

```
##       Mean
## 1 1.348125
```

This **estimATE** comes from the **estimatOR**: $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Notice: the **estimator** is a **RV** while the **estimate** is a **realization** of that RV.

Estimation: Population Mean

Since $\bar{X}$ is an RV, we can talk about the expected value and variance of it:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \mu, \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

$$\mathrm{SD}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}.$$

$\mathrm{SD}(\bar{X})$ is often called SE (Standard Error)

Estimation: Good estimators

What makes an estimator a good estimator? We will consider two properties of estimators:

- Unbiasedness:
  - an estimator is said to be **unbiased** if $E(\hat{\theta}) = \theta$.
  - I.e. the estimator gives us the correct value *on average*
  - the **bias** of an estimator is $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- Shrinking/small variance.
  - unbiased estimator with huge variance is not very reliable
  - if choosing between multiple unbiased estimators, choose smallest variance!

Estimation: Good estimators

Example of good estimator: $\hat{\mu} = \bar{X}$.

- Unibased.
- Can be shown to have smallest variance of ALL unbiased estimators...

Other examples of good estimators:

- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
  .
  - Actually unbiased
    - if divided by n, would not be!
- $\hat{\sigma} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
  .
  - Biased!

Remember, we have

- sample variance:
  $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- population variance:
  $\sum_{i=1}^{n} P(X = x_i)(x_i - E(X))^2$

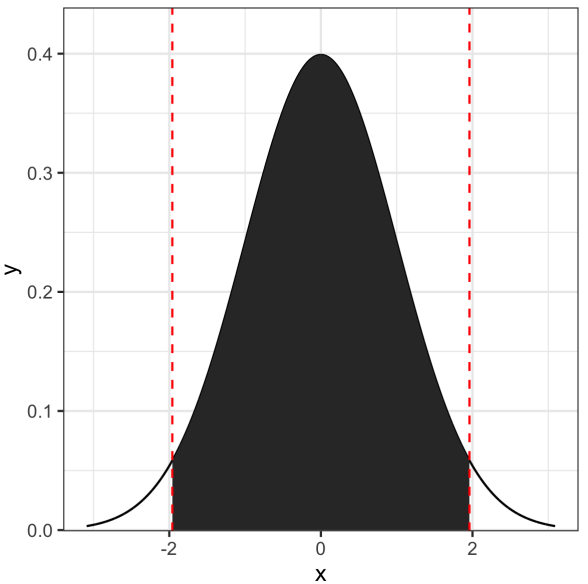
Estimation: Confidence Intervals

If X_i's are iid $N(\mu, \sigma^2)$, what is the distribution of $\bar{X}$?

$\bar{X} \sim N(\mu, \sigma^2/n)$. So what is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$?

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

Now, we can find values x_1, x_2 such that $P(x_1 \leq Z \leq x_2) = 1-\alpha$.
Let's for simplicity use $\alpha = 0.05$. I.e. we want to find x_1, x_2 such that this area is 0.95.



Estimation: Confidence Intervals

If X_i's are iid $N(\mu, \sigma^2)$, what is the distribution of $\bar{X}$?

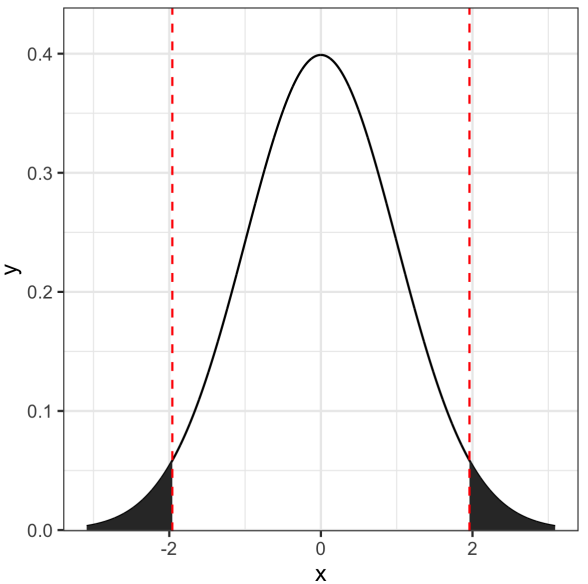
$\bar{X} \sim N(\mu, \sigma^2/n)$. So what is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$?

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

Now, we can find values x_1, x_2 such that $P(Z \leq x_1) + P(Z \geq x_2) = \alpha$. Let's for simplicity use $\alpha = 0.05$. I.e. we want to find x_1, x_2 such that this area is 0.05.

If we decide the two areas in the tails are the same, $x_1 = -x_2$.

x_2 is by definition the $\alpha/2$ (0.025 in this case) critical value, $z_{\alpha/2}$ - it cuts off $\alpha/2$ (0.025) to the right!



Estimation: Confidence Intervals

So,

$$
\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X}-\mu \leq z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(-\bar{X}-z_{\alpha/2}\sigma/\sqrt{n} \leq -\mu \leq -\bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(\bar{X}+z_{\alpha/2}\sigma/\sqrt{n} \geq \mu \geq \bar{X}-z_{\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right)
\end{aligned}
$$

Estimation: Confidence Intervals

The interval $[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ is called a $(1-\alpha)\cdot 100\%$ Confidence Interval:

We are $(1-\alpha)\cdot 100\%$ confident that an interval constructed this way will contain the true value μ!

Estimation: Confidence Intervals

Example

Through years, and years of experience, the car manufacturer has learned that the true standard deviation of the paint thickness is 0.34. We can now construct a 95% confidence interval for the true mean.

First, we find $z_{\alpha/2}$. When looking for a 95% CI, $\alpha = 0.05$. $z_{0.025}$ cuts off 0.025 on the right hand side. I.e. it cuts off $1 - 0.025 = 0.975$ to the left. So $P(X \leq z_{0.025}) = 0.975$, which makes $z_{0.025}$ the 0.975 quantile:

```
Z <- Normal()

z_crit <- quantile(Z, 0.975)
z_crit
```

```
## [1] 1.959964
```

Estimation: Confidence Intervals

We can find the confidence interval as

```
paint_thickness %>%
  summarize(mean = mean(thickness),
            LL = mean - z_crit*0.34/sqrt(n()),
            UL = mean + z_crit*0.34/sqrt(n()))
```

```
##       mean       LL       UL
## 1 1.348125 1.181528 1.514722
```

or

```
xbar <- mean(paint_thickness$thickness)
n <- nrow(paint_thickness)
```

```
xbar - z_crit*0.34/sqrt(n)
```

```
## [1] 1.181528
```

```
xbar + z_crit*0.34/sqrt(n)
```

```
## [1] 1.514722
```

Estimation: Confidence Intervals

We are 95% confident that the true mean thickness is between 1.18 and 1.51.

What is $P(1.18 \leq \mu \leq 1.51)$? 0 or 1. We do not know which, but those are the only possible values.

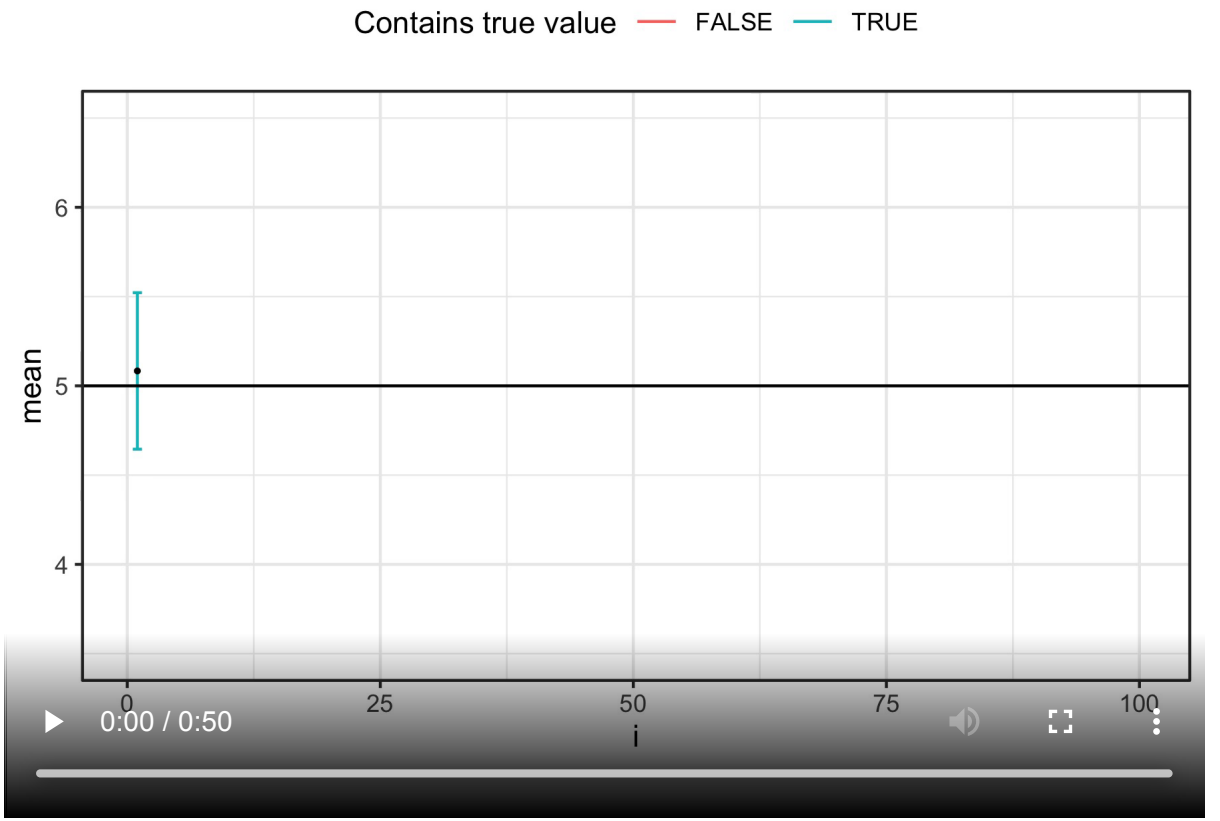
We can think of LL and UL as random variables: new sample -> new CI. So makes sense to say $P(LL \leq \mu \leq UL) = 1 - \alpha$.

But as soon as we get a sample, observe values, and find the realization of the RVs, no more randomness. Hence, not meaningful to talk about probabilities anymore.

What does it mean that we are "95% confident the true value is in the interval?"

Estimation: Confidence Intervals

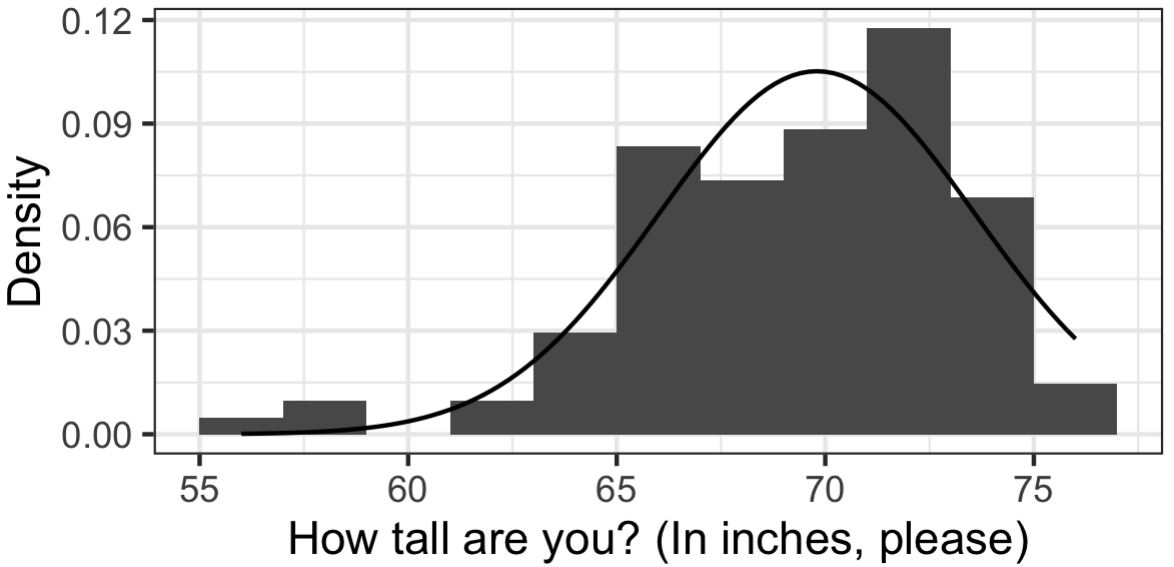
If we repeat the process, the interval we get will contain the true value 95% of the time.



Estimation: Confidence Intervals

Example: student heights

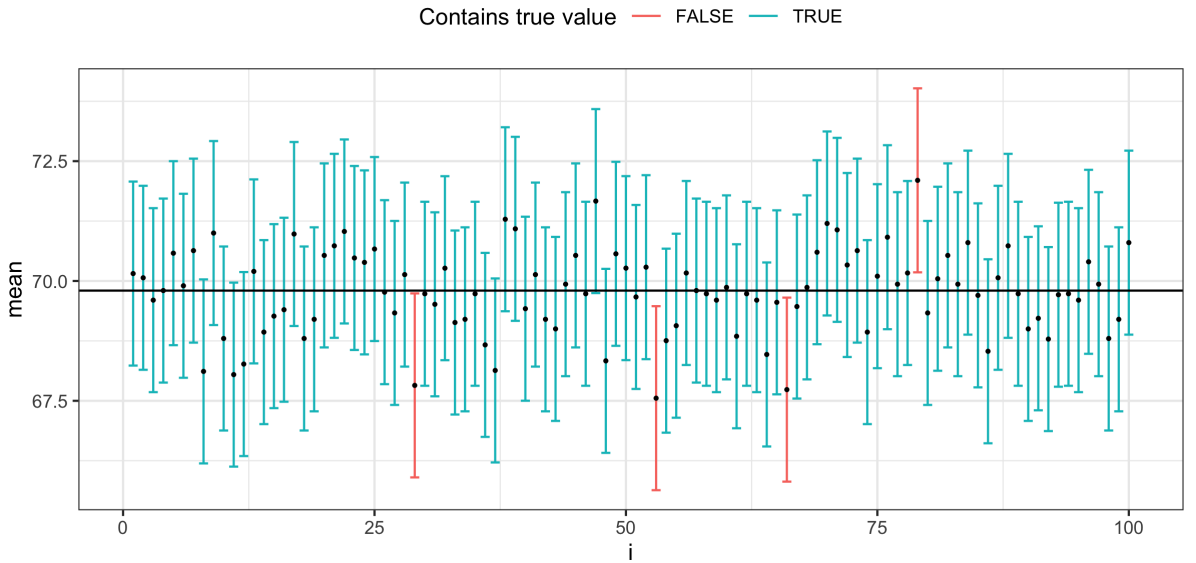
The true population:



Estimation: Confidence Intervals

Example: student heights

Take a sample of 15 students, calculate 95% CI, repeat 100 times.



Estimation: Confidence Intervals

Example: Framingham Heart Study

Total cholesterol in adults.

```
fram <- read_csv(here::here("csv_data/framingham.csv")) %>%
  filter(!is.na(totChol))
```

Remember, so far we need the data to be normal.

Histogram:

```
ggplot(fram,
       aes(x = totChol)) +
  geom_histogram(bins = 35)
```

QQ-plot:

```
ggplot(fram, aes(sample = totCho
  geom_qq() +
  geom_abline(aes(slope = sd(tot
```

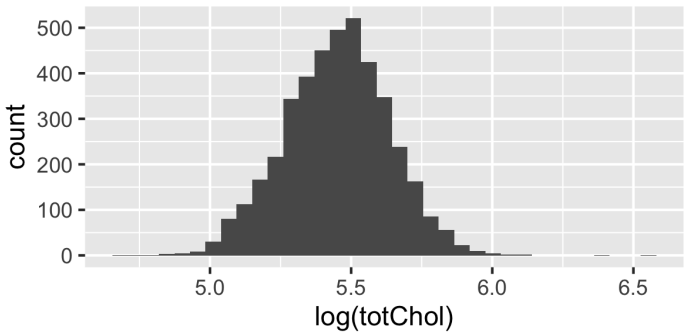
Estimation: Confidence Intervals

Example: Framingham Heart Study

Not convincing, but if we log transform:

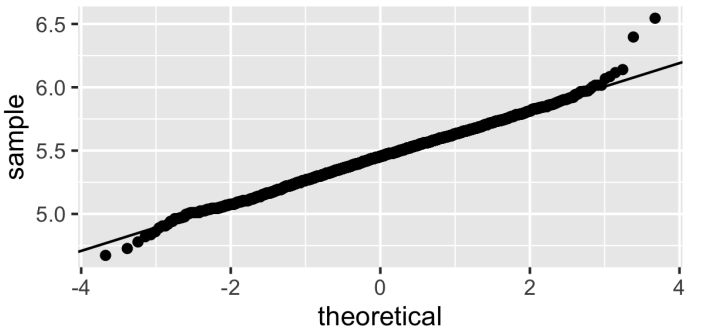
Histogram:

```
ggplot(fram,
       aes(x = log(totChol))) +
  geom_histogram(bins = 35)
```



QQ-plot:

```
ggplot(fram, aes(sample = log(to
  geom_qq() +
  geom_abline(aes(slope = sd(log
```



Estimation: Confidence Intervals

Example: Framingham Heart Study

So we can construct a 95% confidence interval for the log(totChol)... only we do not know the true value of σ?! Let's assume for know that we do know it, and it is 0.19.

```
mean(log(fram$totChol))
```

```
## [1] 5.449583
```

We are 95% confident that the true *population mean* of the log total cholesterol is in this interval.

Lower limit:

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 5.45 - 1.96\frac{0.19}{\sqrt{4190}}$$
$$= 5.444$$

Upper limit:

Estimation: Confidence Intervals

All of this build on some key assumptions:

1. $\bar{X}$ normally distributed
2. Know the true standard deviation.

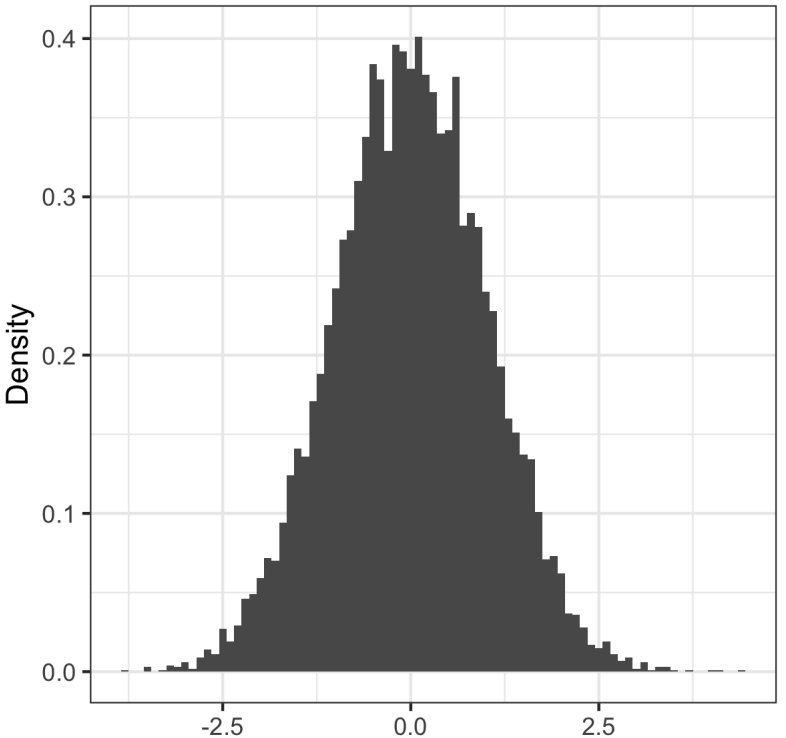
When both satisfied, $\frac{\bar{X}-E(\bar{X})}{\text{SD}(\bar{X})} \sim N(0,1)$.

We said "if $X_1, \ldots, X_n$ are normal, then $\bar{X}$ normal". Let's check. Actually, let's check that if $X_1, \ldots, X_n$ are normal, then $\frac{\bar{X}-\mu}{\text{SD}(\bar{X})} \sim N(0,1)$.

How would we go about that? Get many, many samples, calculate the mean for each, and plot a histogram. We'll use $\mu = 5$, $\sigma = 10$, and $n = 5$.

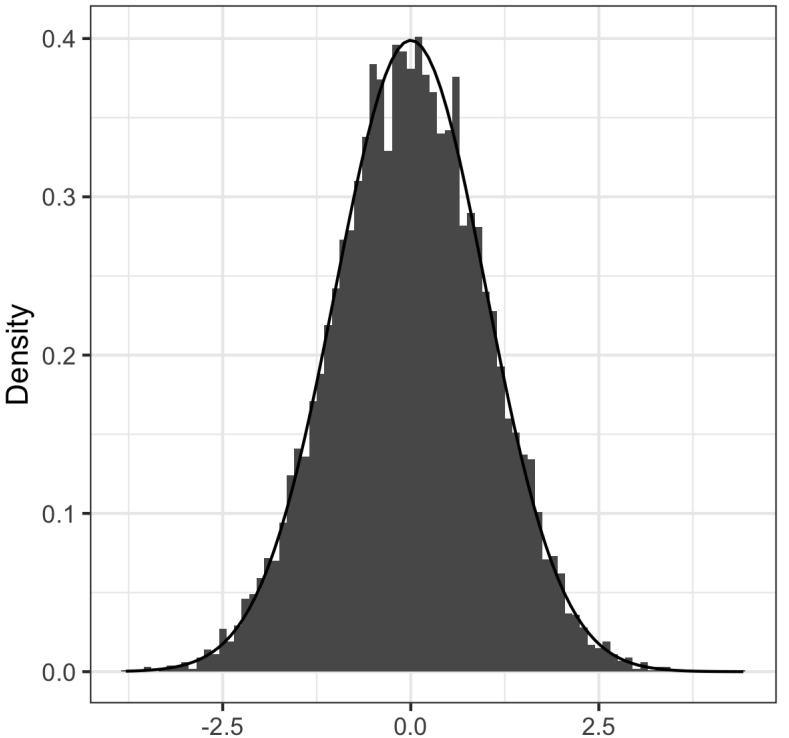
Estimation: Confidence Intervals

Histogram of $\frac{\bar{x}-5}{10/\sqrt{5}}$ for 10000 samples:



Estimation: Confidence Intervals

Histogram of $\frac{\bar{x}-5}{10/\sqrt{5}}$ for 10000 samples:



Estimation: Confidence Intervals

What if we do not know the true value of σ? We would use $\hat{\sigma} = S$... but the distribution of $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ is NOT $N(0,1)$:

Estimation: Confidence Intervals

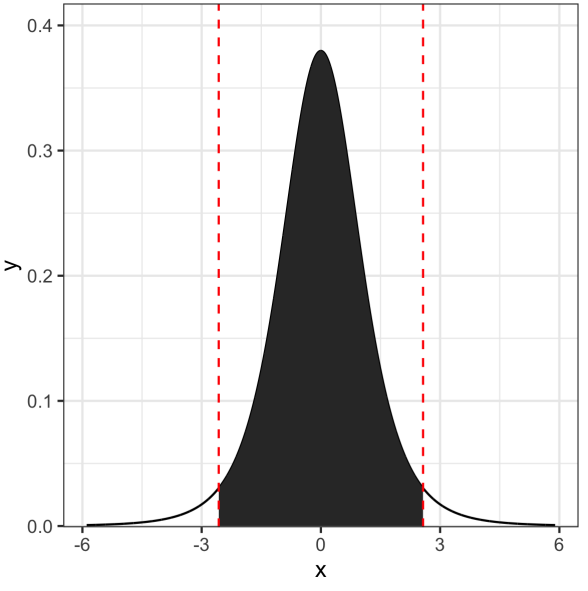
The distribution of $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}}$ is a so-called t-distribution with $n-1$ degrees of freedom:

Estimation: Confidence Intervals

Does this mean all is lost? No, of course not! We just need to adjust a bit. We now have

$$\begin{aligned}\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} &= \frac{\bar{X}-E(\bar{X})}{\widehat{\mathrm{SD}}(\bar{X})} \\ &= T_{n-1} \sim\ t_{n-1}\end{aligned}$$

Now, we can find values x_1, x_2 such that $P\left(x_1 \leq T_{n-1} \leq x_2\right) = 1-\alpha$. Let's for simplicity use $\alpha = 0.05$. I.e. we want to find x_1, x_2 such that this area is 0.95.



Estimation: Confidence Intervals

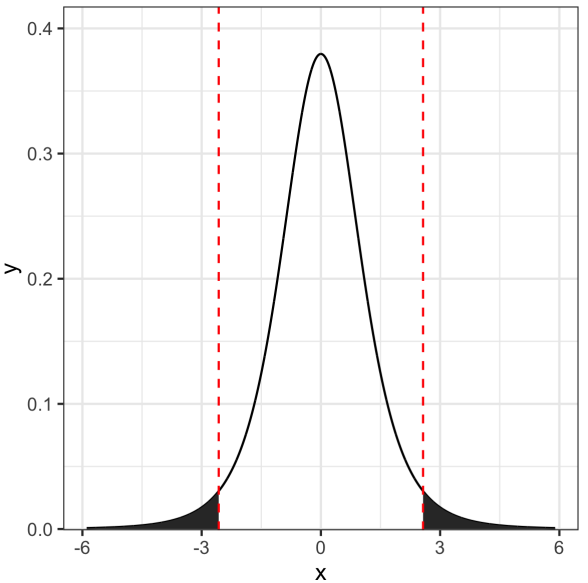
Does this mean all is lost? No, of course not! We just need to adjust a bit. We now have

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{X} - E(\bar{X})}{\widehat{\mathrm{SD}}(\bar{X})}$$
$$= T_{n-1} \sim t_{n-1}$$

Now, we can find values x_1, x_2 such that
$P(T_{n-1} \leq x_1) + P(T_{n-1} \geq x_2) = \alpha$
. Let's for simplicity use $\alpha = 0.05$.
I.e. we want to find x_1, x_2 such that this area is 0.05.

If we decide the two areas in the tails are the same, $x_1 = -x_2$.

x_2 is by definition the $\alpha/2$ (0.025 in this case) critical value *in the t-*



Estimation: Confidence Intervals

So,

$$
\begin{aligned}
1-\alpha &= P(-t_{n-1,\alpha/2} \leq T_{n-1} \leq t_{n-1,\alpha/2}) \\
&= P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) \\
&= P\left(-t_{n-1,\alpha/2}\sigma/\sqrt{n} \leq \bar{X}-\mu \leq t_{n-1,\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(-\bar{X}-t_{n-1,\alpha/2}\sigma/\sqrt{n} \leq -\mu \leq -\bar{X}+t_{n-1,\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(\bar{X}+t_{n-1,\alpha/2}\sigma/\sqrt{n} \geq \mu \geq \bar{X}-t_{n-1,\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(\bar{X}-t_{n-1,\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X}+t_{n-1,\alpha/2}\sigma/\sqrt{n}\right)
\end{aligned}
$$

Estimation: Confidence Intervals

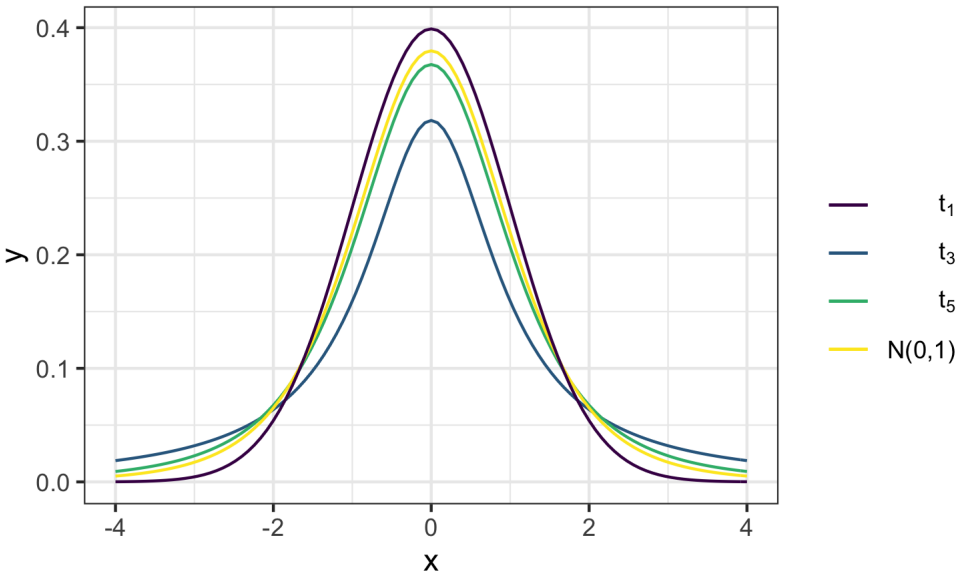
When $\bar{X}$ is normal, but the true value of σ is unknown, the interval $[\bar{X} - t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}]$ is called a $(1-\alpha)\cdot 100\%$ Confidence Interval:

We are $(1-\alpha)\cdot 100\%$ confident that an interval constructed this way will contain the true value μ!

New Distribution: t-distribution

The t-distribution is very similar to the standard normal.

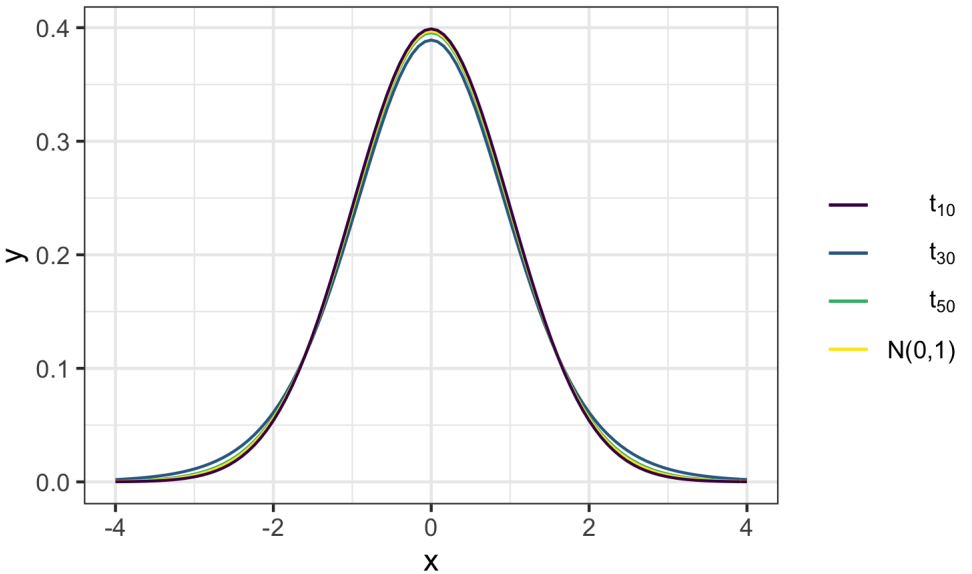
It is defined by a single parameter called the *degrees of freedom* (denoted df). We will use T_{df} as notation for a random variable that follows the t-distribution with df degrees of freedom, i.e. $T_{df} \sim t_{df}$.



New Distribution: t-distribution

The t-distribution is very similar to the standard normal.

It is defined by a single parameter called the *degrees of freedom* (denoted df). We will use T_{df} as notation for a random variable that follows the t-distribution with df degrees of freedom, i.e. $T_{df} \sim t_{df}$.



Actually, if " $\mathrm{df} = \infty$ ", the t-distribution is *exactly* the standard normal.