MEDB 5501, Module08

2024-10-09

Topics to be covered

- What you will learn
  - The two-sample t-test
  - The t-distribution
  - Critical values and p-values
  - R code for the t-test
  - Confidence intervals
  - Sample size justification
  - Your homework

Population model

- Population 1
  - $X_{11}, X_{12}, \ldots, X_{1N_1}$
  - X_{1i} are independent $N(\mu_1, \sigma_1)$
- Population 2
  - $X_{21}, X_{22}, \ldots, X_{2N_2}$
  - X_{2i} are independent $N(\mu_2, \sigma_2)$

Speaker notes

The two sample t-test is based on a population model where there are N_1 observations in the first population and N_2 observations in the second population. In general, the size of the two populations, N_1 and N_2 are assumed to be very large.

Sample values

- Sample 1
  - $X_{11}, X_{12}, \ldots, X_{1n_1}$
  - Calculate $\bar{X}_1$ and S_1
- Sample 2
  - $X_{21}, X_{22}, \ldots, X_{2n_2}$
  - Calculate $\bar{X}_2$ and S_2

Speaker notes

Because the populations are so large, you need to take a sample (hopefully a representative sample) from each population. With the sample, you can calculate sample statistics.

Hypothesis and test statistic

- $H_0: \ \mu_1 - \mu_2 = 0$
- $H_1: \ \mu_1 - \mu_2 \neq 0$
  - Accept H_0 if $\bar{X}_1 - \bar{X}_2$ is close to zero

Speaker notes

The null hypothesis for the two-sample t-test is that the population means, μ_1 and μ_2 are equal, which is the same as saying that the difference between the two population means is equal to zero.

The population means are unknown, but you can use the sample means,

How close is close?

- standard error
  - $se = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  - Pooled standard deviation
    - $S_p = \sqrt{\frac{n_1 \ S_1^2 + n_2 \ S_2^2}{n_1 + n_2}}$
- Not valid with heterogeneity
  - That is, $\sigma_1 \neq \sigma_2$

Speaker notes

You measure how close $\bar{X}_1 - \bar{X}_2$ is to zero by using the standard error. The standard error is a measure of how much sampling error you have when using $\bar{X}_1 - \bar{X}_2$ to estimate $\mu_1 - \mu_2$.

This standard error relies on equal variation in both groups. You'll hear more discussion of this issue later in the presentation.

Break #1

- What you have learned
  - The two-sample t-test
- What's coming next
  - The t-distribution

The t distribution

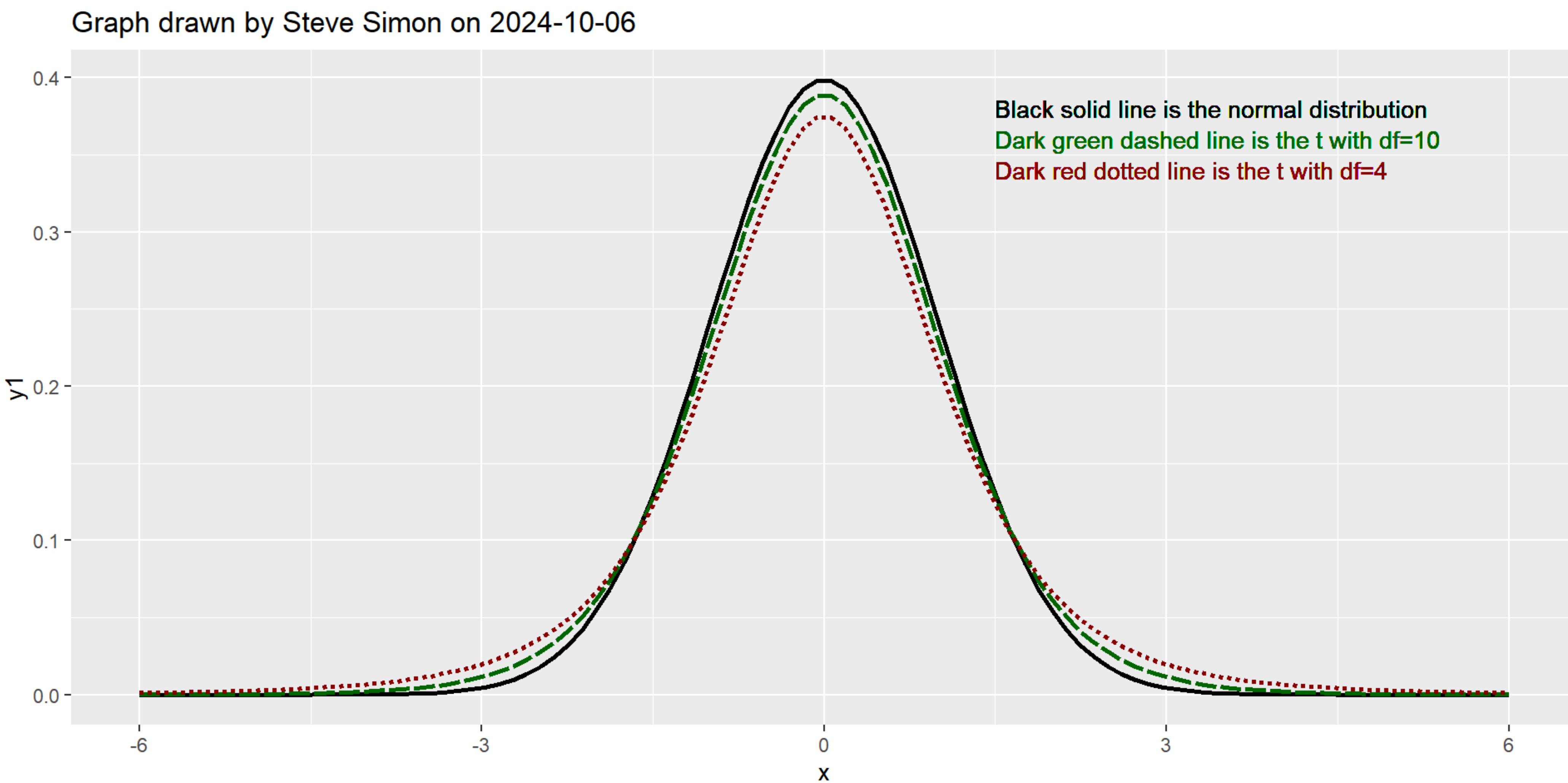
- $T = \frac{\bar{X}_1 - \bar{X}_2}{se}$
  - Variation in the numerator AND the denominator
  - Use a t-distribution, not a normal distribution
    - $df = n_1 + n_2 - 2$

Speaker notes

The test statistic, T, is the ratio of the difference in sample means to the standard error. This statistic has variation both in the numerator and the denominator. This produces a statistic that is not normally distributed, but close to normal. It is the t-distribution.

The t-distribution has degrees of freedom. A large degrees of freedom means very little sampling error in the denominator. It is the total sample size $(n_1 + n_2)$ minus two degrees of freedom associated with the two estimated means used in the standard deviation calculation.

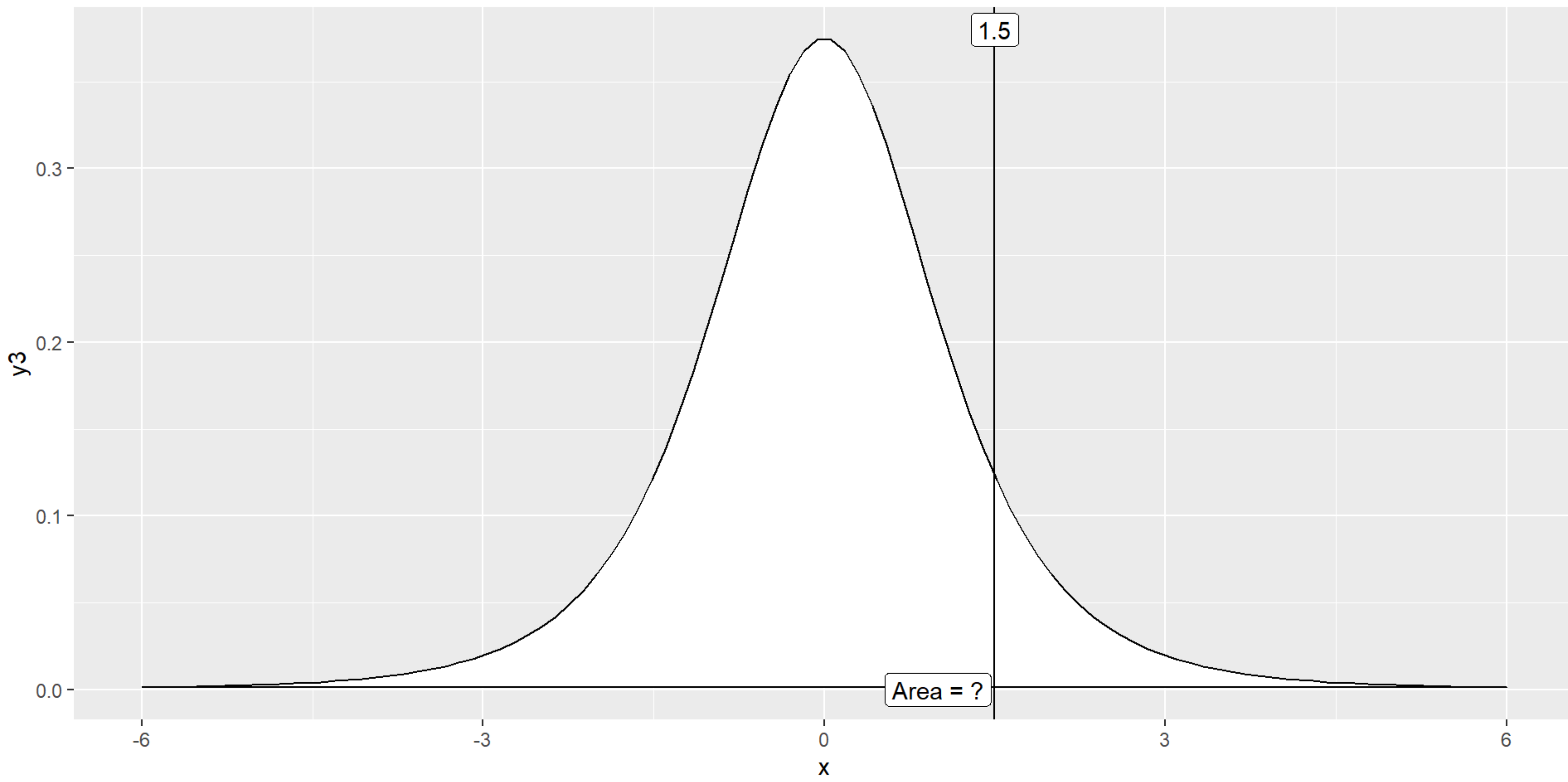
Comparing the t and normal distributions



Speaker notes

This graph compares the normal distribution to a t distribution with 10 degrees of freedom and a t distribution with 4 degrees of freedom. Both the normal and the t-distributions are symmetric. The t-distributions have a little bit less probability near zero and a bit more probability at the extremes.

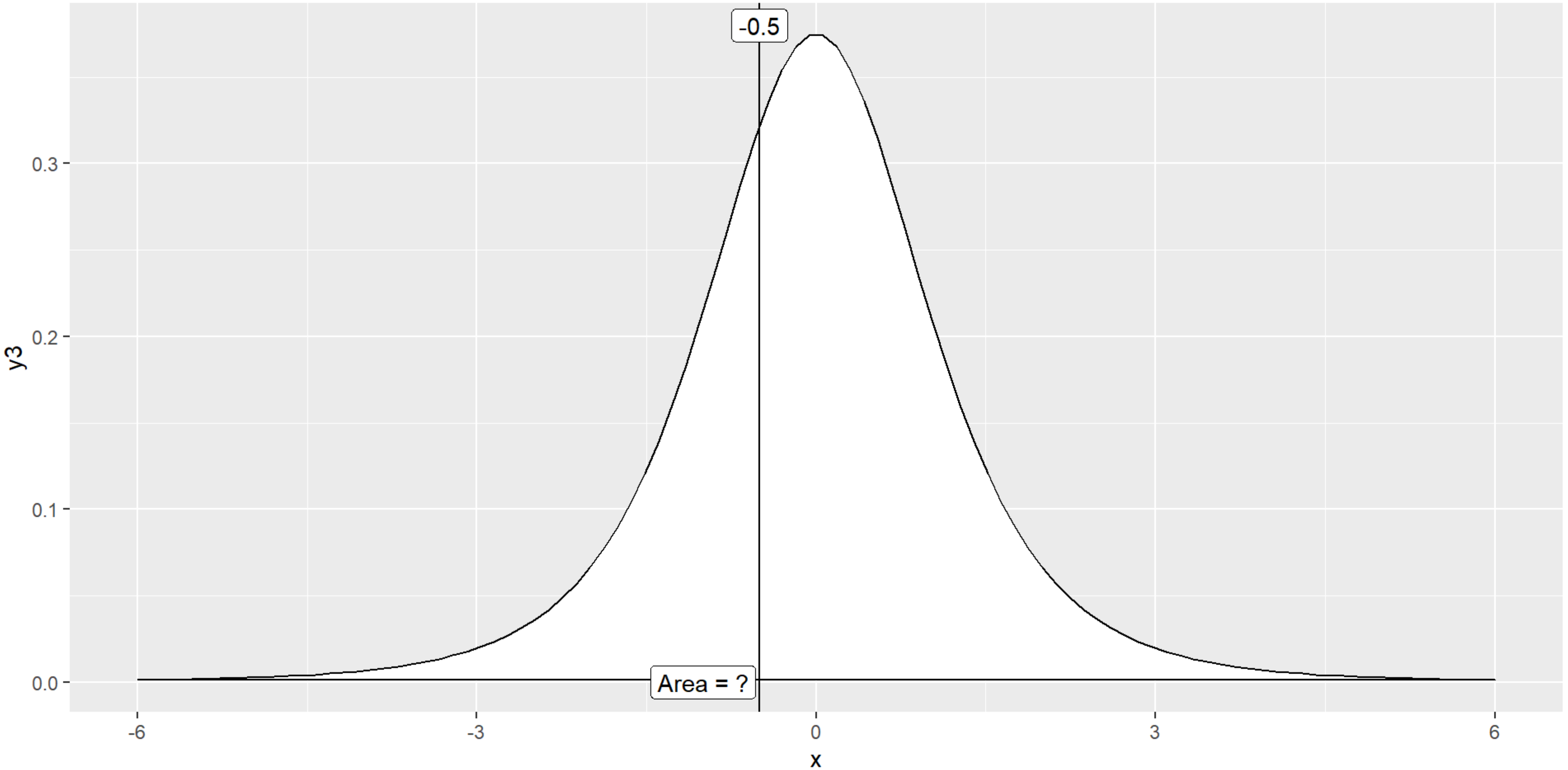
P[t(4) < 1.5]




```
pt(1.5, 4)
```

```
[1] 0.896
```

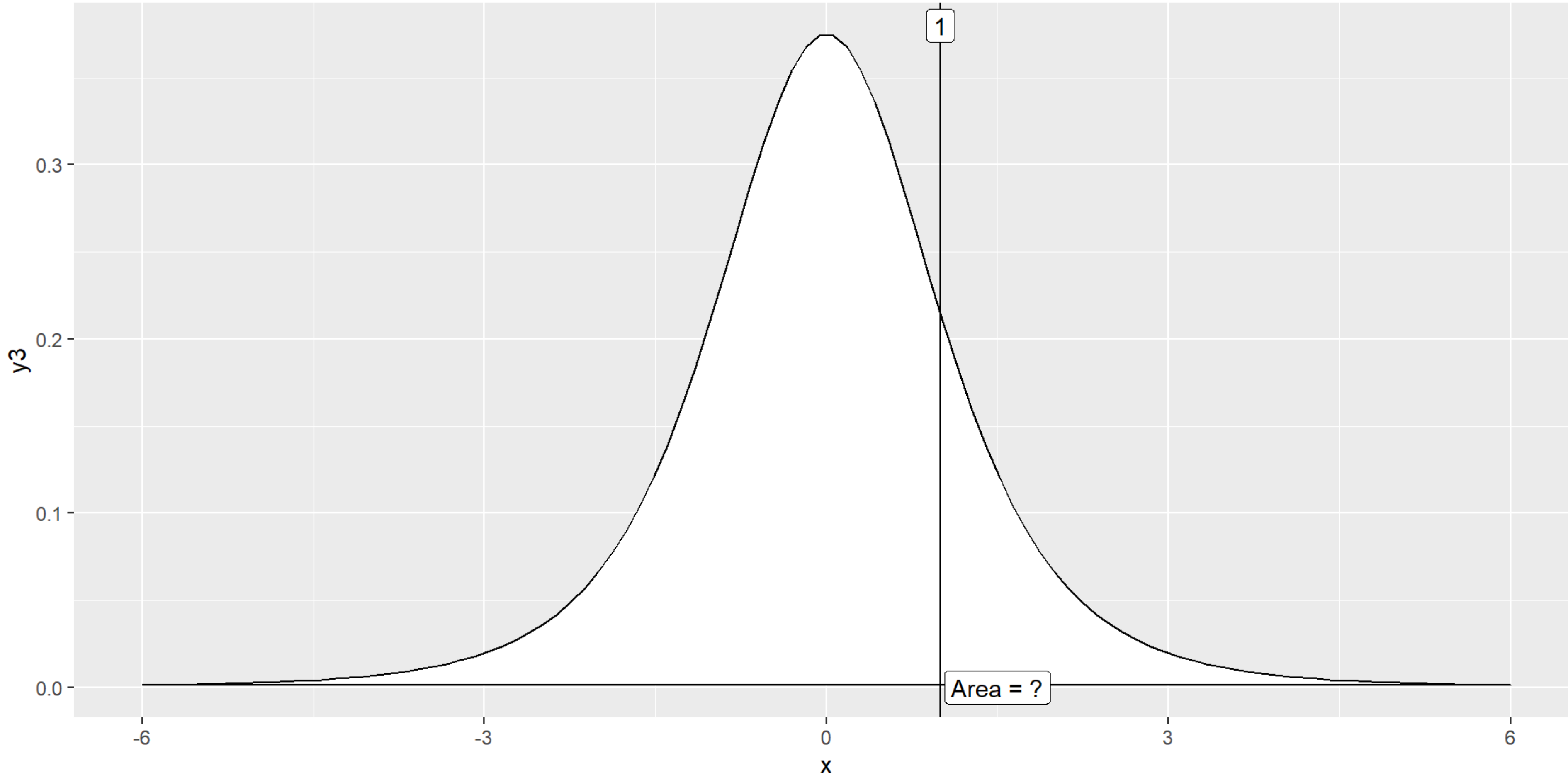
P[t(4) < -0.5]




```r
pt(-0.5, 4)
```

```
[1] 0.321665
```

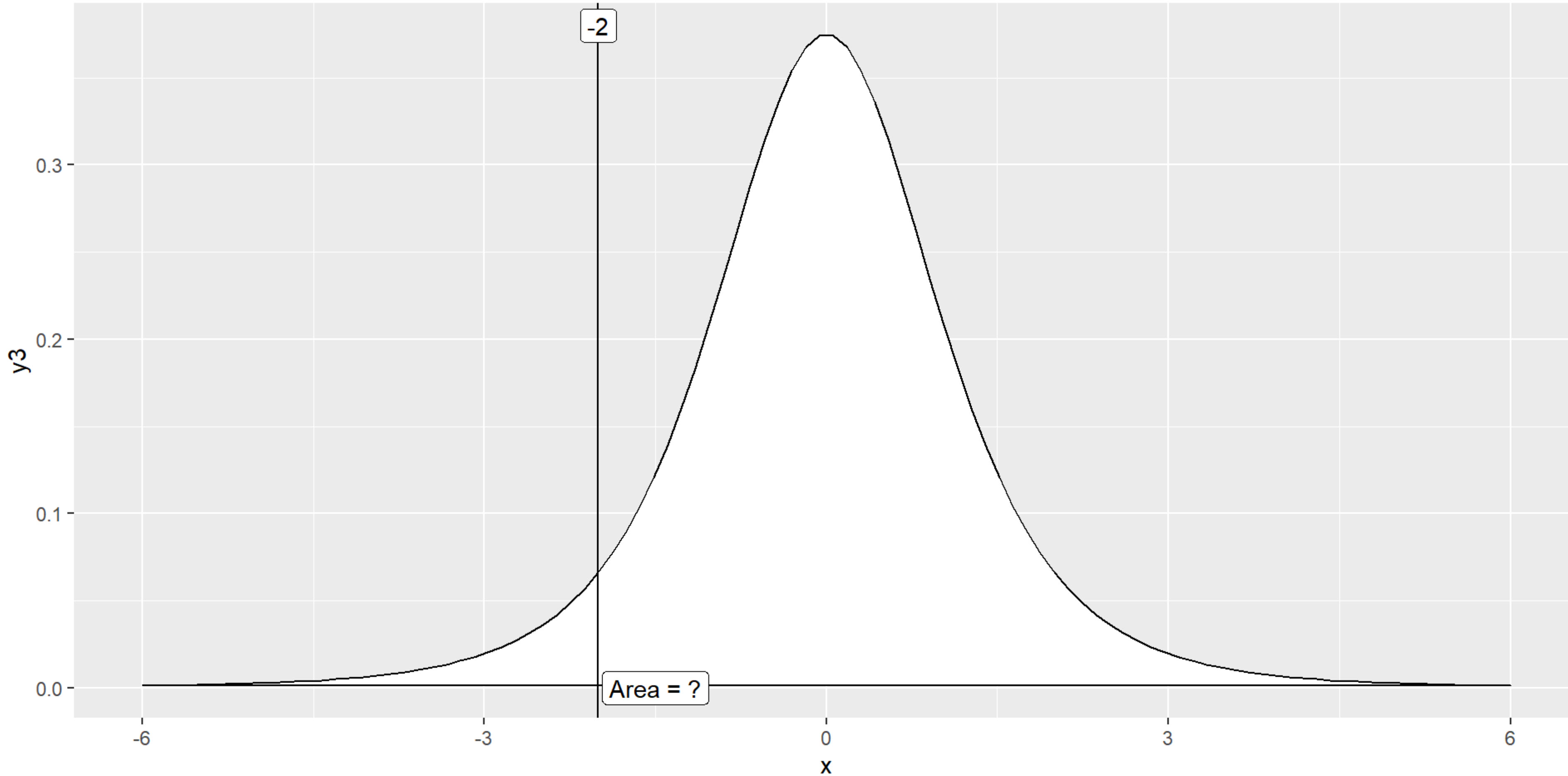
P[t(4) > 1]




```r
1 - pt(1, 4)
```

```
[1] 0.1869505
```

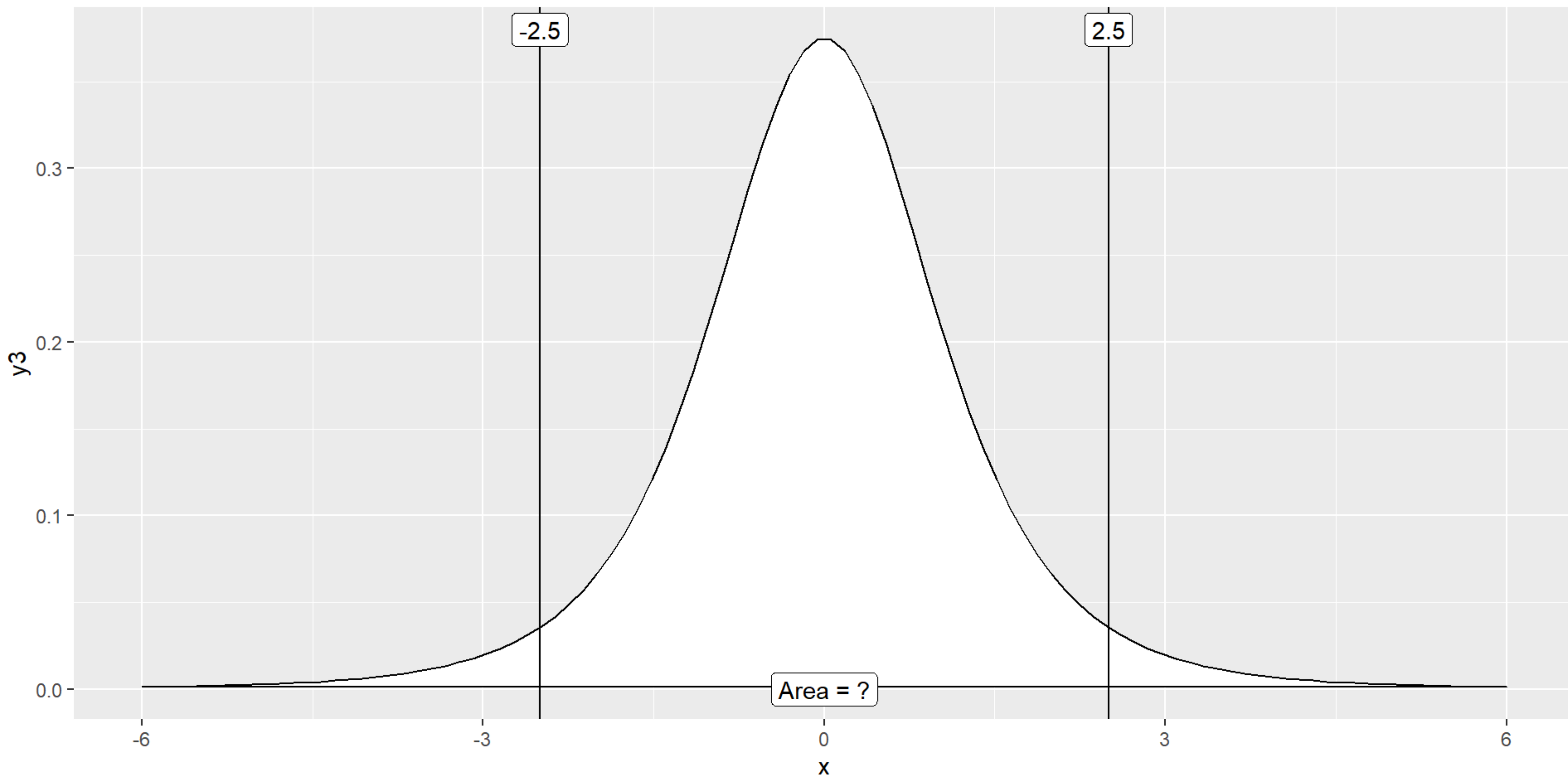
P[t(4) > -2]




```
1 - pt(-2, 4)
```

```
[1] 0.9419417
```

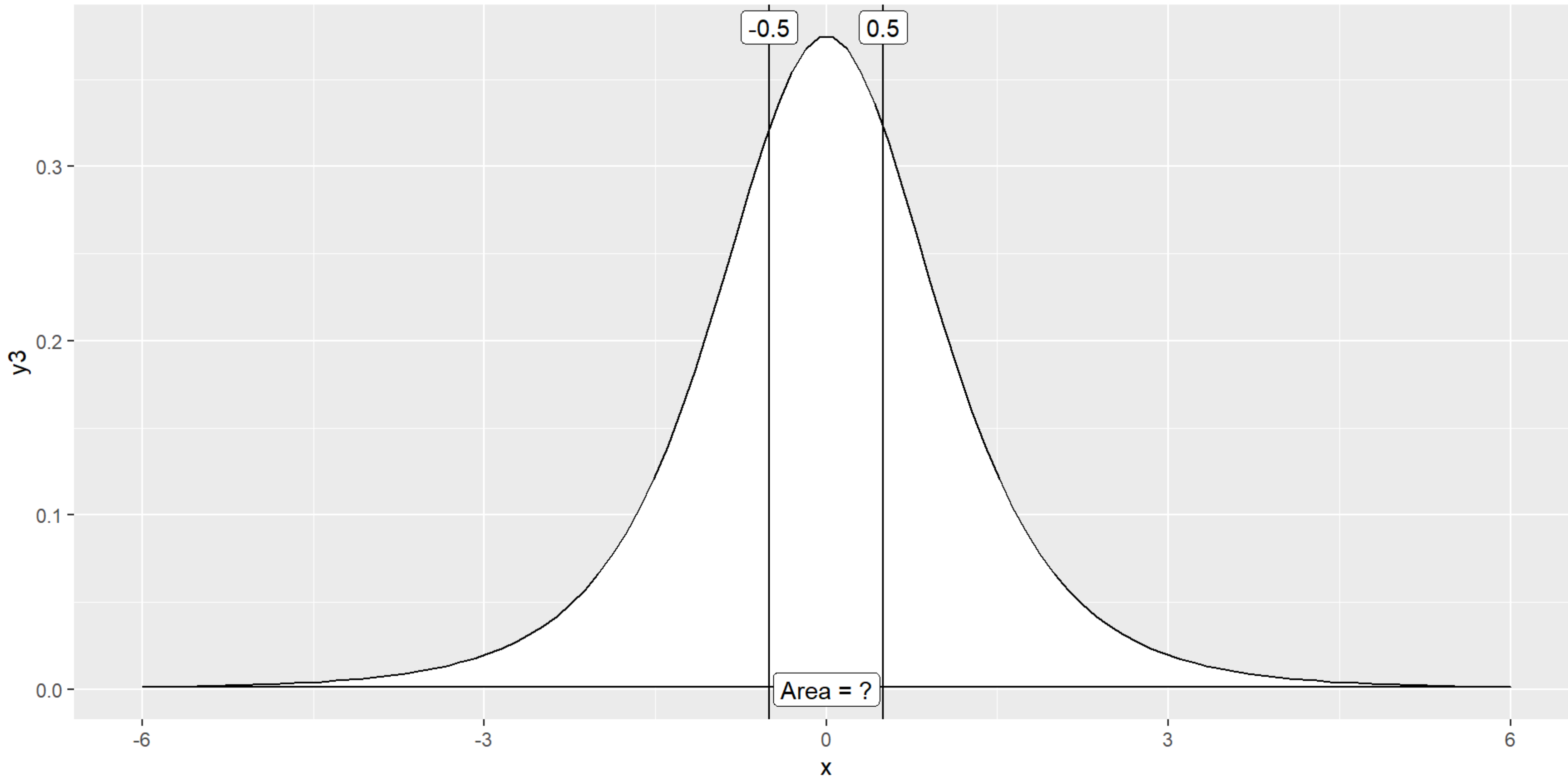
P[-2.5 < t(4) < 2.5]




```
pt(2.5, 4) - pt(-2.5, 4)
```

```
[1] 0.9332335
```

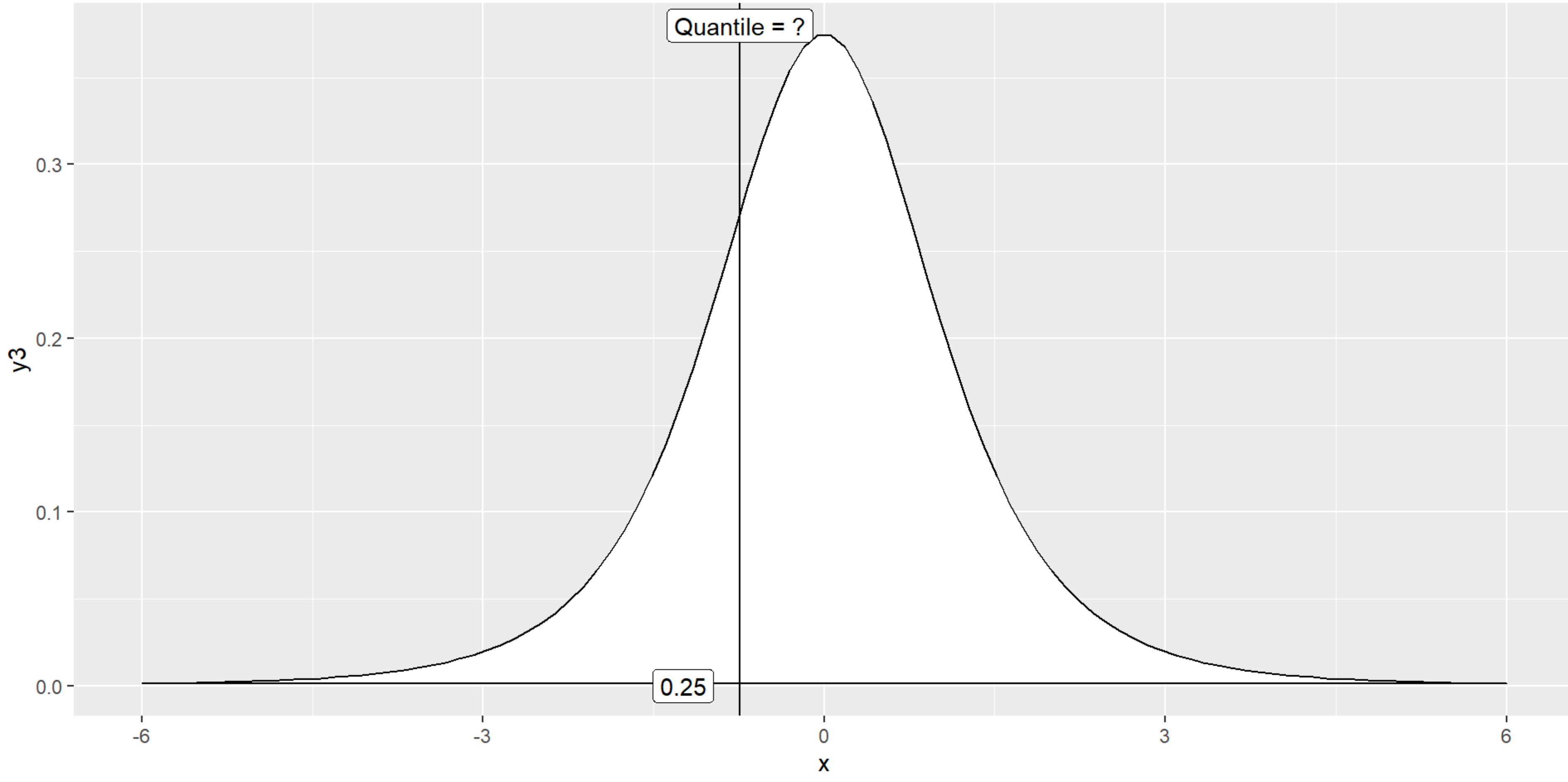
P[-0.5 < t(4) < 0.5]




```r
pt(0.5, 4) - pt(-0.5, 4)
```

```
[1] 0.35667
```

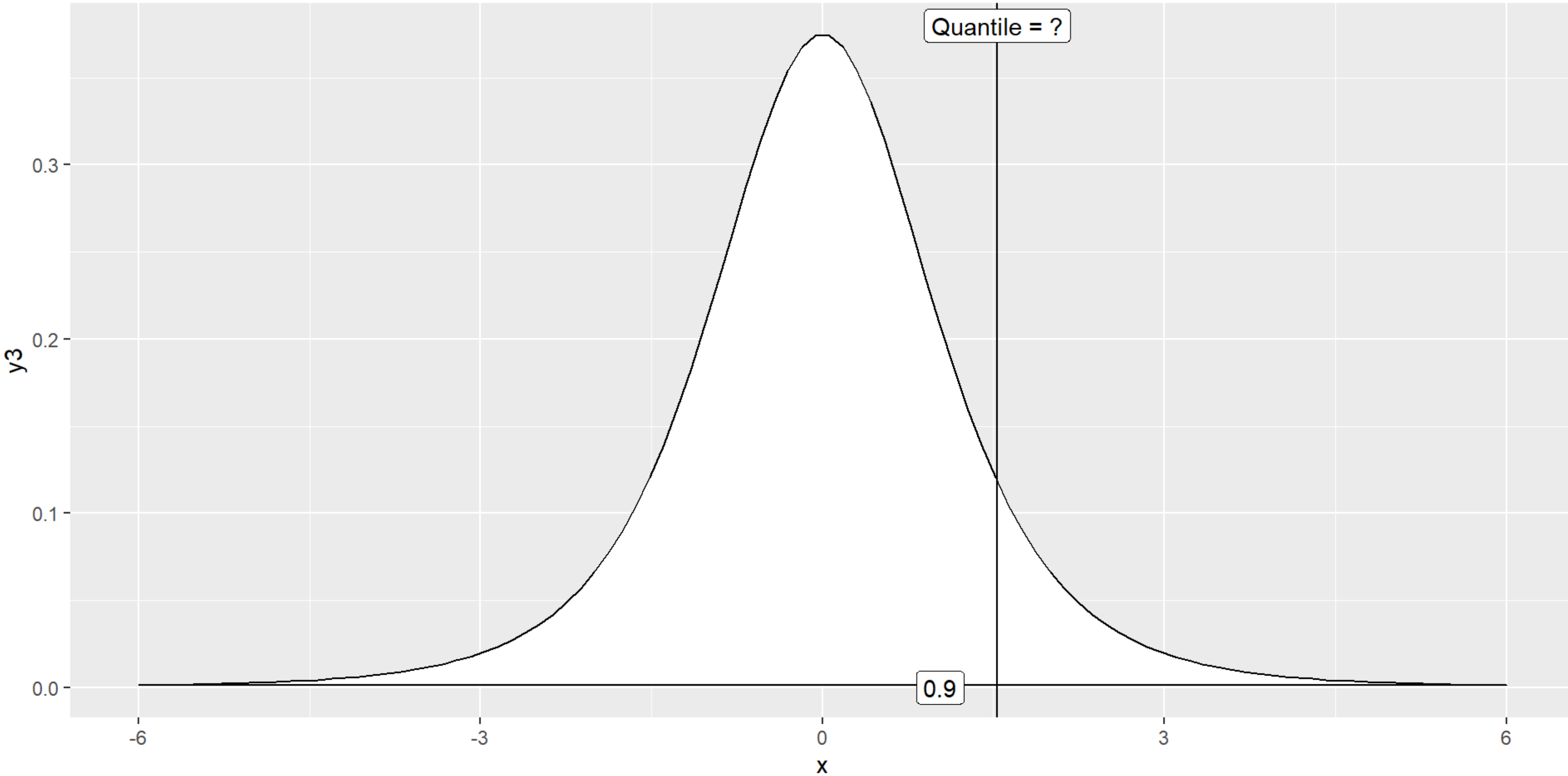
25th percentile of t(4)




```
qt(0.25, 4)
```

```
[1] -0.7406971
```

90th percentile of t(4)




```r
qt(0.9, 4)
```

```
[1] 1.533206
```

Break #2

- What you have learned
  - The t-distribution
- What's coming next
  - Critical values and p-values

Type I and Type II errors

- Type I error, Rejecting the null hypothesis when the null hypothesis is true
  - α is the probability of a Type I error
- Type II error, Accepting the null hypothesis when the null hypothesis is false
  - β is the probability of a Type II error
  - Power = 1 - β

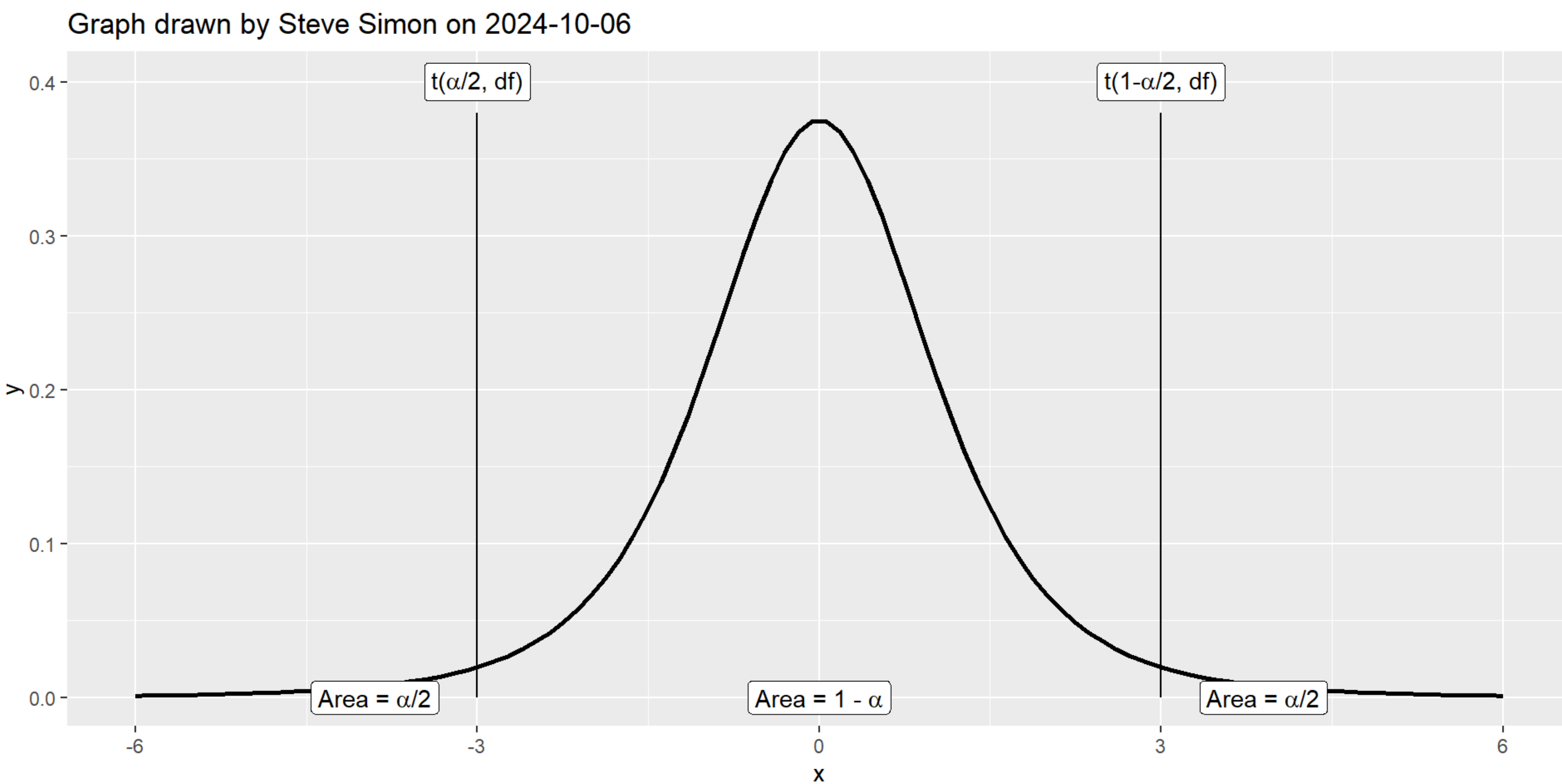
Critical values, 1

- $H_0: \ \mu_1 - \mu_2 = 0$
- $H_1: \ \mu_1 - \mu_2 \neq 0$
  - $T = \frac{\bar{X}_1 - \bar{X}_2}{se}$
  - Accept H_0 if $t(\alpha/2, df) < T < t(1 - \alpha/2, df)$

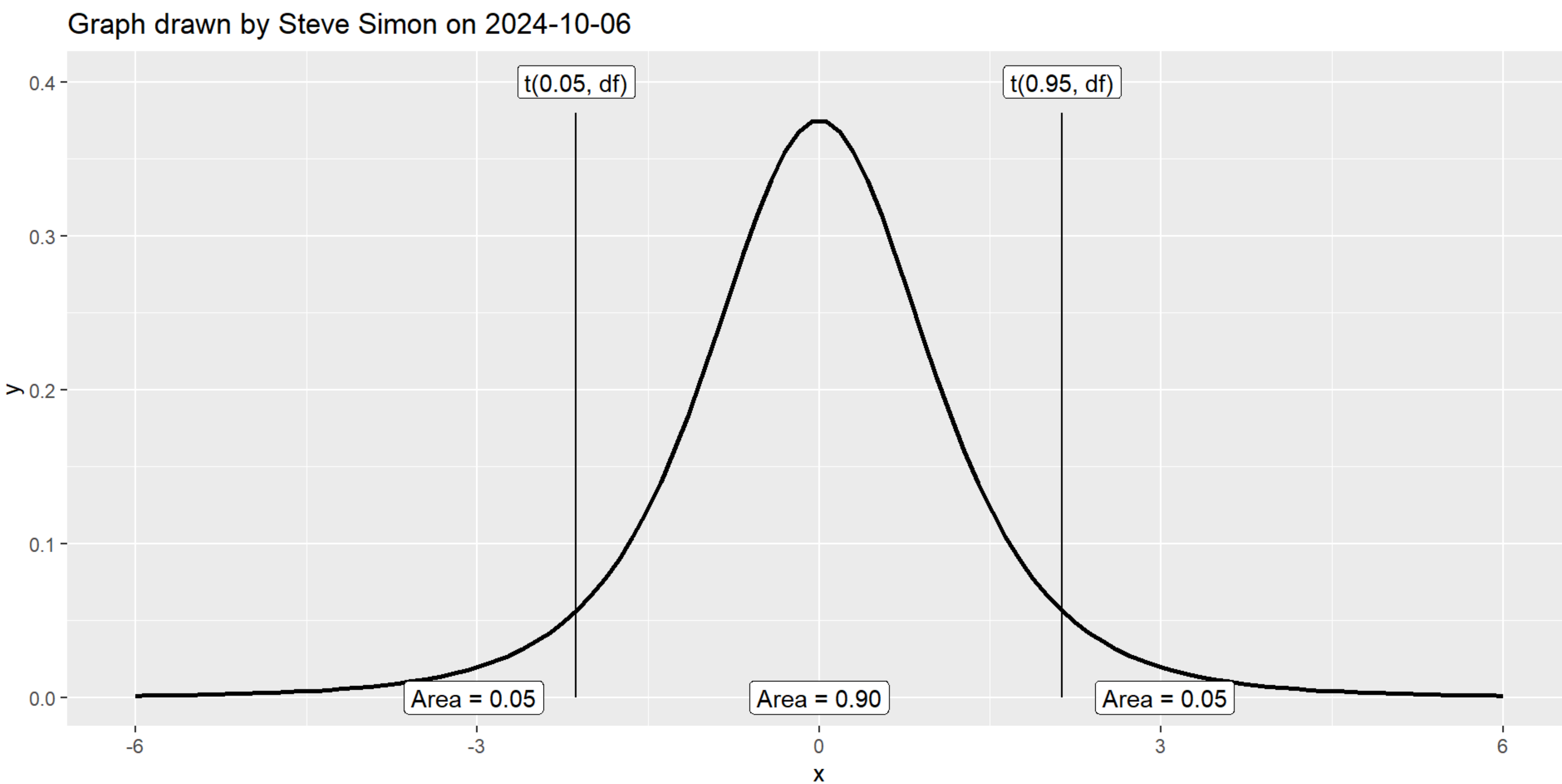
Speaker notes

The formal test of hypothesis looks at whether the test statistic T is close to zero. Close means that it falls between the $\alpha/2$ and $1 - \alpha/2$ percentiles of a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

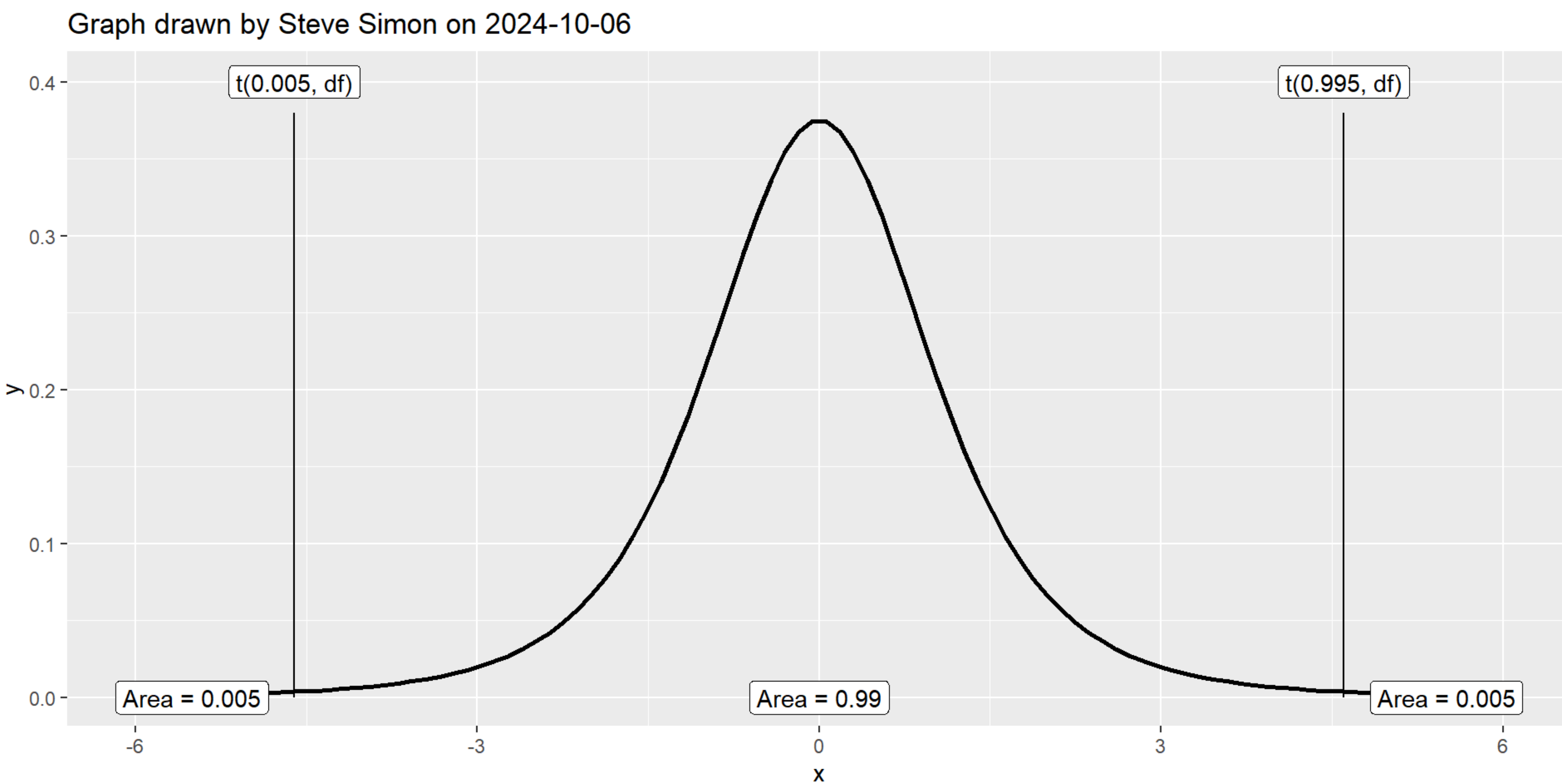
Critical values, 2



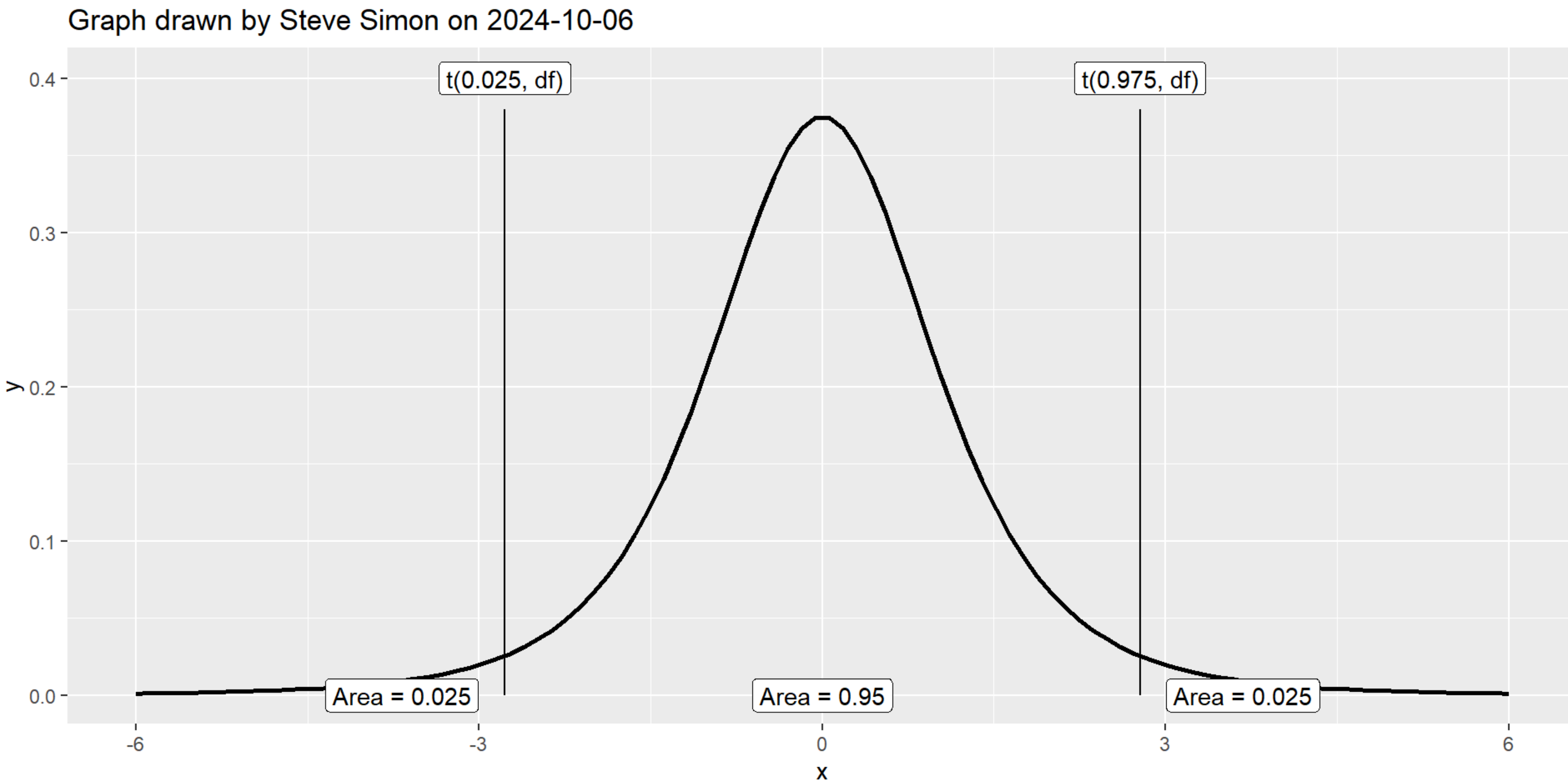
Critical values, 3



Critical values, 4



Critical values, 5



P-value

- p-value = $2P[t(n_1 + n_2 - 2) > |T|]$
  - Probability of sample results or results more extreme
  - Accept H_0 if p-value > α
- Why the 2?
  - Measuring extremity in either direction

Postural sway data, 1

```
# A tibble: 2 × 4
  age      fb_mn fb_sd     n
  <chr>    <dbl> <dbl> <int>
1 Elderly   26.3  9.77     9
2 Young     18.1  4.09     8
```

Postural sway data, 2

- Calculate pooled standard deviation
  - $S_p = \sqrt{\dfrac{n_1\ S_1^2 + n_2\ S_2^2}{n_1 + n_2}}$
  - $S_p = \sqrt{\dfrac{9\ (9.77)^2 + 8\ (4.09)^2}{9 + 8}}$
  - S_p = 7.64

Postural sway data, 3

- Calculate standard error
  - se = $S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  - se = $7.64\sqrt{\frac{1}{9} + \frac{1}{8}}$
  - se = 3.71

Postural sway data, 4

- Calculate T
  - T = $\frac{\bar{X}_1 - \bar{X}_2}{se}$
  - T = $\frac{26.3 - 18.1}{3.72}$
  - T = 2.2

Postural sway data, 5

- Calculate critical values
  - t(0.025, 15) = qt(0.025, 15) = -2.13
  - t(0.975, 15) = qt(0.975, 15) = 2.13
  - Since T is outside the two critical values, reject H_0

Postural sway data, 6

- Calculate p-value
  - p.value = $P[t(n_1 + n_2 - 2) > |T|]$
  - p.value = $2\ P[t(15) > |2.2|]$ = 2*(1-pt(2.2, 15)) = 0.044

Postural sway data, 7

```
    Two Sample t-test

data:  fbsway by age
t = 2.2044, df = 15, p-value = 0.04353
alternative hypothesis: true difference in means between group Elderly and
group Young is not equal to 0
95 percent confidence interval:
  0.2715453 16.1451214
sample estimates:
mean in group Elderly   mean in group Young
             26.33333              18.12500
```

Break #3

- What you have learned
  - Critical values and p-values
- What's coming next
  - R code for the t-test

postural-sway data dictionary, 1

```
data_dictionary: postural-sway.txt

copyright: |
  The author of the jse article holds the copyright, but does not list
conditions under which it can be used. Individual use for educational purposes
is probably permitted under the Fair Use provisions of U.S. Copyright laws.

description: |
  Postural sway is a measure of how well patients can balance. The postural
sway was measured using a force plate in two groups of subjects, elderly or
young. Sway was measured in the forward/back direction and in the side-to-side
direction.

additional_description: https://gksmyth.github.io/ozdasl/general/balaconc.html
```

postural-sway data dictionary, 2

```
download_url: https://gksmyth.github.io/ozdasl/general/balaconc.txt

format: tab delimited

varnames: Included in the first line

missing_value_code: not needed

size:
  rows: 17
  columns: 3
```

postural-sway data dictionary, 3

```
Age:
  scale: nominal
  values:
  - Elderly
  - Young
FBSway:
  label: Sway in the forward-backward direction
  scale: ratio
  range: positive real
SideSway:
  label: Sway in the side-to-side direction
  scale: ratio
  range: positive real
```

simon-5501-08-sway.qmd, 1

```
---
title: "Analysis of postural sway data"
author: "Steve Simon"
format:
  html:
    embed-resources: true
date: 2024-10-07
---

This program reads data and runs a two-sample t-test. Consult the [data
dictionary][dic] for information about the data itself.

[dic]: https://github.com/pmean/data/blob/main/files/postural-sway.yaml
```

simon-5501-08-sway.qmd, 2

````
## Libraries

```{r setup}
#| message: false
#| warning: false
library(broom)
library(tidyverse)
```
````

simon-5501-08-sway.qmd, 3

````
## Read data

```{r read-memory}
sway <- read_tsv(
  file="../data/postural-sway.txt",
  col_types="cnn")
names(sway) <- tolower(names(sway))
glimpse(sway)
```
````

simon-5501-08-sway.qmd, 4

````
## Boxplot of fbsway by age

```{r}
#| fig.height: 1.5
sway |>
  ggplot(aes(age, fbsway)) +
    geom_boxplot() +
    ggtitle("Graph drawn by Steve Simon on 2024-10-05") +
    xlab("Treatment group") +
    ylab("Front to back sway") +
    coord_flip()
```

The elderly patients have generally higher sway values. Both groups have
similar variation except for an extreme outlier in the elderly patients.
````

simon-5501-08-sway.qmd, 5

```
## Descriptive statistics for memory by treatment

```{r group-means}
sway |>
  group_by(age) |>
  summarize(
    fb_mn=mean(fbsway),
    fb_sd=sd(fbsway),
    n=n())
```

The average front-to-back sway is higher in the elderly patients. There is
more variation in the elderly group, possibly caused by the extreme outlier.
```

simon-5501-08-sway.qmd, 6

````
## Two-sample t-test

```{r t-test}
m1 <- t.test(
  fbsway ~ age,
  data=sway,
  alternative="two.sided",
  var.equal=TRUE)
m1
```
````

simon-5501-08-sway.qmd, 7

```
## Equivalent analysis using linear regression

```{r lm}
m2 <- lm(fbsway ~ age, data=sway)
tidy(m2)
confint(m2)
```
```

Break #4

- What you have learned
  - R code for the t-test
- What's coming next
  - Confidence intervals

Confidence interval for difference in means

- $\bar{X}_1 - \bar{X}_2 \pm t(1-\alpha/2, n_1+n_2-2)se$
    - Range of plausible values for $\mu_1 - \mu_2$

```
	Two Sample t-test

data:  fbsway by age
t = 2.2044, df = 15, p-value = 0.04353
alternative hypothesis: true difference in means between group Elderly and
group Young is not equal to 0
95 percent confidence interval:
  0.2715453 16.1451214
sample estimates:
mean in group Elderly   mean in group Young 
             26.33333              18.12500 
```

Speaker notes

The formula for a confidence interval for the difference between two means is shown here. The calculations are tedious, but not difficult.

Interpretation

- Statement about population mean difference ($\mu_1 - \mu_2$)
- Range of plausible values
- Not a probability statement
  - 95% confidence does not mean 95% probability
- If you collected 100 independent samples,
  - Roughly 95 would contain $\mu_1 - \mu_2$

Break #5

- What you have learned
  - Confidence intervals
- What's coming next
  - Sample size justification

Three things you need to justify your sample size

1. Research hypothesis
2. Measure of variability
3. Minimum clinically important difference (MCID)

Scenario, 1

- Replicate postural sway study
  - Different populations
  - Same outcome measure
- Research hypothesis, $H_0\ \mu_1 - \mu_2 = 0$
- Standard deviations: 9.77, 4.09
- MCID = 4

Scenario, 2

```r
sample_size_estimate <- power.t.test(
  n=NULL,
  delta=4,
  sd=9.8,
  sig.level=0.05,
  power=0.9,
  type="two.sample",
  alternative="two.sided")
```

Scenario, 3

```
     Two-sample t test power calculation

              n = 127.1097
          delta = 4
             sd = 9.8
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Break #6

- What you have learned
  - Sample size justification
- What's coming next
  - Your homework

simon-5501-08-directions.md, 1

```
---
title: "Directions for 5501-08 programming assignment"
format:
  html:
    embed-resources: true
---

This programming assignment was written by Steve Simon on 2024-10-08 and is
placed in the public domain.
```

simon-5501-08-directions.md, 2

```
## Program

-   Download the [sway program][sway] and the [sample size program][sample]
    -   Store it in your src folder
-   Modify the file names
    -   Use your last name instead of "simon"
-   Modify the documentation headers
    -   Add your name
    -   Optional: change the copyright statement

[sway]: https://github.com/pmean/classes/blob/master/general/simon-5501-08-sway.qmd
[sample]: https://github.com/pmean/classes/blob/master/general/simon-5501-08-sample-size.qmd
```

simon-5501-08-directions.md, 3

```
## Data

-   Download the [data][dat] file
    -   Store it in your data folder
-   Refer to the [data dictionary][dic], if needed.

[dat]: https://github.com/pmean/data/blob/main/files/postural-sway.txt
[dic]: https://github.com/pmean/data/blob/main/files/postural-sway.yaml
```

simon-5501-08-directions.md, 4

```
## Question 1

Replicate the analysis shown in the program, using side-to-side sway. Modify
all the interpretations appropriately.
```

simon-5501-08-directions.md, 5

```
## Question 2

Evaluate the sample size calculations for all the scenarios listed
```

simon-5501-08-directions.md, 6

```
## Your submission

-   Save the output in html format
-   Convert it to pdf format.
-   Make sure that the pdf file includes
    -   Your last name
    -   The number of this course
    -   The number of this module
-   Upload the file
```

simon-5501-08-directions.md, 7

```
## If it doesn't work

If your program has any errors or fails to produce the output that you desire
and you can't resolve the problem, upload the program file along with the pdf
file to help us figure out what went wrong. You will get a chance to resubmit
the assignment if needed.
```

Summary

- What you have learned
  - The two-sample t-test
  - The t-distribution
  - Critical values and p-values
  - R code for the t-test
  - Confidence intervals
  - Sample size justification
  - Your homework