NHST II

PSYC 2020-A01 / PSYC 6022-A01 | 2025-10-10 | Lab 8

Jessica Helmer

Outline

- Assignment 7 Review
- NHST II Review
- Import Deep Dive

Learning objectives:
R: Importing

Hypothesis Testing Demo

We start by assuming the null hypothesis (H_0) is true. Let's say our null hypothesis is that the population mean of some happiness scale is 10 ($\mu = 10$). In this situation, we also know that the population variance on this scale is 4 ($\sigma^2 = 4, \sigma = 2$).

Let's set this up:

Hypothesis Testing Demo

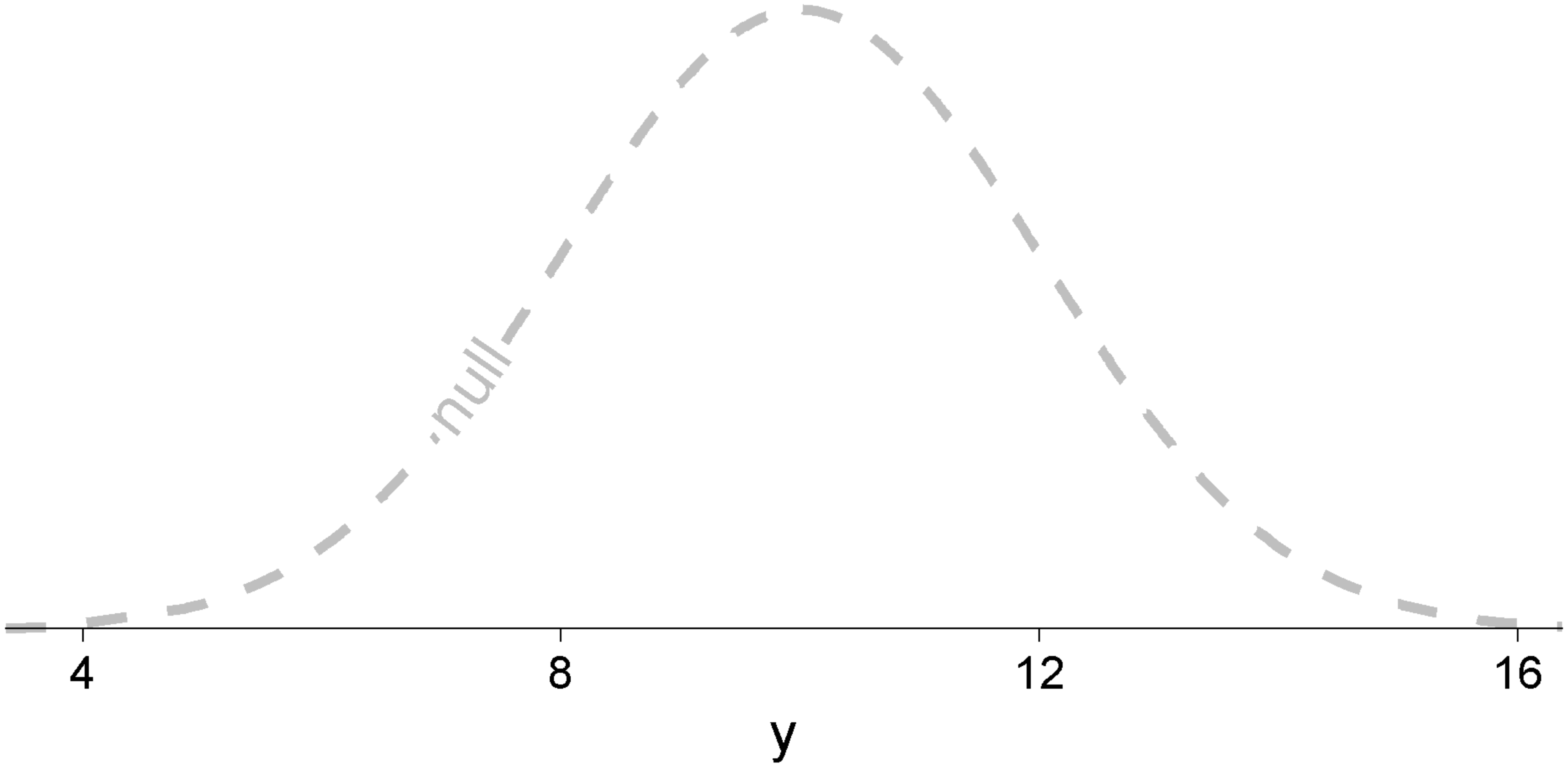
```r
n_pop <- 1000
pop_mean <- 10
pop_sd <- 2
```

```r
null <- data.frame(x = seq(-20, 20, length.out = n_pop)) |>
  mutate(y = rnorm(n_pop, pop_mean, pop_sd),
         dens = dnorm(y, pop_mean, pop_sd))
```

If the null hypothesis were true, this is what we’d expect to see:

Hypothesis Testing Demo: Scenario 1

Plot Code



Hypothesis Testing Demo: Scenario 1

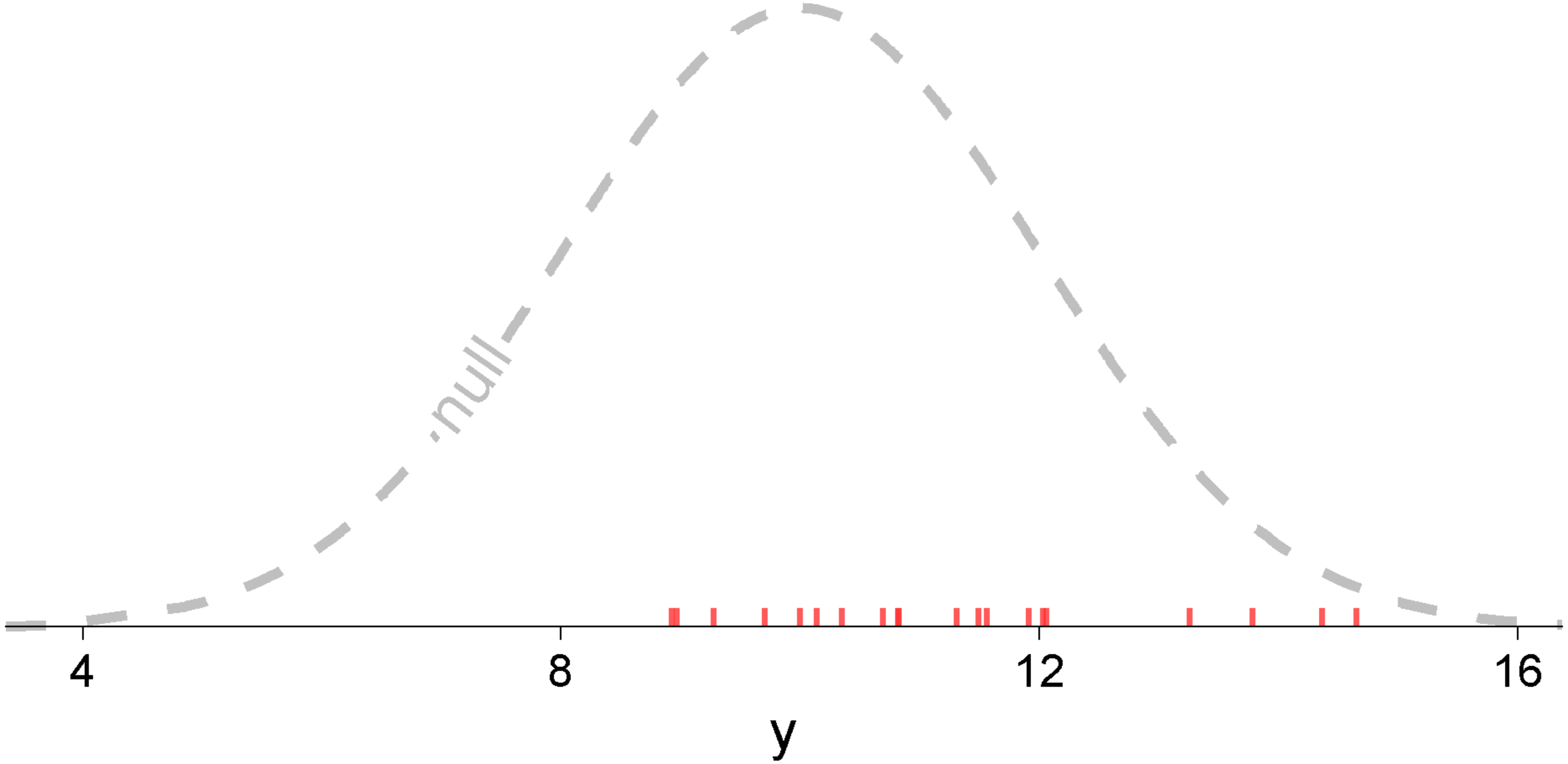
Let’s then collect a sample of 20 people. Say they have a sample mean of about 12.

```r
n_samp <- 20
samp_mean <- 12
```

```r
alt <- data.frame(x = seq(-20, 20, length.out = n_samp)) |>
  mutate(y = rnorm(n_samp, samp_mean, pop_sd),
         dens = dnorm(y, samp_mean, pop_sd))
```

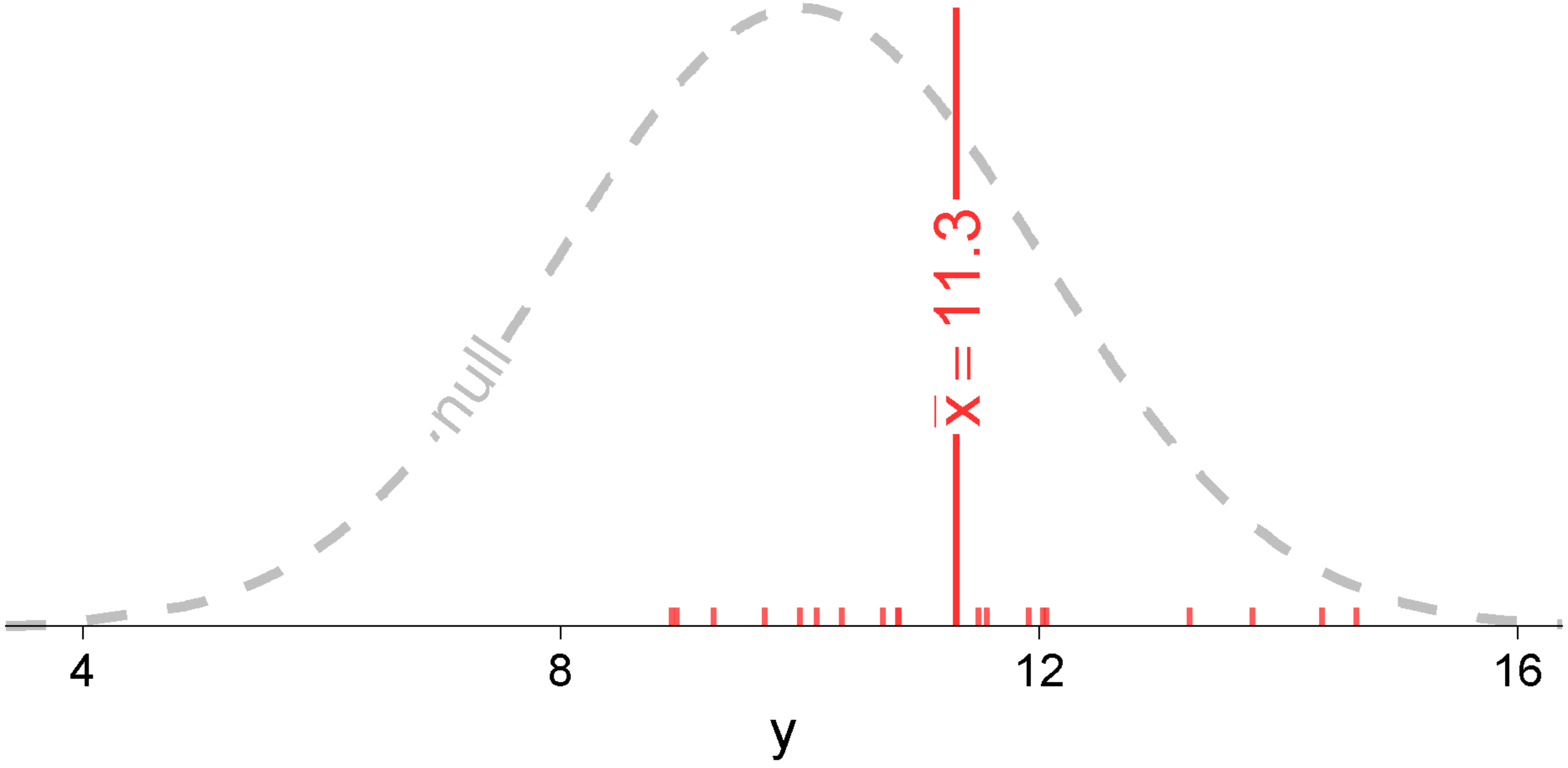
Hypothesis Testing Demo: Scenario 1

Plot Code



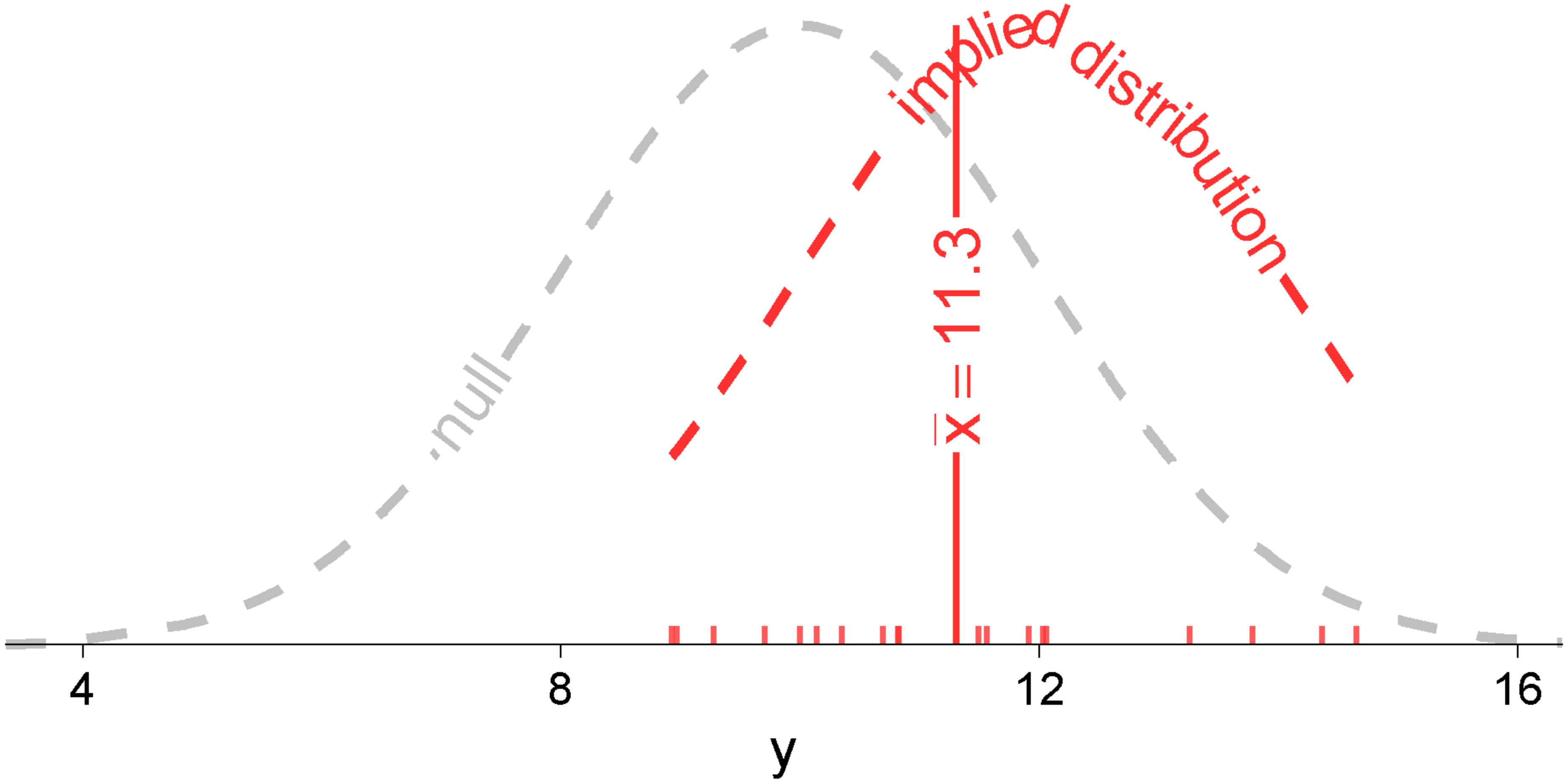
Hypothesis Testing Demo: Scenario 1

Plot Code



Hypothesis Testing Demo: Scenario 1

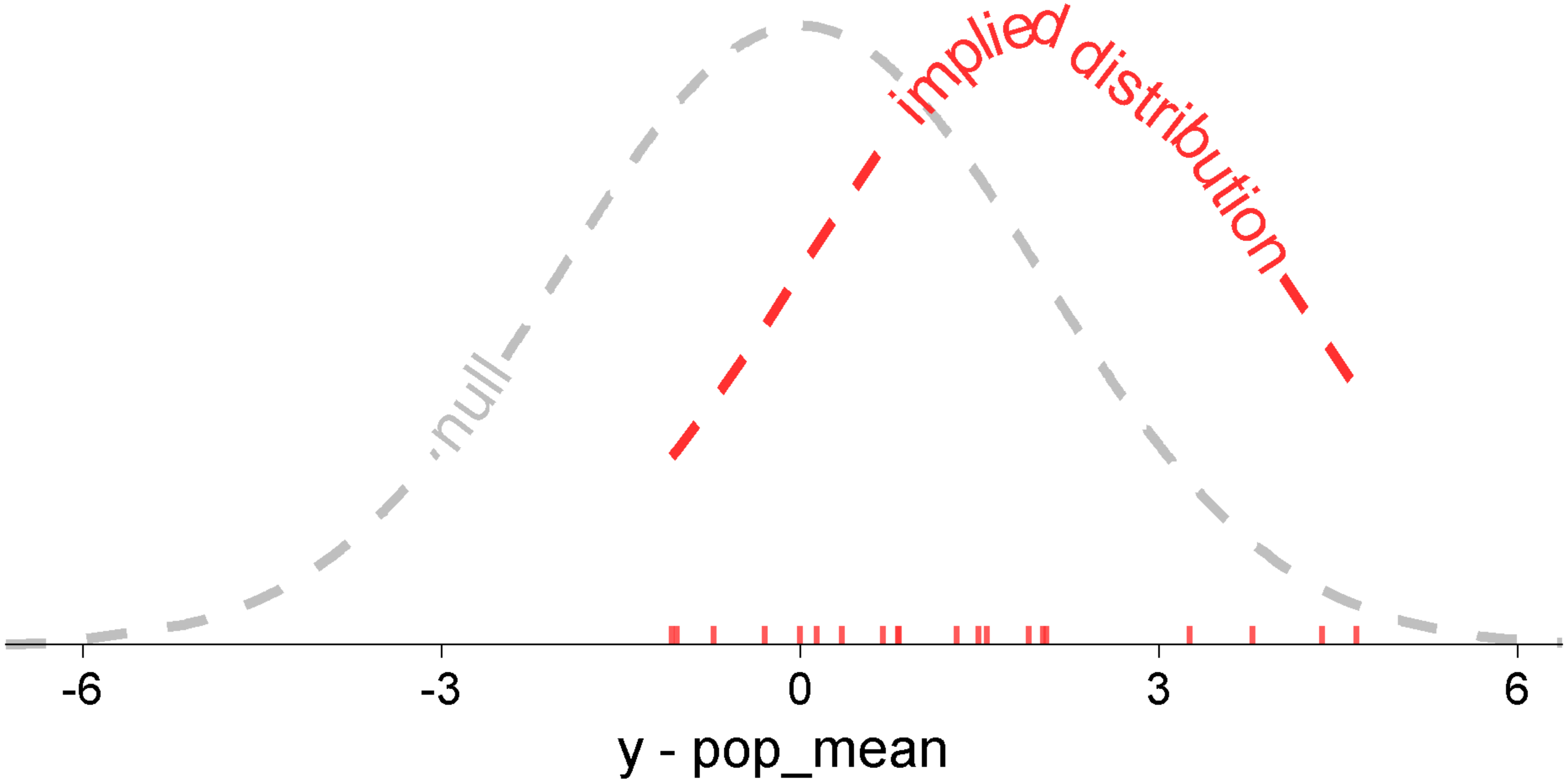
Plot Code



Hypothesis Testing Demo: Scenario 1

Plot

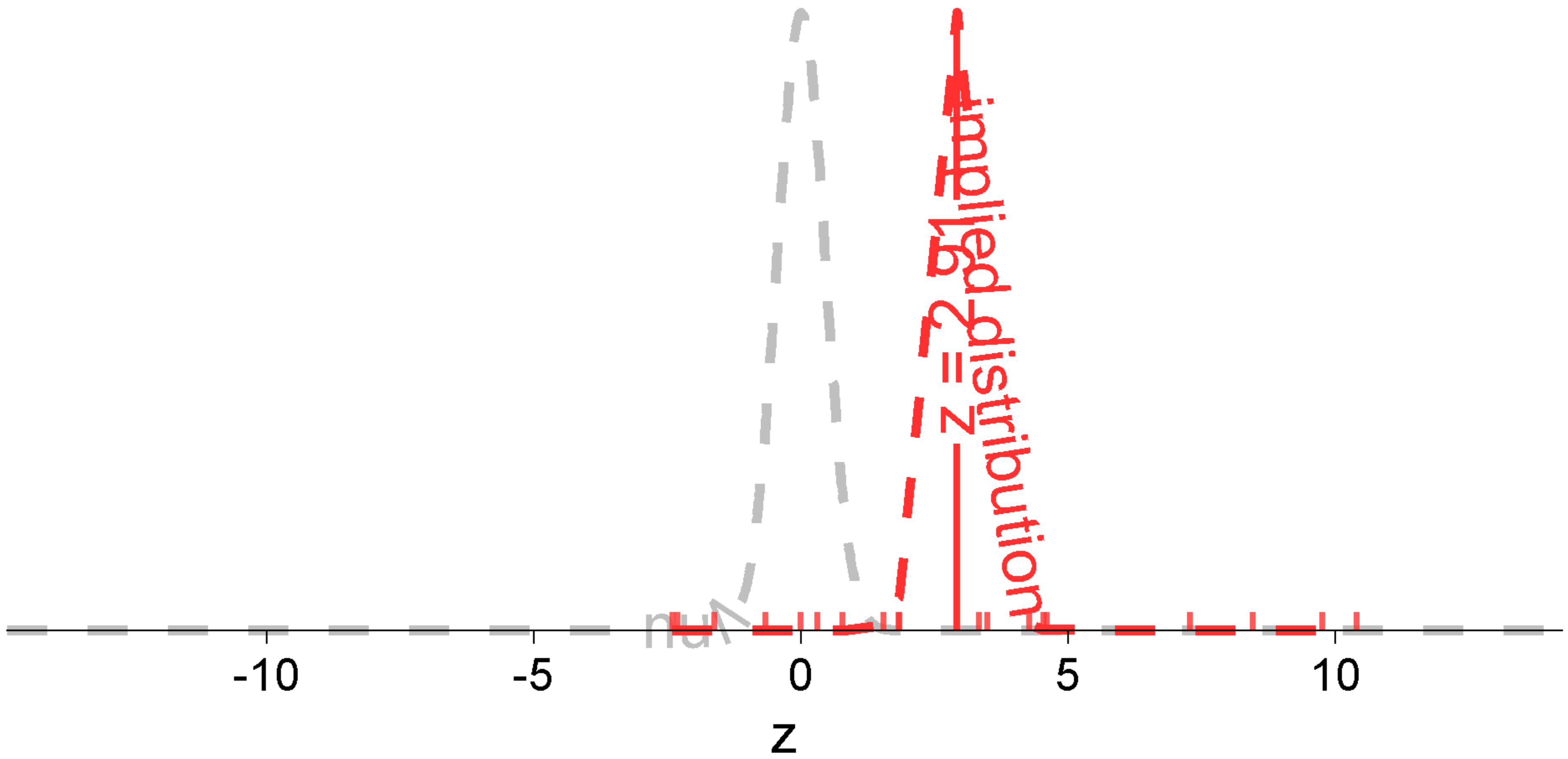
Code



Hypothesis Testing Demo: Scenario 1

Plot

Code



Hypothesis Testing Demo: Scenario 2

Now let’s try an example with the same mean difference, but the population variance is larger

$$\mu = 10, \sigma^2 = 64, \sigma = 8$$

Let’s then collect a sample of 20 people. Say they have a sample mean of about 12.

```r
pop_sd <- 8
n_samp <- 20
samp_mean <- 12
```

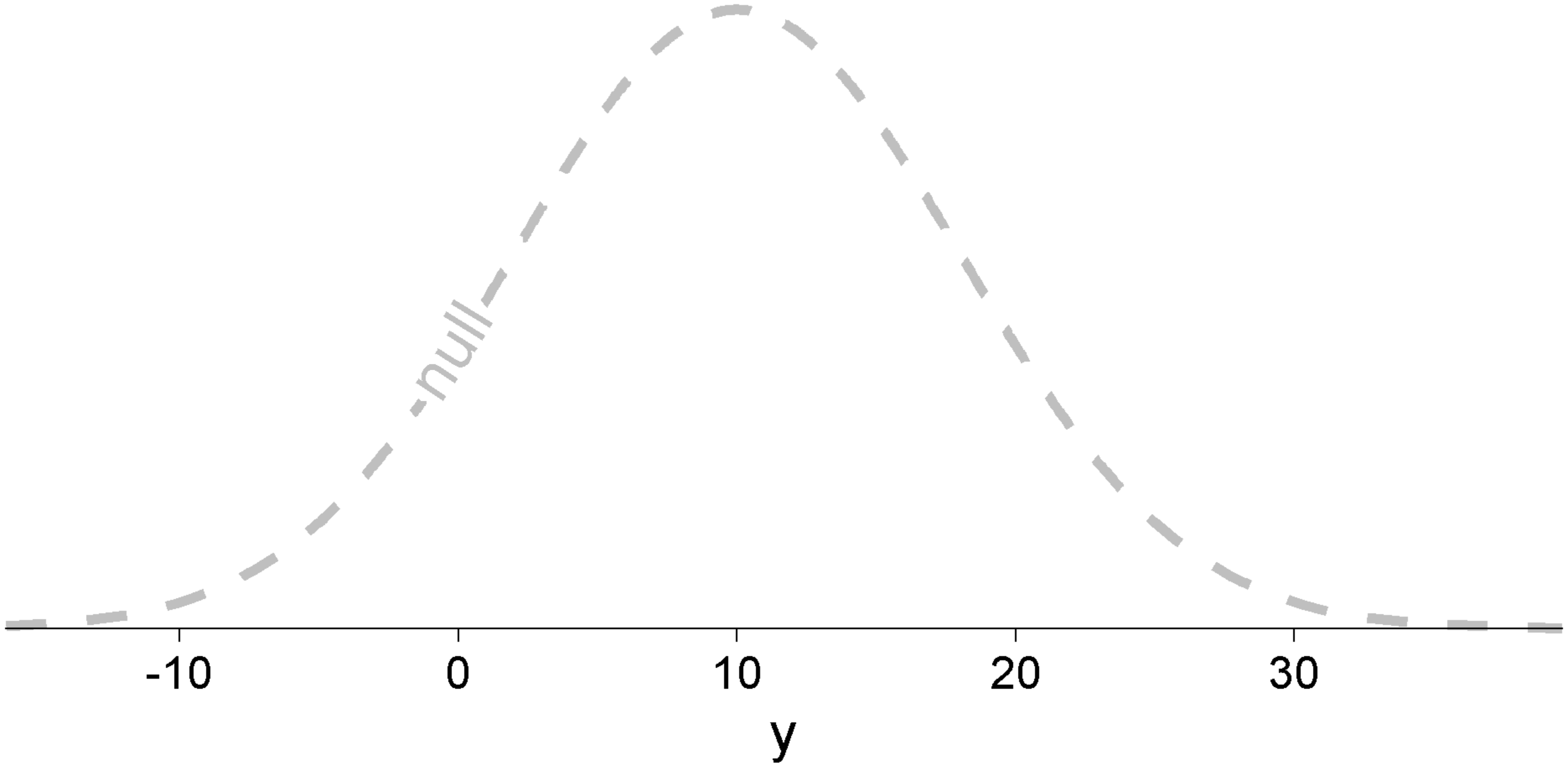
```r
null <- data.frame(x = seq(-20, 20, length.out = n_pop)) |>
  mutate(y = rnorm(n_pop, pop_mean, pop_sd),
         dens = dnorm(y, pop_mean, pop_sd))

alt <- data.frame(x = seq(-20, 20, length.out = n_samp)) |>
```

```
mutate(y = rnorm(n_samp, samp_mean, pop_sd),
       dens = dnorm(y, samp_mean, pop_sd))
```

Hypothesis Testing Demo: Scenario 2

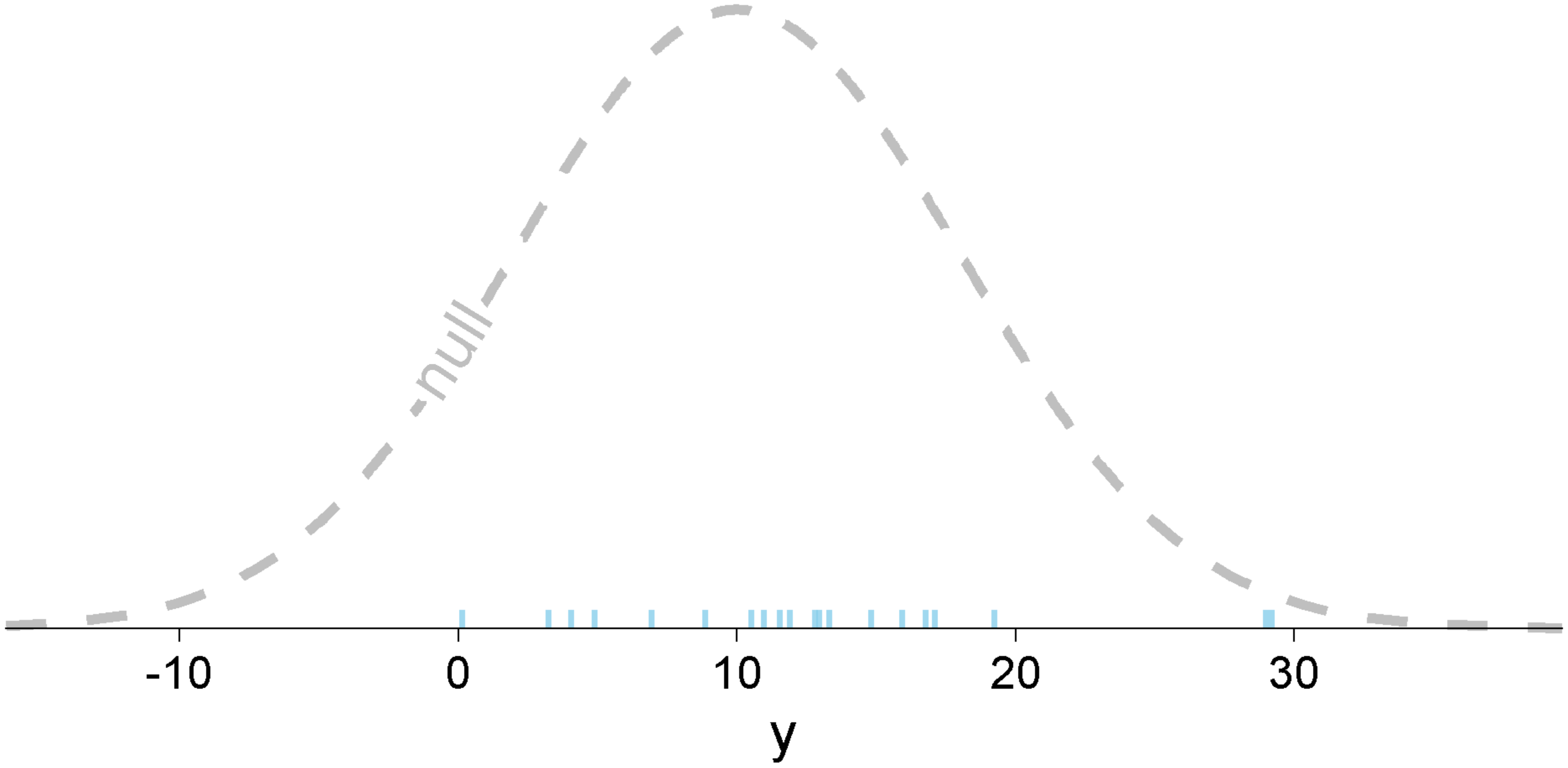
Plot Code



Hypothesis Testing Demo: Scenario 2

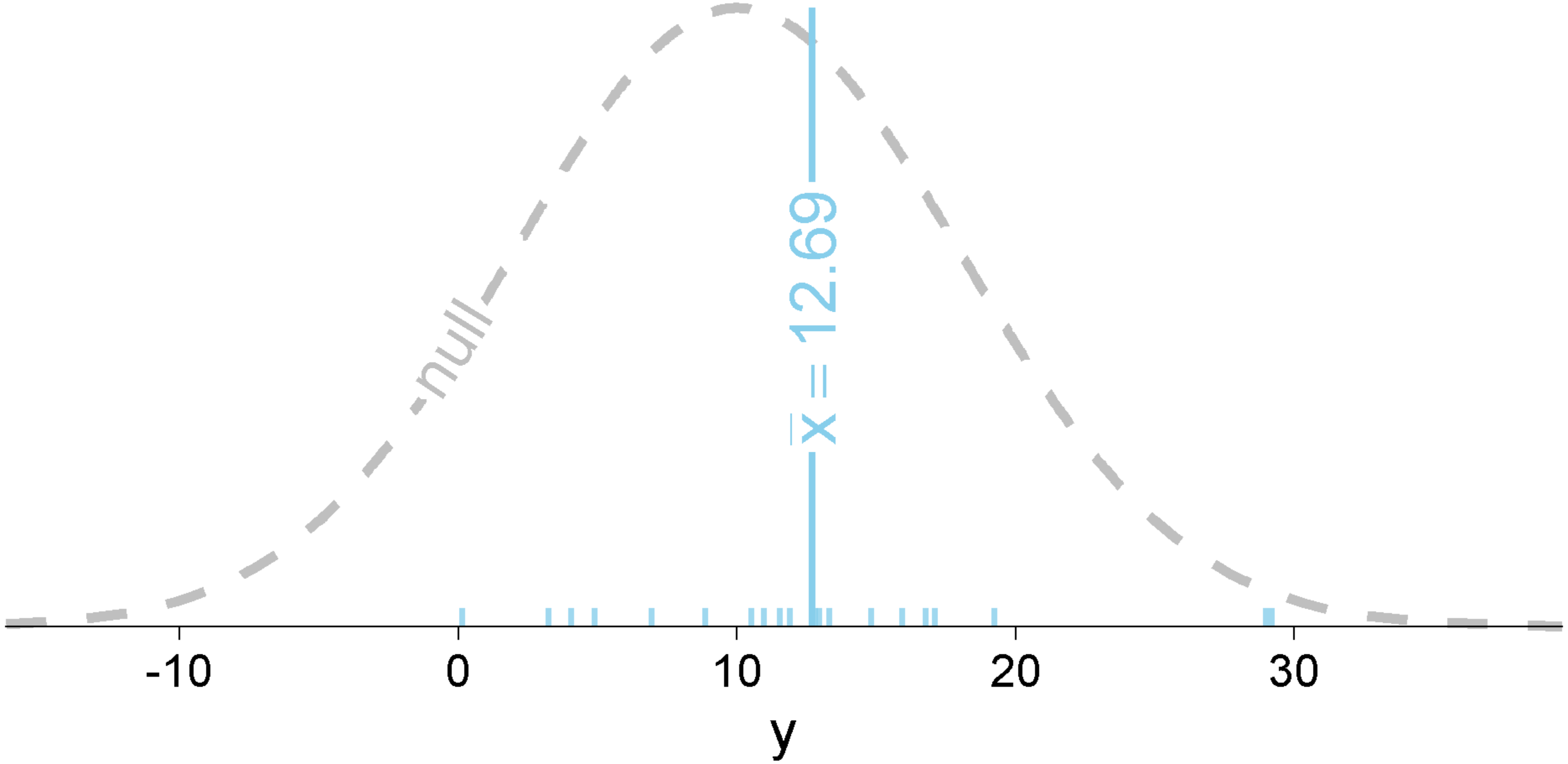
Plot

Code



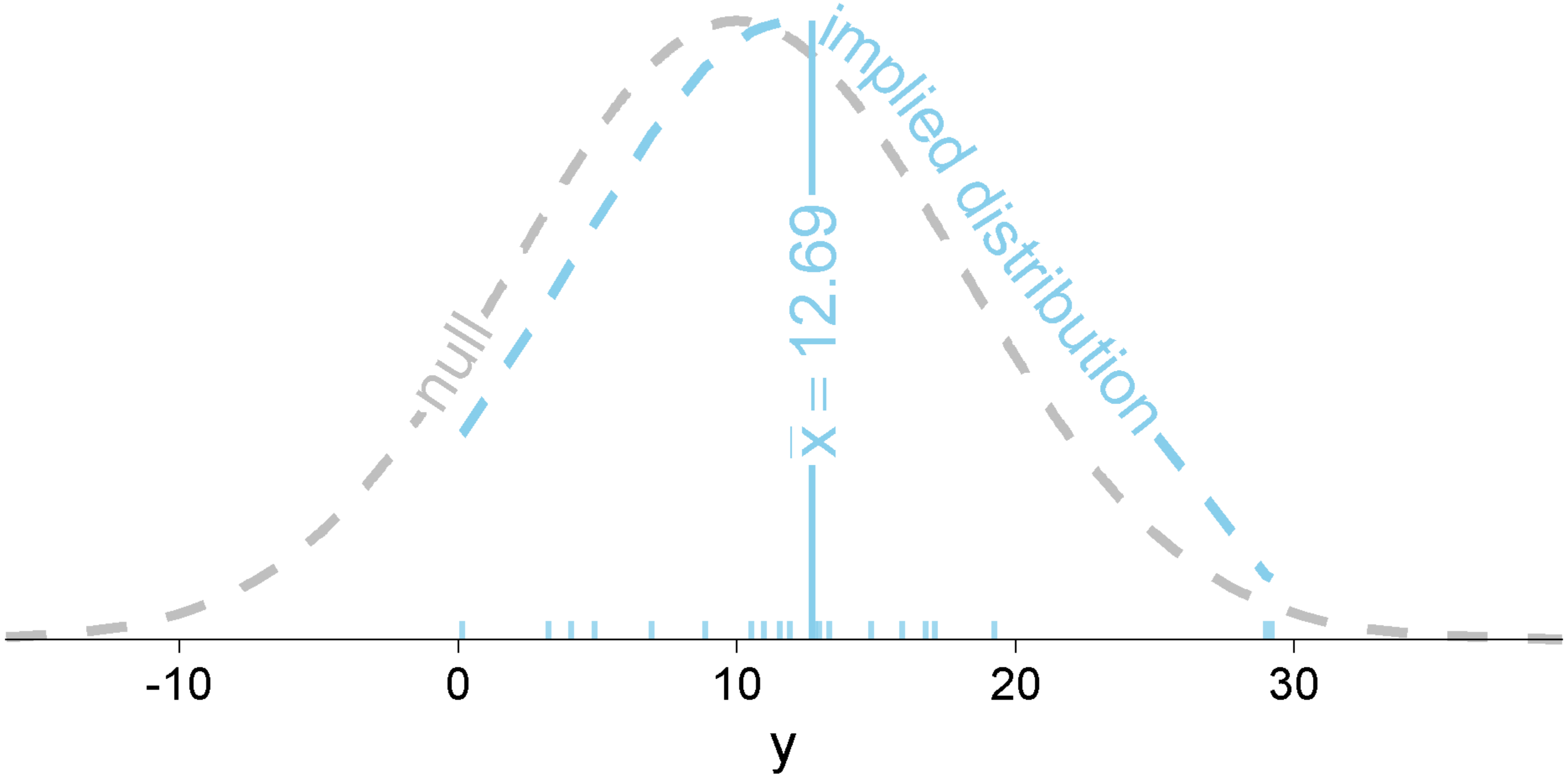
Hypothesis Testing Demo: Scenario 2

Plot Code



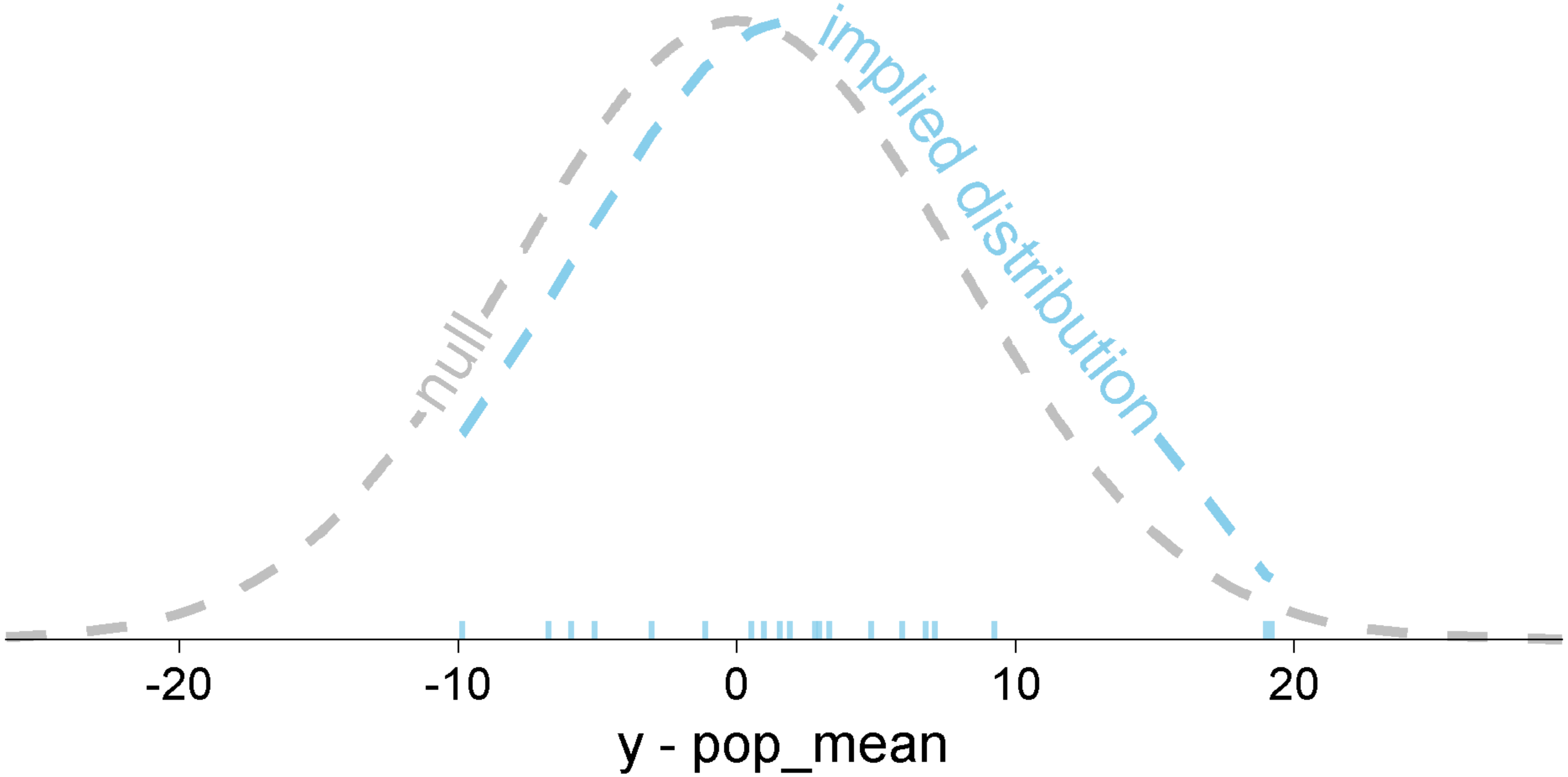
Hypothesis Testing Demo: Scenario 2

Plot | Code



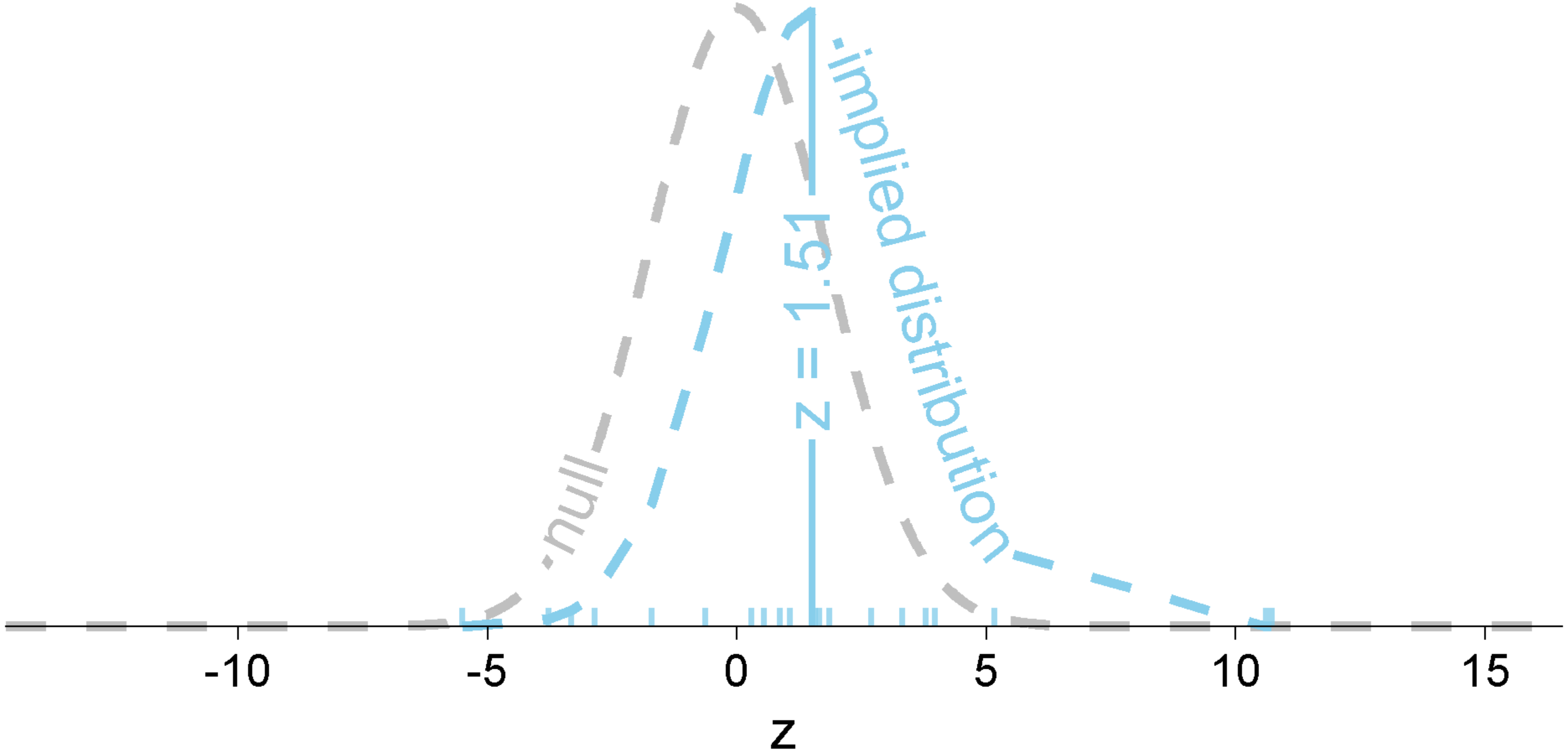
Hypothesis Testing Demo: Scenario 2

Plot Code



Hypothesis Testing Demo: Scenario 2

Plot Code



Hypothesis Testing Demo: Scenario 3

Let’s try the same population parameters, but take a bigger sample!

Let’s then collect a sample of 200 people. Say they have a sample mean of about 12.

```r
n_samp <- 200
```

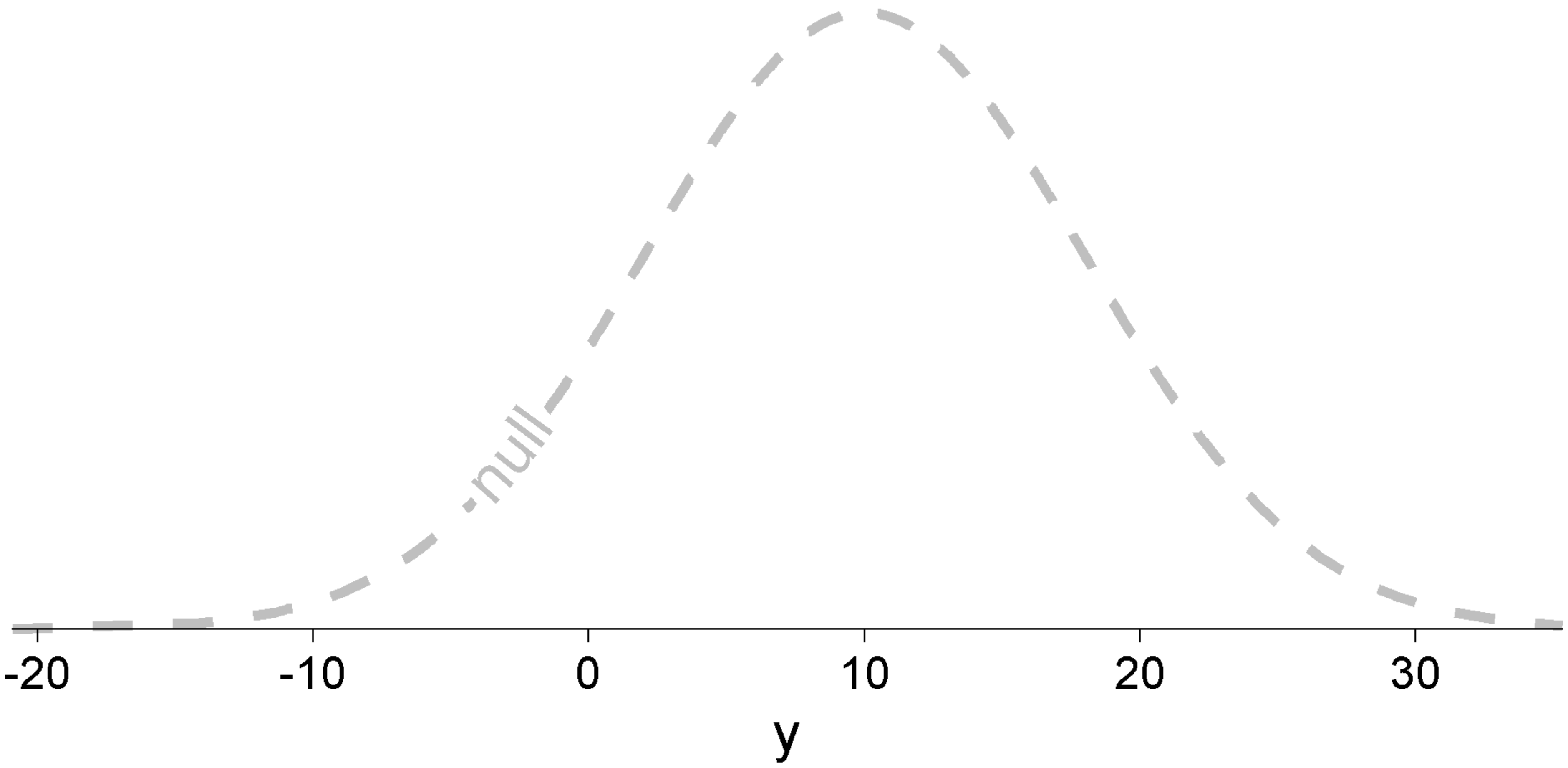
```r
null <- data.frame(x = seq(-20, 20, length.out = n_pop)) |>
  mutate(y = rnorm(n_pop, pop_mean, pop_sd),
         dens = dnorm(y, pop_mean, pop_sd))

alt <- data.frame(x = seq(-20, 20, length.out = n_samp)) |>
  mutate(y = rnorm(n_samp, samp_mean, pop_sd),
         dens = dnorm(y, samp_mean, pop_sd))
```

Hypothesis Testing Demo: Scenario 3

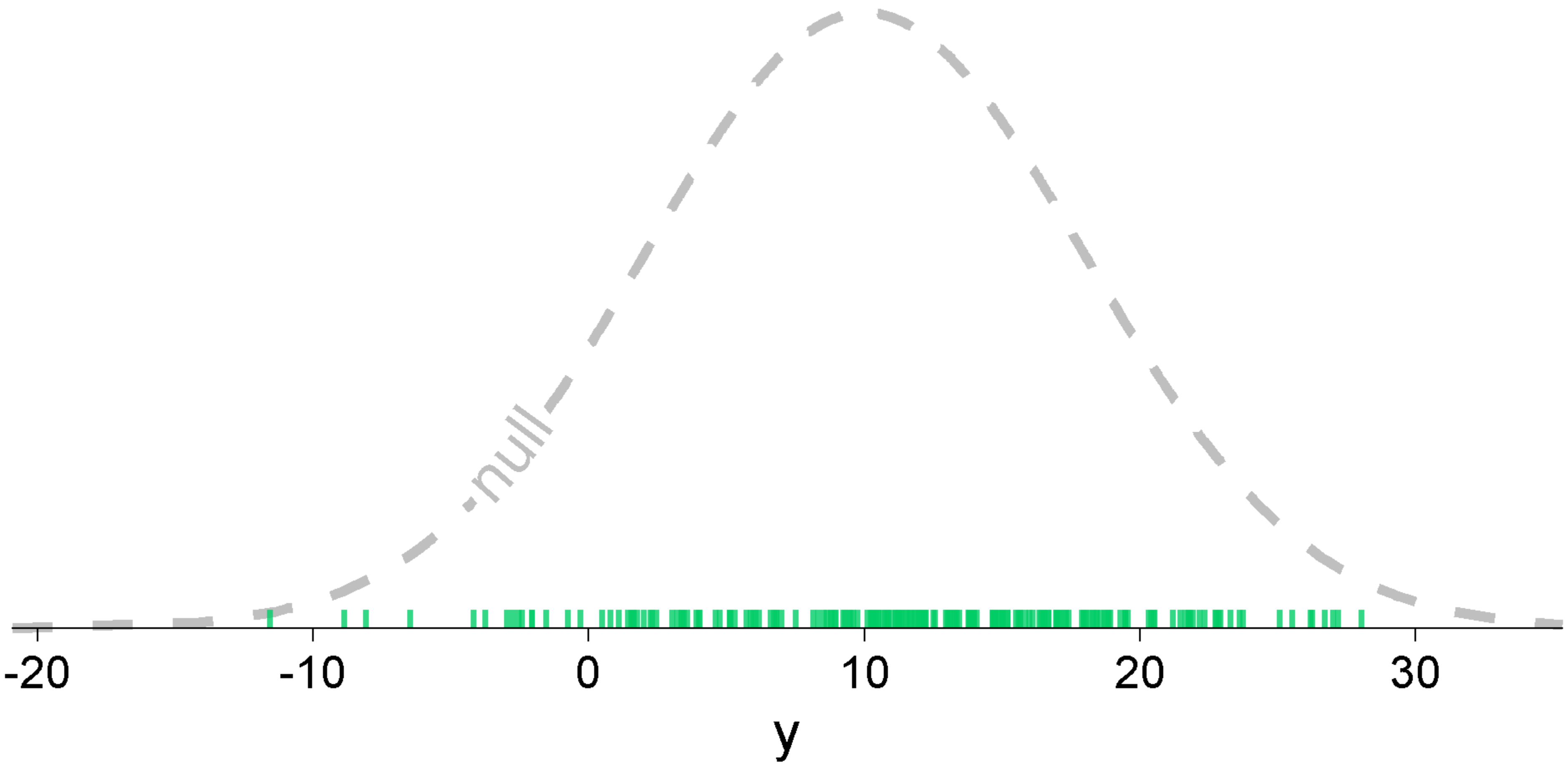
Plot

Code



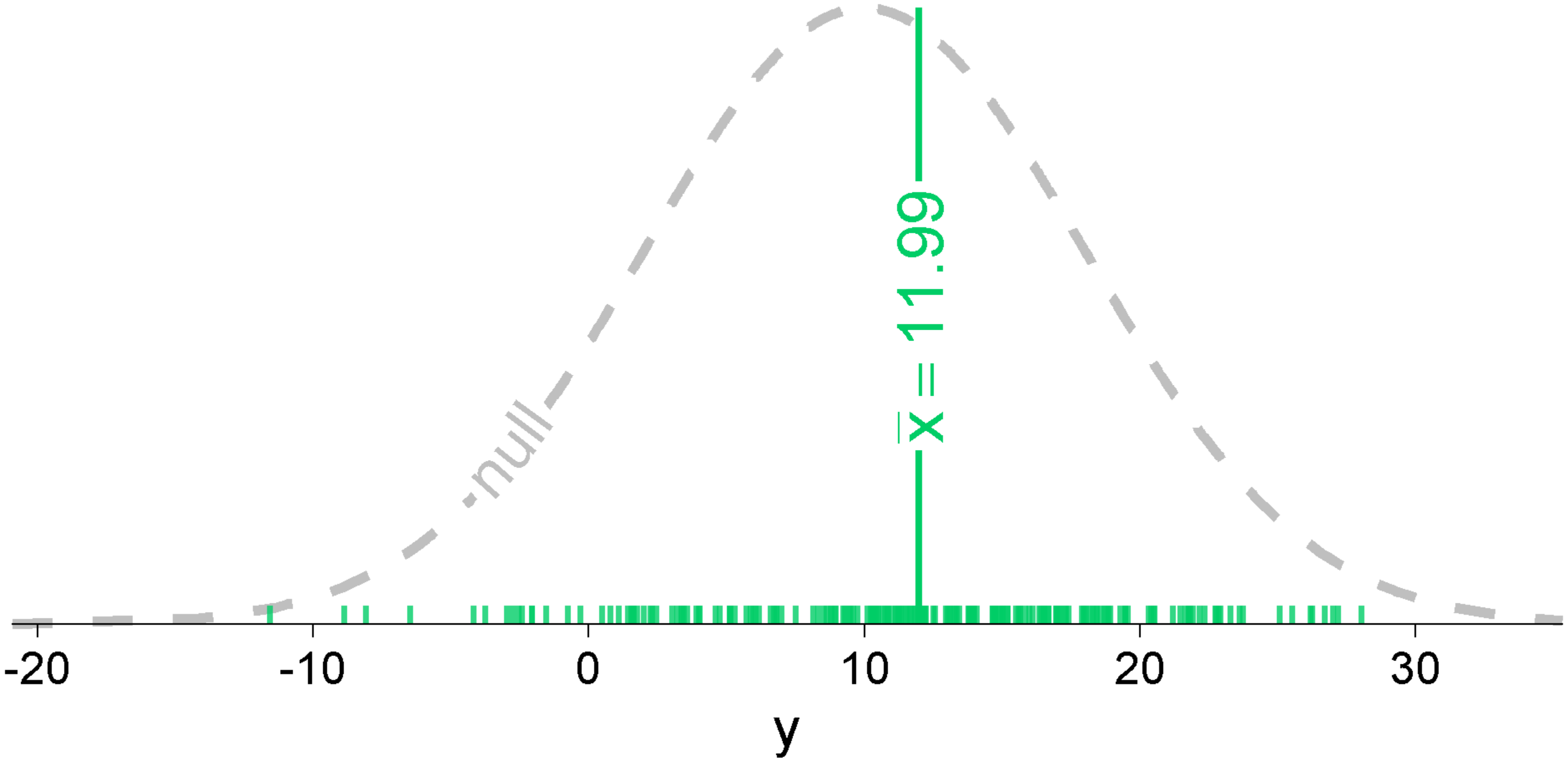
Hypothesis Testing Demo: Scenario 3

Plot Code



Hypothesis Testing Demo: Scenario 3

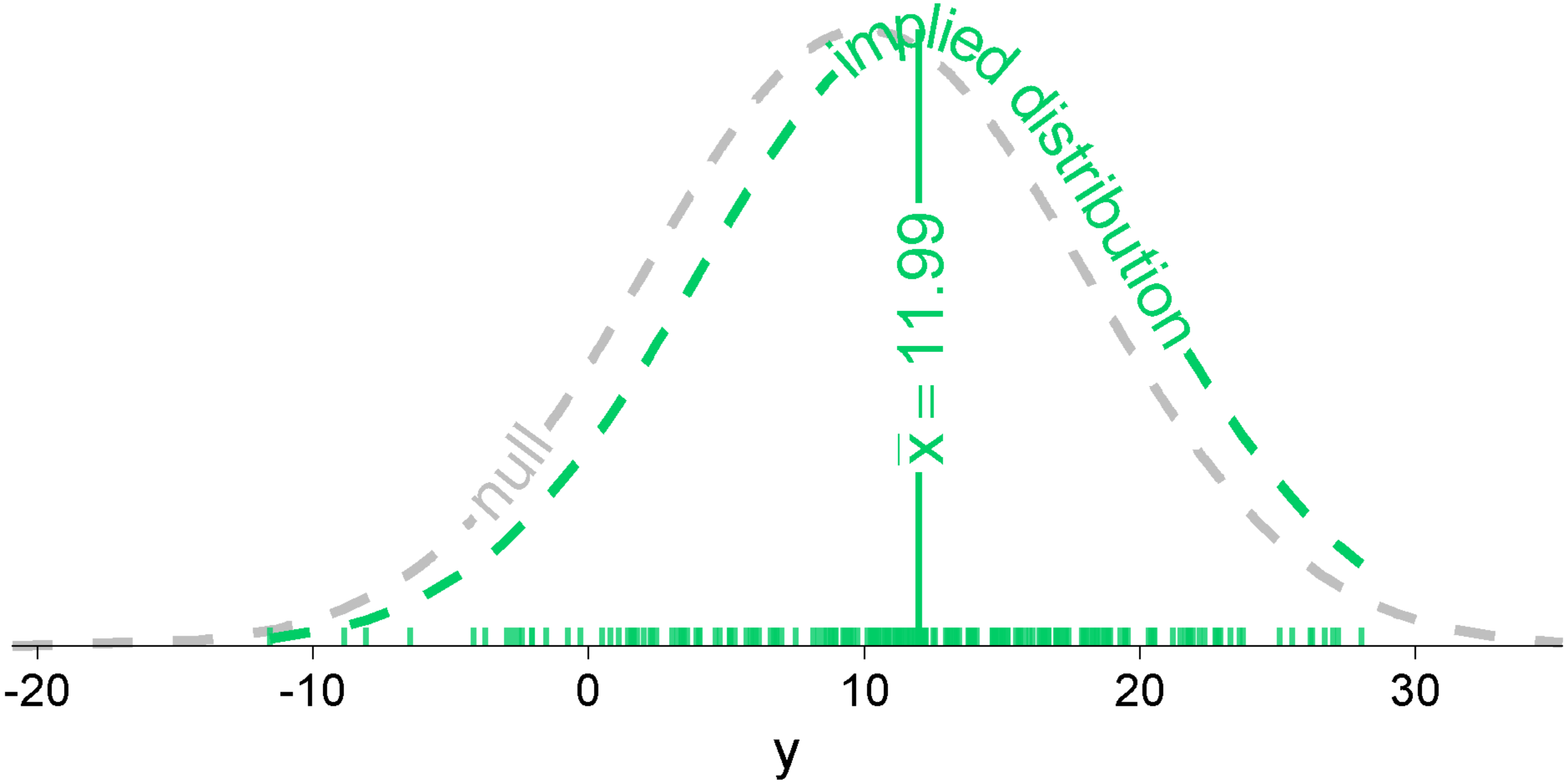
Plot Code



Hypothesis Testing Demo: Scenario 3

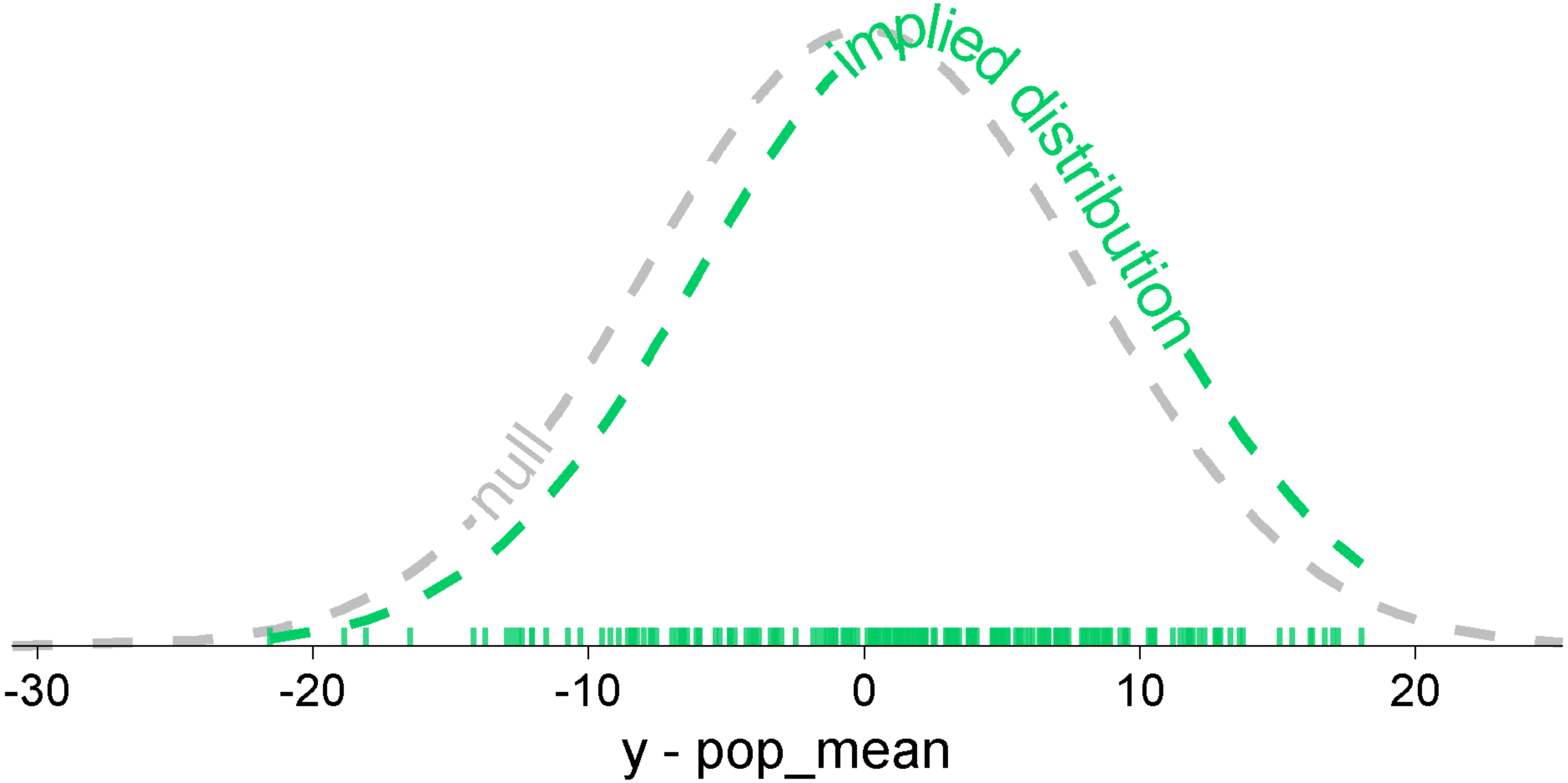
Plot

Code



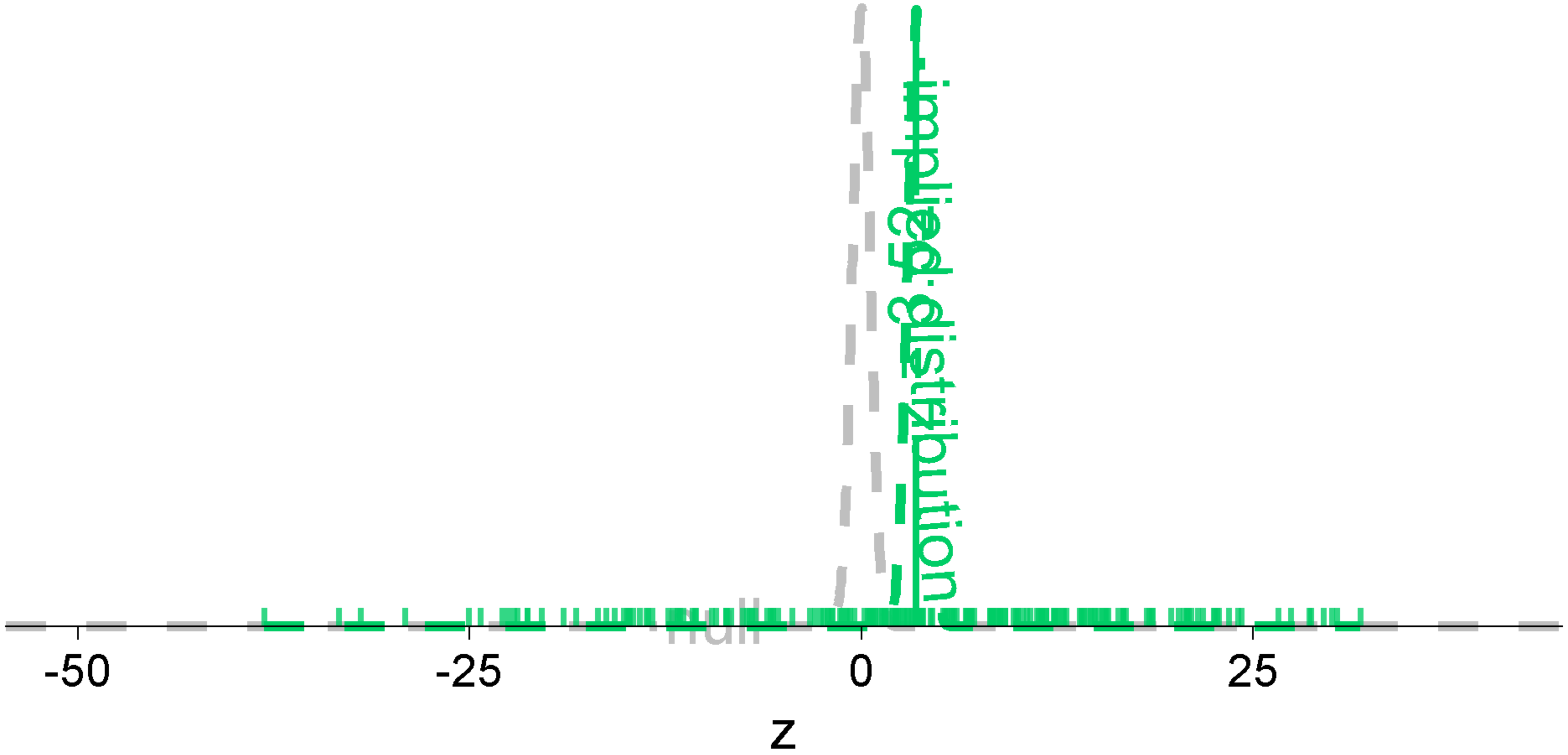
Hypothesis Testing Demo: Scenario 3

Plot Code



Hypothesis Testing Demo: Scenario 3

Plot | Code



Hypothesis Testing Demo: Scenario 4

Now let’s say there’s a huge difference between our population mean and sample mean. But, we only took a sample of 5 people.

$$\mu = 10, \sigma^2 = 4, \sigma = 2$$

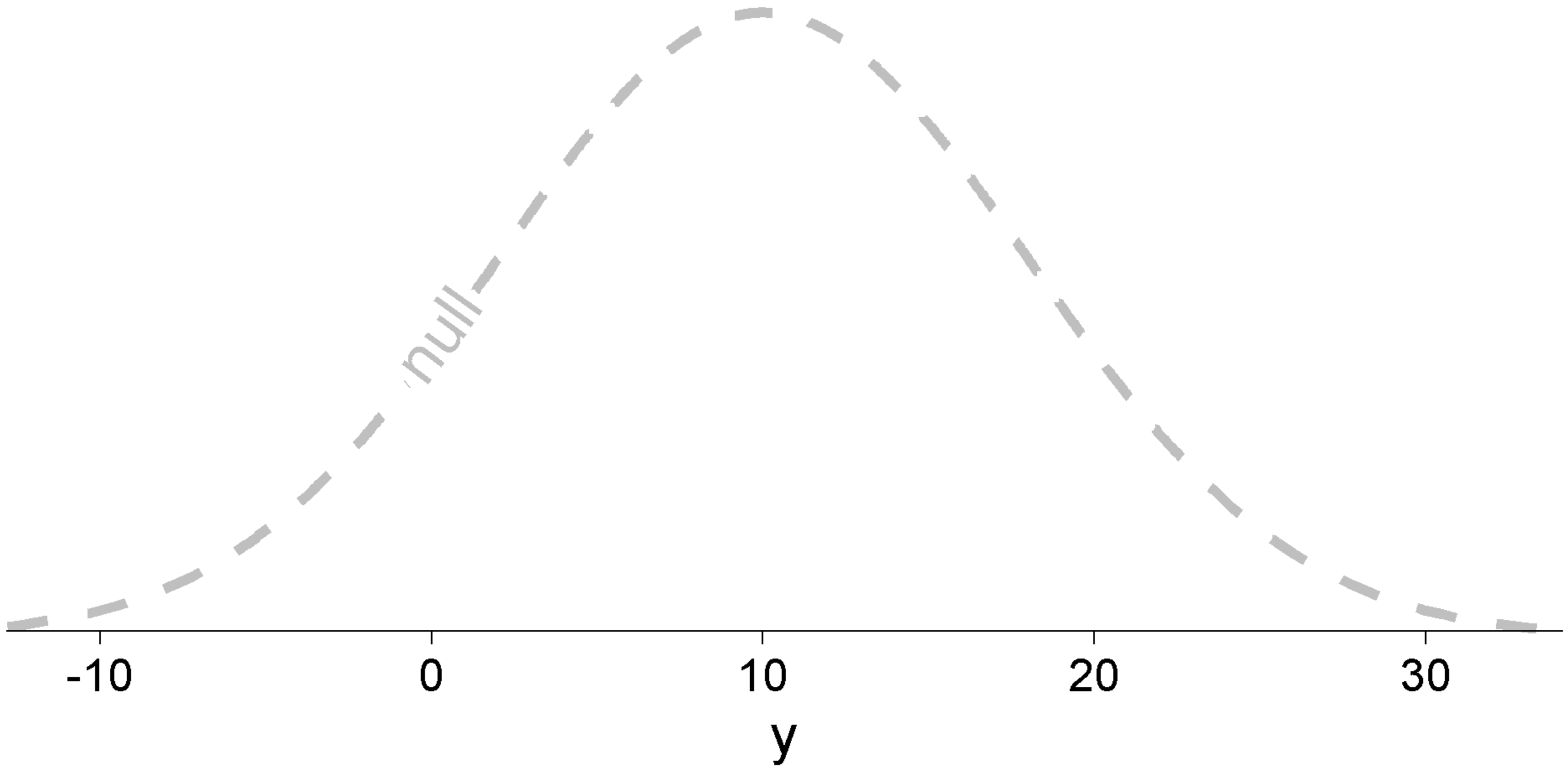
```r
n_samp <- 5
samp_mean <- 20
```

```r
null <- data.frame(x = seq(-20, 20, length.out = n_pop)) |>
  mutate(y = rnorm(n_pop, pop_mean, pop_sd),
         dens = dnorm(y, pop_mean, pop_sd))

alt <- data.frame(x = seq(-20, 20, length.out = n_samp)) |>
  mutate(y = rnorm(n_samp, samp_mean, pop_sd),
         dens = dnorm(y, samp_mean, pop_sd))
```

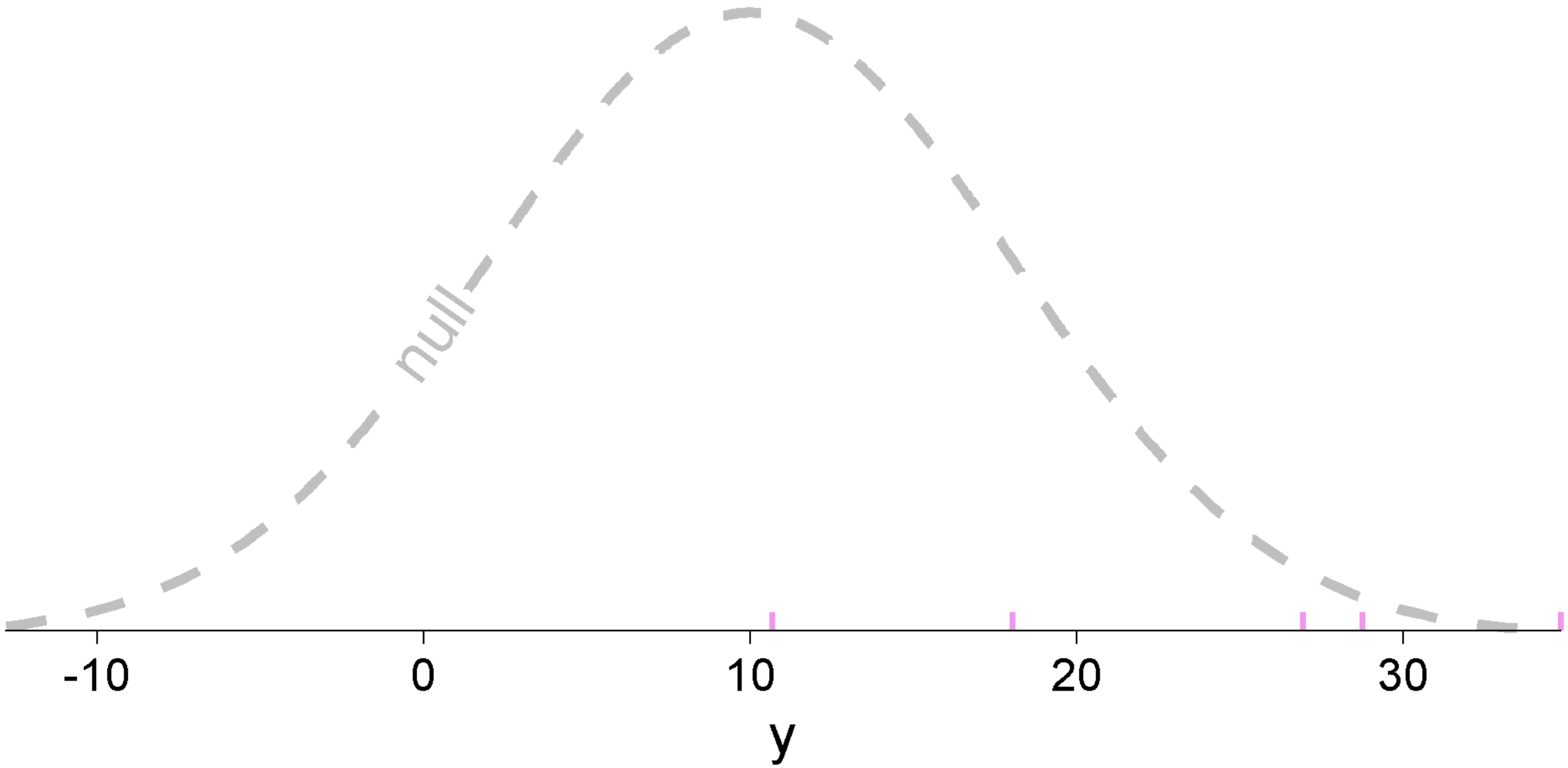
Hypothesis Testing Demo: Scenario 4

Plot Code



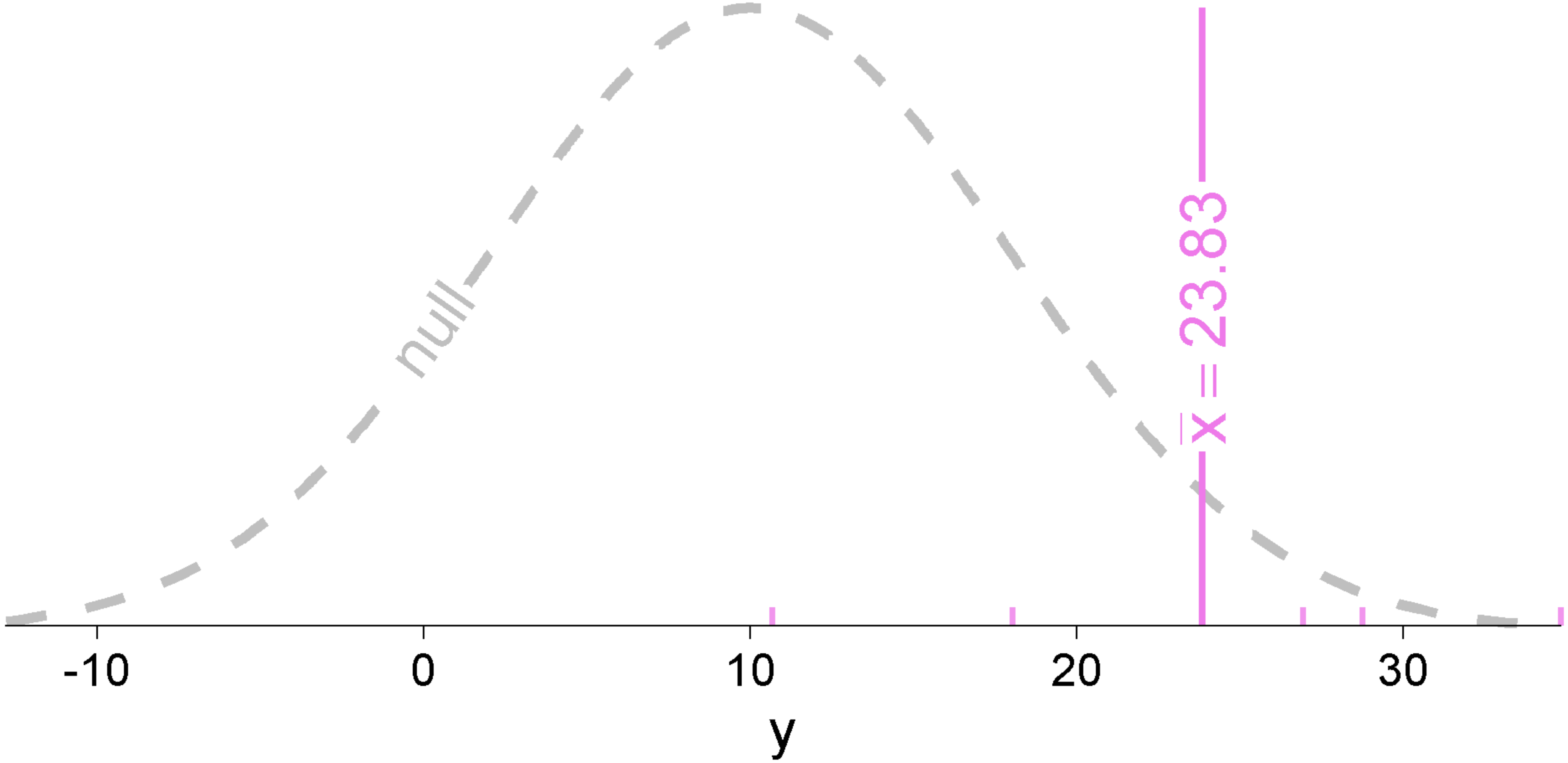
Hypothesis Testing Demo: Scenario 4

Plot Code



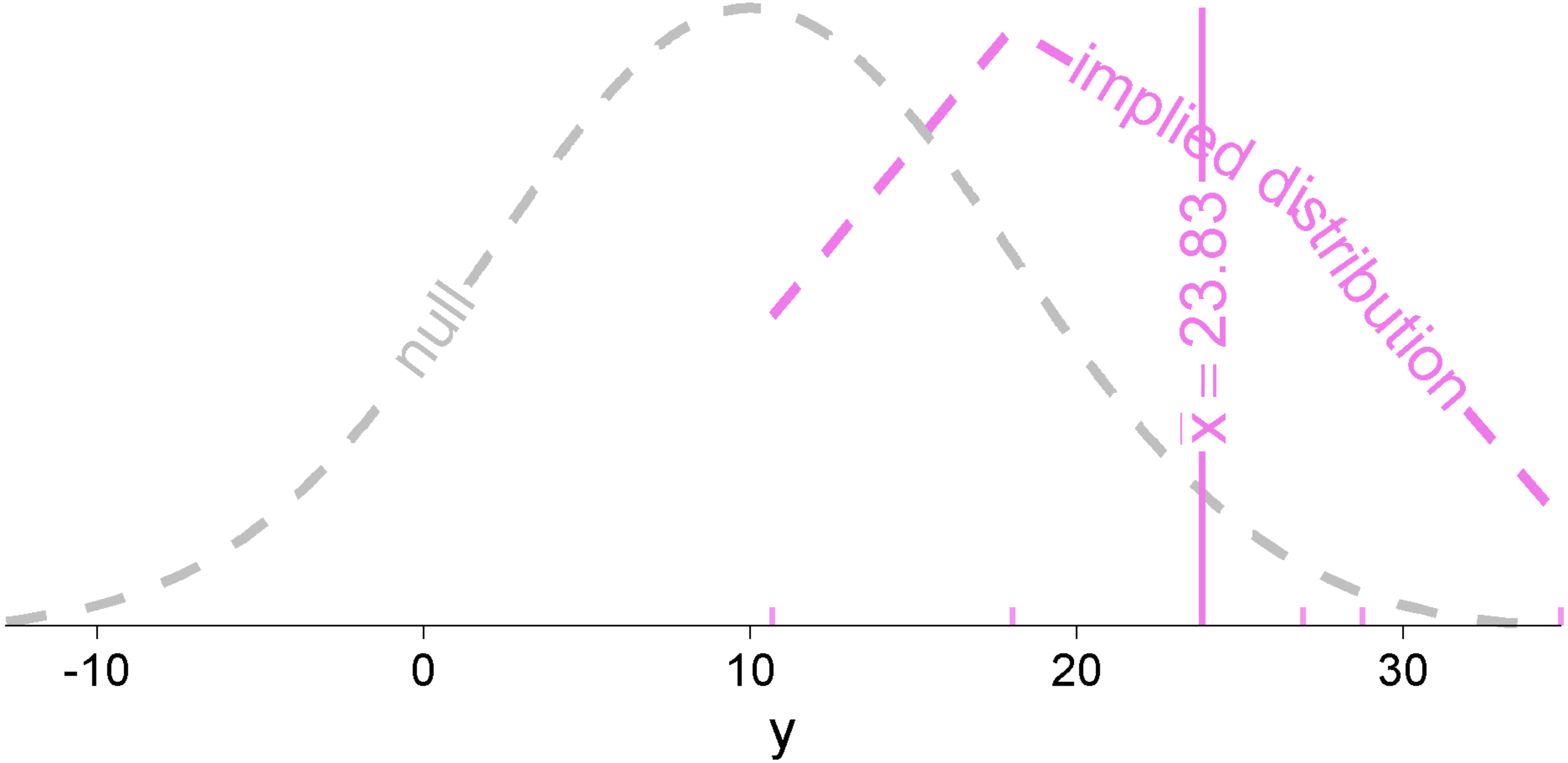
Hypothesis Testing Demo: Scenario 4

Plot Code



Hypothesis Testing Demo: Scenario 4

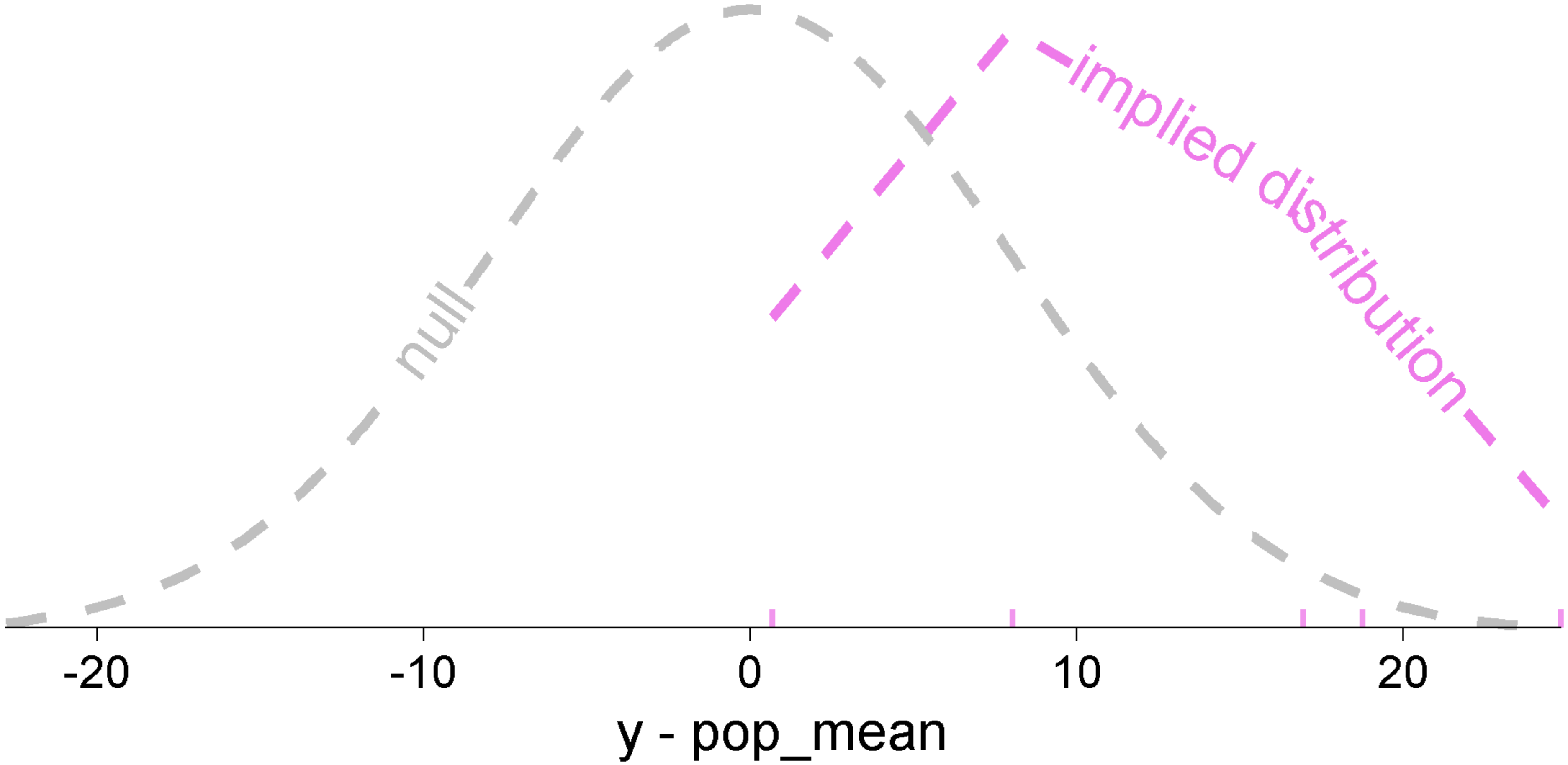
Plot Code



Hypothesis Testing Demo: Scenario 4

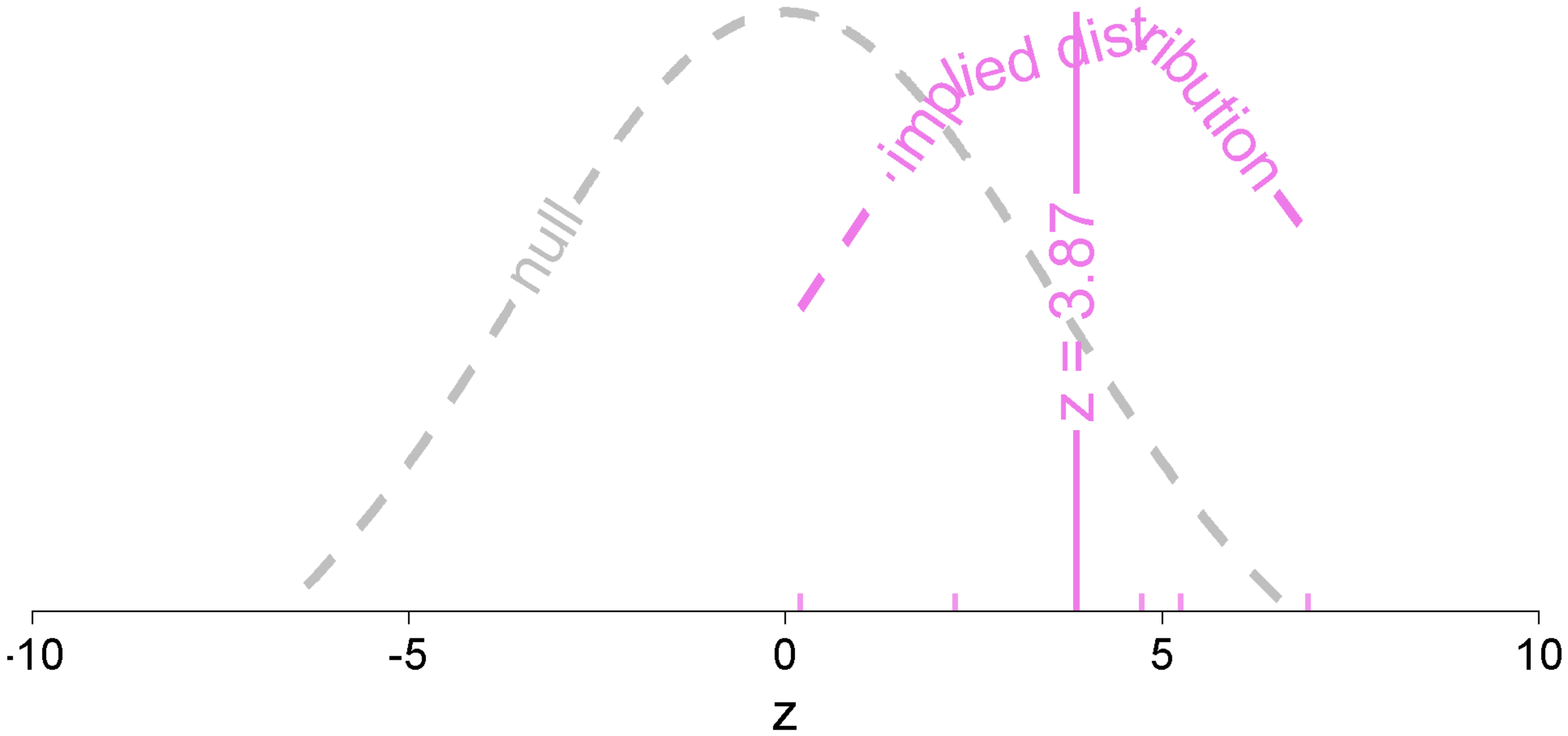
Plot

Code



Hypothesis Testing Demo: Scenario 4

Plot Code



Example Z-test in R

Let's do an example in R. Since we're so familiar with it, let's use the `iris` dataset.

Suppose we have a friend who's a renowned flower researcher. This friend tells us that the population mean for irises is 4cm for petal lengths and 3cm for petal widths. The population variance for both is 4. This friend has collected the iris data and would like us to test if the sample means for petal length and petal width are different than the population mean.

Let's set this up!

Example *Z*-test in R

First: what is our null hypothesis? Alternative?

Petal Length

$H_0 : \mu_{length} = 4$

$H_A : \mu_{length} \neq 4$

Petal Width

$H_0 : \mu_{width} = 3$

$H_A : \mu_{length} \neq 3$

Example *Z*-test in R

What are our estimates and parameters?

Petal Length

N: `length(iris$Petal.Length)` = 150

$\bar{x}$: `mean(iris$Petal.Length)` = 3.76

$\mu_{length} : 4$

$\sigma : \sqrt{\sigma^2} = \sqrt{4} = 2$

Petal Width

N: `length(iris$Petal.Width)` = 150

$\bar{x}$: `mean(iris$Petal.Width)` = 1.2

$\mu_{length} : 3$

$\sigma : \sqrt{\sigma^2} = \sqrt{4} = 2$

> **Note**
>
> `length()` vs `nrow()`: `length()` returns the length of a vector, and `nrow()` returns the number of rows in a dataframe. For both, what might go wrong here?

Example *Z*-test in R

Let’s calculate these *z*-statistics.

Petal Length

```
samp_mean <- mean(iris$Petal.Length)
pop_mean <- 4
pop_sd <- sqrt(4)
samp_n <- sum(!is.na(iris$Petal.Length))

z_Petal.Length <- (samp_mean - pop_mean) /
  (pop_sd / sqrt(samp_n))
z_Petal.Length
```

```
[1] -1.481941
```

Petal Width

```
z_Petal.Width <- (mean(iris$Petal.Width) - 3) /
  (sqrt(4) / sqrt(sum(!is.na(iris$Petal.Length)))
z_Petal.Width
```

```
[1] -11.02679
```

Interpretation?

Z-Distribution in R

Since a standard normal distribution. and *z*-distribution are the same, we can use the `norm()` function family to examine the distribution.

Sampling from *z*-distribution

```r
rnorm(10)
```

```
 [1]  0.51594139  1.79543362
-0.54505247  1.76562800
-0.06737896 -1.39235310
 [7]  0.87729246  1.02919715
2.01662311 -0.80411997
```

Probability of a given *z*-value

```r
pnorm(c(-1.96, 1.96))
```

```
[1] 0.0249979 0.9750021
```

```r
pnorm(3)
```

```
[1] 0.9986501
```

```r
pnorm(3, lower.tail = F)
```

```
[1] 0.001349898
```

Note

`pnorm()` gives the probability at or less than the value supplied. `p`-values look at the probability of that or more extreme values, so we use `lower.tail=FALSE` to get a *p*-value for a *z*-value

z-value at a given probability

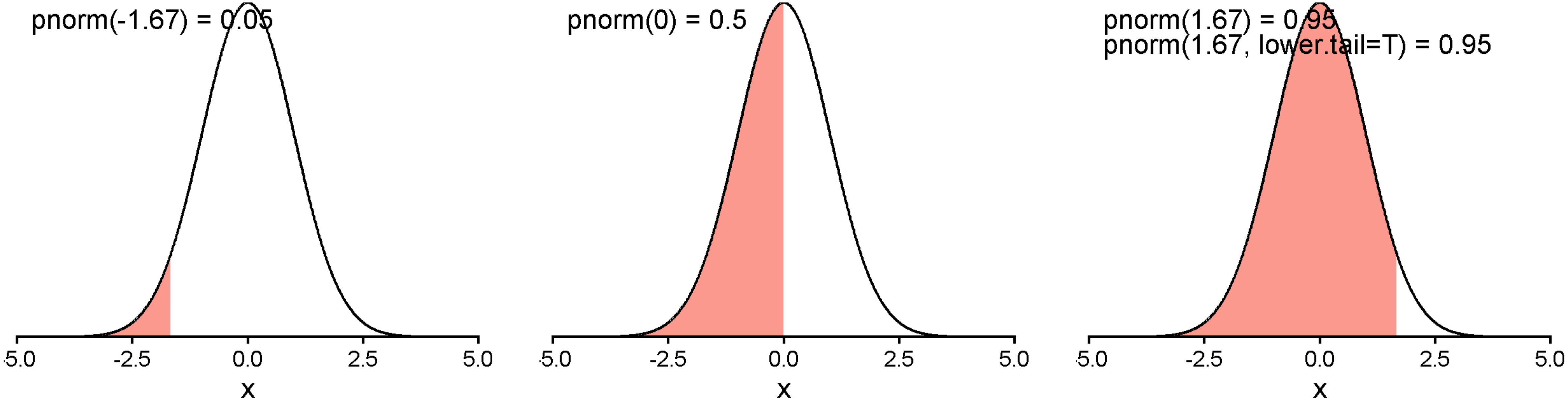
```r
qnorm(c(.025, .975))
```

```
[1] -1.959964  1.959964
```

```r
qnorm(.5)
```

```
[1] 0
```

Quick `lower.tail` Demo



Example *Z*-test in R

Let’s calculate these *z*-statistics.

Petal Length

```r
z_Petal.Length
```

```
[1] -1.481941
```

```r
pnorm(z_Petal.Length)
```

```
[1] 0.06917796
```

Petal Width

```r
z_Petal.Width
```

```
[1] -11.02679
```

```r
pnorm(z_Petal.Width)
```

```
[1] 1.419134e-28
```

Interpretation?

Working Directory Workshop

Working Directory Review

Our *working directory* is where R starts looking for files when we ask it to

In RMarkdown, `knitr` (which knits our document) overrides working directory changes within a chunk

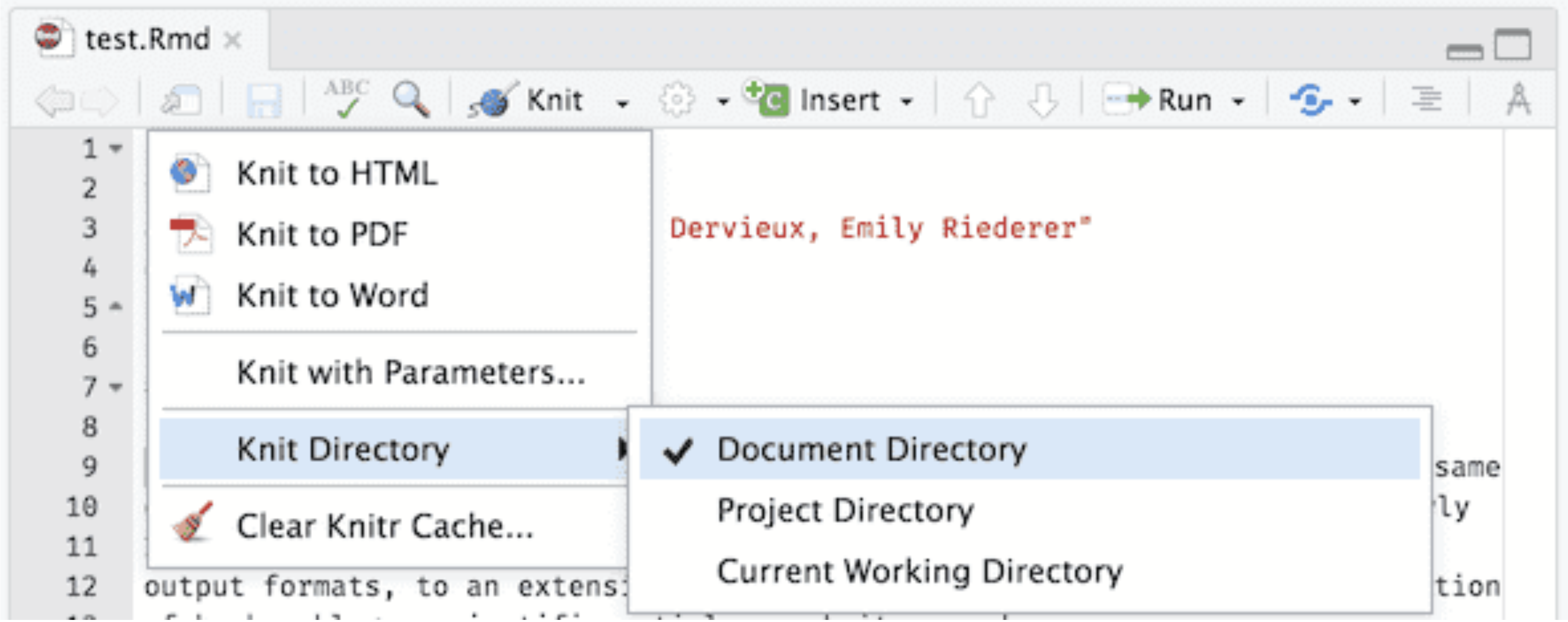
Relative paths let us specify paths to a file from a starting point folder

By default,

- RMarkdown's starting point is the document's folder
- `here::here()`'s starting point is the project's folder
- R script's starting point is the project's folder

Working Directory Review

You can change where RMarkdown knits to (this will also change where the .html output appears)



```
Dervieux, Emily Riederer"
same
ly
output formats, to an extens
tion
```

Working Directory Review

How would we specify paths to each of these files?

```
Documents/PSYC2020L
├── PSYC2020L.Rproj
├── Lab 1
│   ├── data1.csv
│   ├── script1.R
│   ├── rmdfile1.Rmd
├── Lab 2
│   ├── data2.csv
└───┴── rmdfile2.Rmd
```

`script1.R` to `data1.csv`

`"Lab 1/data1.csv"` or `here::here("Lab 1", "data1.csv")`

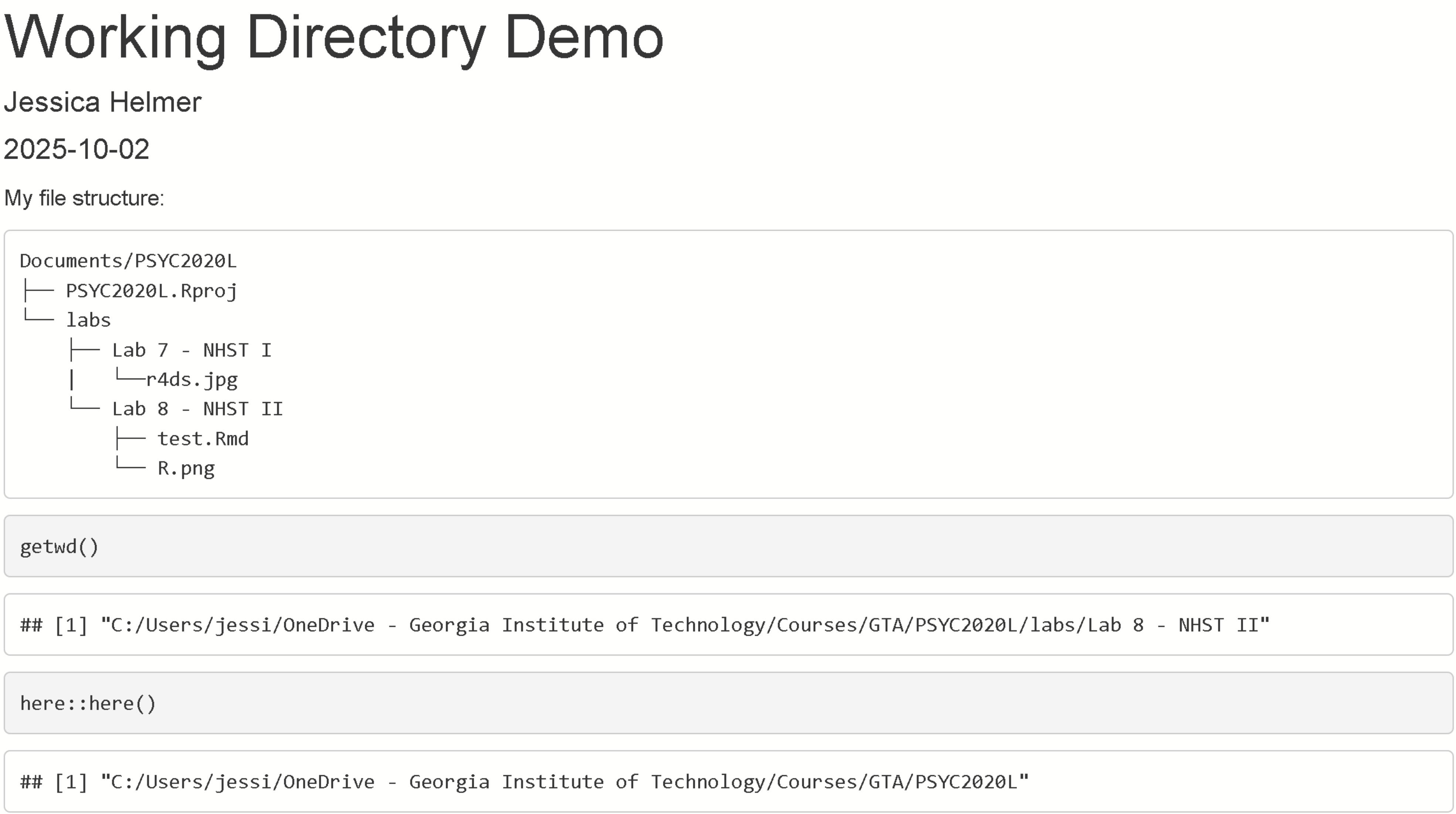
`rmdfile1.Rmd` to `data1.csv`

`"data1.csv"` or `here::here("Lab 1", "data1.csv")`

`rmdfile2.Rmd` to `data1.csv`

`"../data2.csv"` or `here::here("Lab 1", "data1.csv")`

Working Directory Review



Working Directory Demo

Jessica Helmer

2025-10-02

My file structure:

```
Documents/PSYC2020L
├── PSYC2020L.Rproj
└── labs
    ├── Lab 7 - NHST I
    |   └──r4ds.jpg
    └── Lab 8 - NHST II
        ├── test.Rmd
        └── R.png
```

```
getwd()
```

```
## [1] "C:/Users/jessi/OneDrive - Georgia Institute of Technology/Courses/GTA/PSYC2020L/labs/Lab 8 - NHST II"
```

```
here::here()
```

```
## [1] "C:/Users/jessi/OneDrive - Georgia Institute of Technology/Courses/GTA/PSYC2020L"
```

Working Directory Review

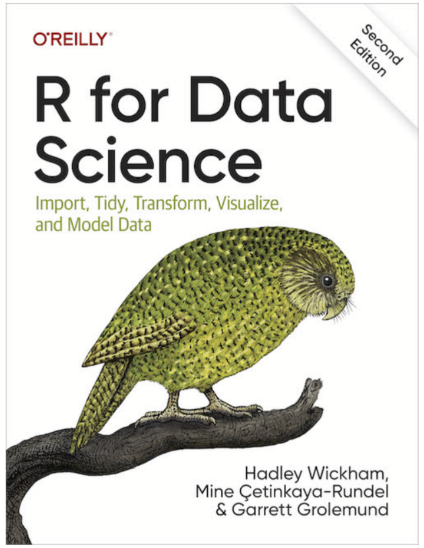
Getting `R` Picture

```
knitr::include_graphics("R.png")
```

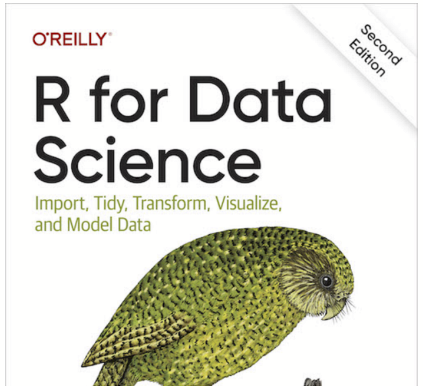
```
knitr::include_graphics(here::here("labs", "Lab 8 - NHST II", "R.png"))
```

Working Directory Review

```
knitr::include_graphics("../Lab 7 - NHST I/r4ds.jpg")
```


O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund

```
knitr::include_graphics(here::here("labs", "Lab 7 - NHST I", "r4ds.jpg"))
```



Working Directory Review

Let’s do this for real!

Download the `testdata.csv` file off Canvas

Make a test R script (could be regular R script if you follow along) and test RMarkdown document

Let’s make sure we can all read files in both formats

Assignment 8