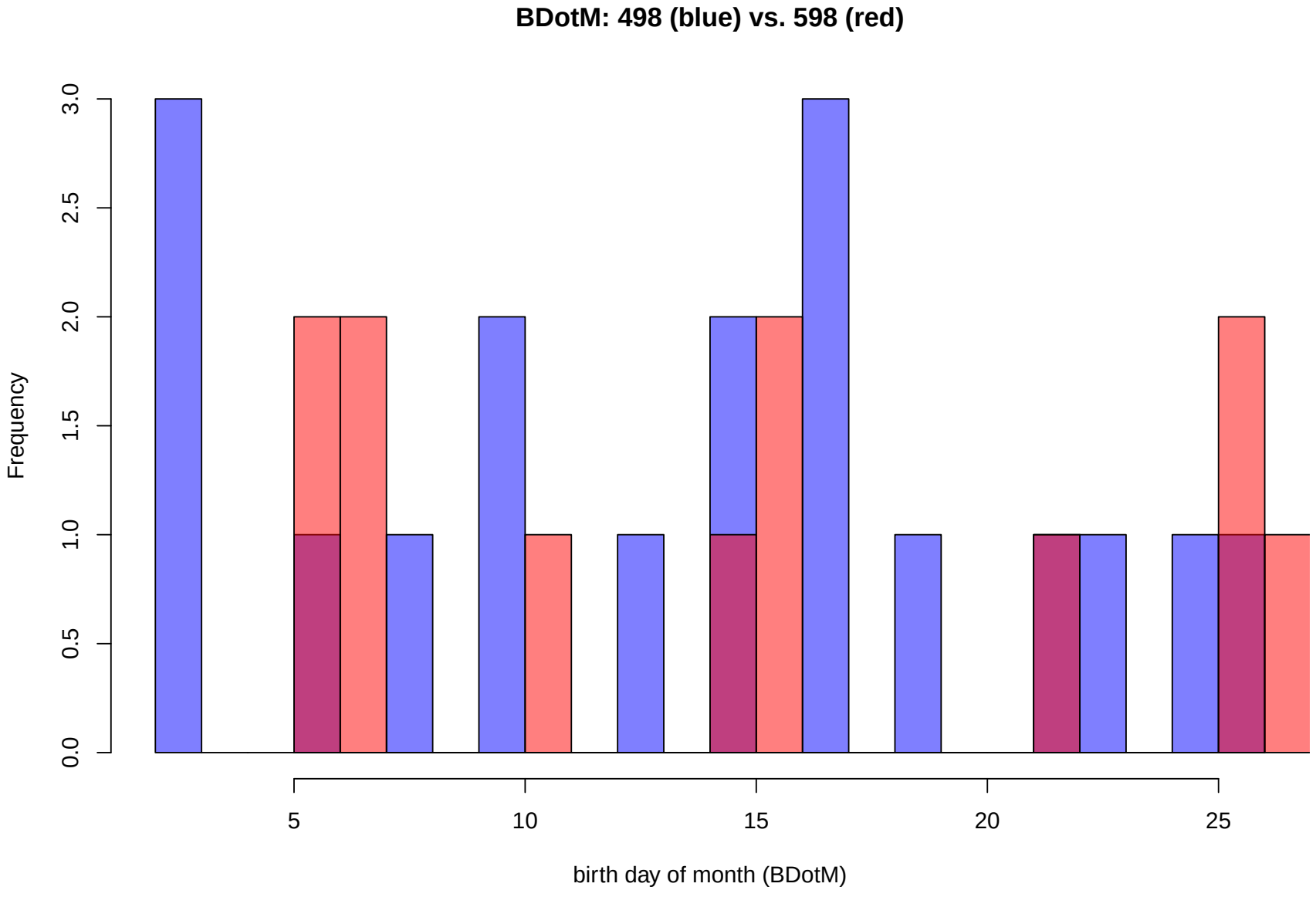
The Bootstrap and Distributions

BIOE 498/598 PJ

Spring 2021

Does BDotM differ for the BIOE 498 and BIOE 598 students?



Goal: Test if the difference between 498 and 598 BDotM's is signficant

```
mean(days498)
```

```
## [1] 13.88889
```

```
mean(days598)
```

```
## [1] 15.25
```

```
test_diff <- mean(days498) - mean(days598)
test_diff
```

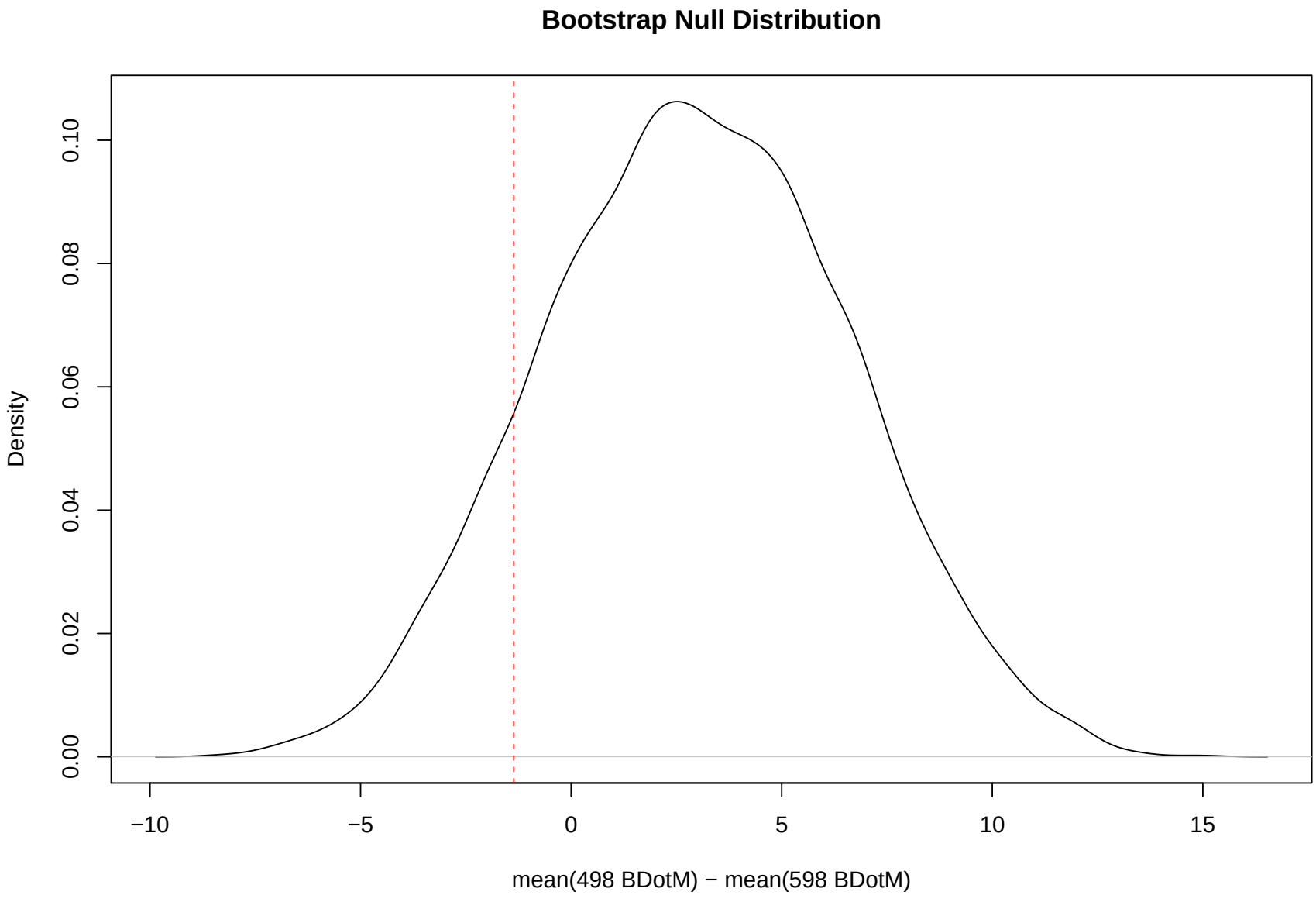
```
## [1] -1.361111
```

Creating a boostrap null distribution

```
# put all the BDotMs in one pool
days <- c(days498, days598)
n498 <- length(days498)
n598 <- length(days598)

diffs <- replicate(10000, {
  sample498 <- sample(n498, days, replace=TRUE)
  sample598 <- sample(n598, days, replace=TRUE)
  # return the difference between the two groups
  mean(sample498) - mean(sample598)
})
```

Creating a boostrap null distribution



Estimating the *p*-value

The *p*-value is the probability that a difference at least as large can be seen randomly.

We can estimate this probability as the fraction of bootstrap samples that are as large as the test difference.

```
mean(abs(diffs) >= abs(test_diff))
```

```
## [1] 0.7683
```

With a *p*-value this large, we cannot reject the null hypothesis that the BDotM is the same for both groups.

Shortcut method: the *t*-test

```
t.test(days498, days598, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  days498 and days598
## t = -0.4488, df = 22.194, p-value = 0.6579
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -7.647553  4.925331
## sample estimates:
## mean of x mean of y
##  13.88889  15.25000
```

Shortcut method: the *t*-test (equal variance)

```
t.test(days498, days598, alternative="two.sided",
       var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  days498 and days598
## t = -0.45729, df = 28, p-value = 0.651
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -7.458096  4.735874
## sample estimates:
## mean of x mean of y
##  13.88889  15.25000
```

Is the sampling distribution normal?

```
# Shapiro-Wilk in R is limited to 5000 points
shapiro.test(diffs[1:5000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diffs[1:5000]
## W = 0.99727, p-value = 7.522e-08
```

We reject the null hypothesis of normality.

Is the sampling distribution normal?