day_3_nov_1_2016_occupancy_ttest

Today I want to run some t-tests for 2 populations (occupied vs not occupied).

First I will create separate variables for when the office is occupied vs not.

```
occupied <- df[df$Occupancy == 1,]
not_occupied <- df[df$Occupancy == 0,]
summary(occupied)
```

```
##       date                        Temperature       Humidity
##  Min.   :2015-02-02 14:19:00   Min.   :20.29   Min.   :22.79
##  1st Qu.:2015-02-03 08:19:44   1st Qu.:21.70   1st Qu.:25.50
##  Median :2015-02-03 12:28:30   Median :22.68   Median :27.12
##  Mean   :2015-02-03 12:06:46   Mean   :22.39   Mean   :27.32
##  3rd Qu.:2015-02-03 16:59:14   3rd Qu.:23.10   3rd Qu.:28.50
##  Max.   :2015-02-04 10:43:00   Max.   :24.41   Max.   :31.47
##      Light             CO2         HumidityRatio        Occupancy
##  Min.   : 217.2   Min.   : 441.6   Min.   :0.003349   Min.   :1
##  1st Qu.: 433.0   1st Qu.: 860.2   1st Qu.:0.004252   1st Qu.:1
##  Median : 461.0   Median :1038.2   Median :0.004578   Median :1
##  Mean   : 499.6   Mean   :1014.5   Mean   :0.004591   Mean   :1
##  3rd Qu.: 538.0   3rd Qu.:1180.2   3rd Qu.:0.005044   3rd Qu.:1
##  Max.   :1697.2   Max.   :1402.2   Max.   :0.005378   Max.   :1
```

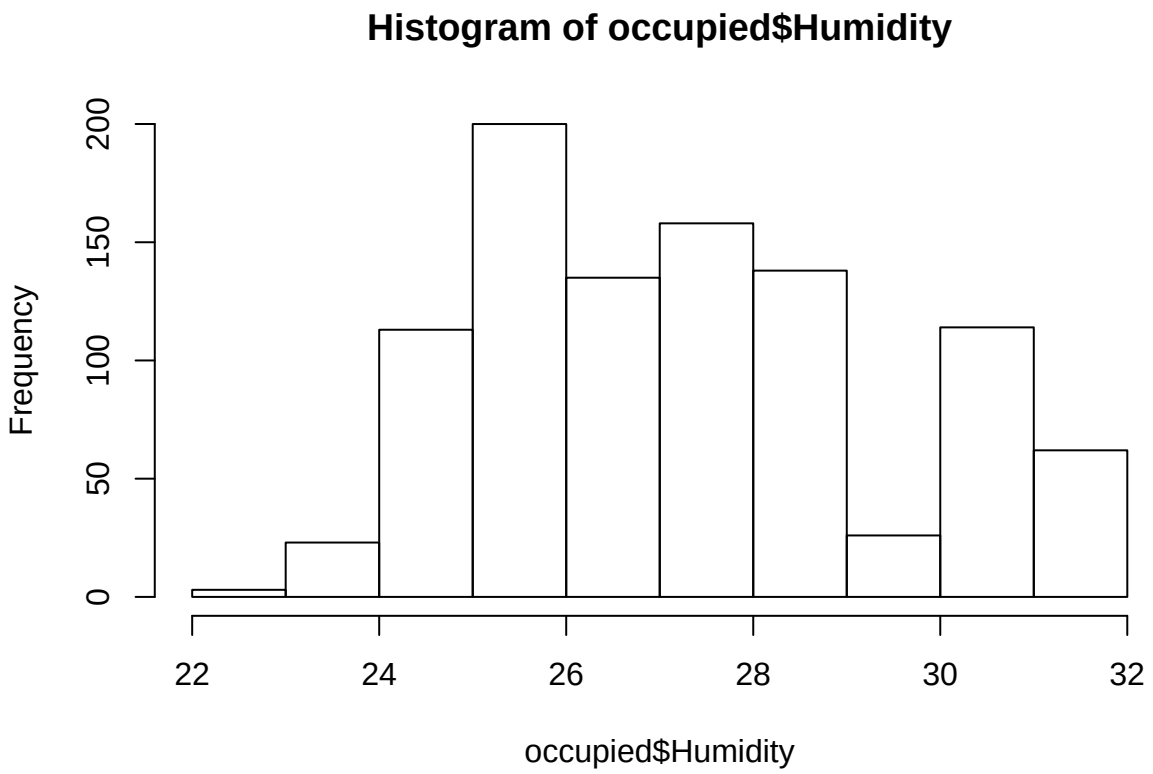
```
summary(not_occupied)
```

```
##       date                        Temperature       Humidity
##  Min.   :2015-02-02 17:34:00   Min.   :20.20   Min.   :22.10
##  1st Qu.:2015-02-03 00:45:00   1st Qu.:20.57   1st Qu.:22.55
##  Median :2015-02-03 13:12:00   Median :20.70   Median :24.20
##  Mean   :2015-02-03 12:44:54   Mean   :20.88   Mean   :24.23
##  3rd Qu.:2015-02-04 00:49:59   3rd Qu.:21.00   3rd Qu.:25.05
##  Max.   :2015-02-04 09:29:00   Max.   :23.29   Max.   :30.12
##      Light             CO2         HumidityRatio        Occupancy
##  Min.   :  0.00   Min.   : 427.5   Min.   :0.003303   Min.   :0
##  1st Qu.:  0.00   1st Qu.: 449.8   1st Qu.:0.003345   1st Qu.:0
##  Median :  0.00   Median : 479.2   Median :0.003661   Median :0
##  Mean   : 17.33   Mean   : 547.6   Mean   :0.003703   Mean   :0
##  3rd Qu.:  0.00   3rd Qu.: 576.5   3rd Qu.:0.003846   3rd Qu.:0
##  Max.   :638.00   Max.   :1205.2   Max.   :0.005114   Max.   :0
```

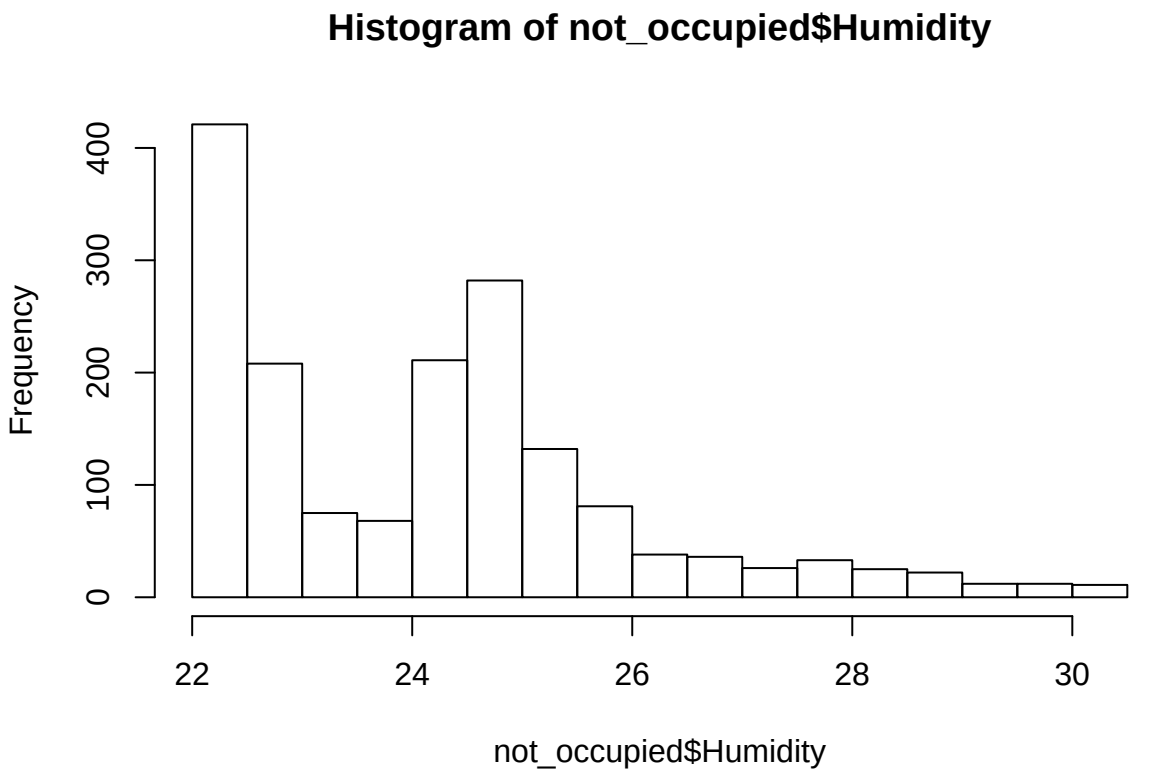
Humidity

We hypothesize that the humidity is different for occupied vs not occupied. Since I have no idea how humidity works, I will just run a 2-tailed test.

```
hist(occupied$Humidity)
```




```
hist(not_occupied$Humidity)
```



```r
t.test(occupied$Humidity, not_occupied$Humidity)
```

```
##
##  Welch Two Sample t-test
##
## data:  occupied$Humidity and not_occupied$Humidity
## t = 37.98, df = 1751.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.931757 3.251040
## sample estimates:
## mean of x mean of y
##  27.31782  24.22642
```

Just eye-balling the p-value, we can tell that it's a very small value. This means that we can confidently reject the null that the humidity is equal amongst occupied vs not occupied.

```r
library(lsr)
cohensD(occupied$Humidity, not_occupied$Humidity)
```

```
## [1] 1.601856
```

Wow I didn't know that cohen's D can go above 1. I usually just see values greater than 0.8.

It's important to remember that cohen's D doesn't actually have any statistical interpretation, just that the difference in the mean is a lot greater than the pooled standard deviation between the two samples.