Paired T-Test

Aaron Stearns

In this R markdown document I will perform a paired t-test on a set of student test scores. The 'pre-test' scores are student exam scores, and the 'post-test' scores are exam scores for the same group of students after having attended a test-prep course. The null hypothesis is that there is no difference in student test scores before and after having taken the test-prep course, and the alternate hypothesis is that there is a difference.

```
library(dplyr)

# Student test scores:
preTest <- c(25, 24, 26, 28, 27, 27, 28, 29, 24, 28, 31, 35, 37, 30)
postTest <- c(36, 38, 32, 33, 38, 43, 31, 32, 33, 38, 37, 36, 38, 40)

# Create data frame:
data <- as.data.frame(cbind(preTest, postTest))

data
```

```
##    preTest postTest
## 1       25       36
## 2       24       38
## 3       26       32
## 4       28       33
## 5       27       38
## 6       27       43
## 7       28       31
## 8       29       32
## 9       24       33
## 10      28       38
## 11      31       37
## 12      35       36
## 13      37       38
## 14      30       40
```

The formula to calculate the t statistic is as follows:

$$t = \frac{\sum D}{\sqrt{\frac{N\sum D^2 - \left(\sum D\right)^2}{N-1}}}$$

The sum of the differences (in this case between pre and post test scores) is divided by the square root of the sum of the differences squared times the number of observations (N), minus the square of the sum of differences, divided by N-1.

```
# Value for n:
n <- length(preTest)
```

```r
# Create column of differences between pre and post-test scores
data <- data %>%
          mutate(difference = preTest - postTest)

# Create squared differences column
data <- data %>%
          mutate(squaredDiff = difference^2)
```

Let's take a look at those new columns:

```r
data
```

```
##    preTest postTest difference squaredDiff
## 1       25       36        -11         121
## 2       24       38        -14         196
## 3       26       32         -6          36
## 4       28       33         -5          25
## 5       27       38        -11         121
## 6       27       43        -16         256
## 7       28       31         -3           9
## 8       29       32         -3           9
## 9       24       33         -9          81
## 10      28       38        -10         100
## 11      31       37         -6          36
## 12      35       36         -1           1
## 13      37       38         -1           1
## 14      30       40        -10         100
```

Now, I'll sum the differences and squared differences, and use those variables in the t-test formula:

```r
sumOfDifferences <- sum(data$difference)

sumOfSquaredDifferences <- sum(data$squaredDiff)

# Calculate t statistic
t <- sumOfDifferences /
        sqrt(((n * sumOfSquaredDifferences) - sumOfDifferences^2) / (n - 1))

# Value for t
t
```

```
## [1] -6.00403
```

I'll compare this with R's built in paired t-test function:

```r
t.test(preTest, postTest, paired = T)
```

```
##
##  Paired t-test
##
## data:  preTest and postTest
## t = -6.004, df = 13, p-value = 4.417e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.295778  -4.847079
## sample estimates:
## mean of the differences
```

```
##                -7.571429
```

And it can be seen that the t values are identical. The p-value is extremely low, and so we can arrive at the conclusion that there is a statistically significant difference between the groups, and that the test-prep course has been effective in raising student scores.