t-test

Virginia Price

7/29/2019

I want you to do a t-test on the pre and post data in order to answer the question, "Is the difference between pre and post data different than zero?" You will need to be careful about what type of data you have and how you might go about doing the ttest.

You can have 1,000,000 bonus points if you re-do the ttest using a linear regression.

Turns out t.test doesn't really work with two datasets of different sizes, so break 'em up, remove all them NA's, and put them into a sweet single dataset.

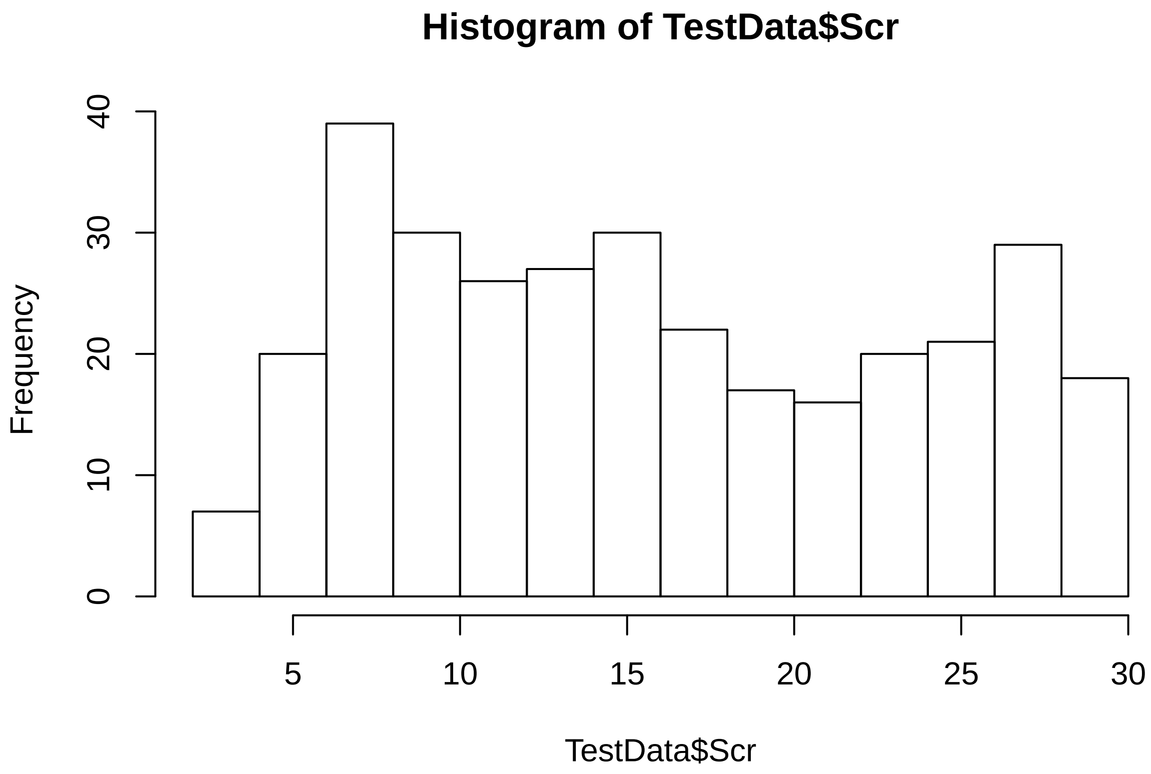
```
PreData <- TestData %>%
  select(ID,PrePost,Scr) %>%
  filter(PrePost == "Pre") %>%
  filter(Scr != 'NA') %>%
  rename(Scr.Pre = "Scr")


PostData <- TestData %>%
  select(ID,PrePost,Scr) %>%
  filter(PrePost == "Post") %>%
  filter(Scr != 'NA') %>%
  rename(Scr.Post = "Scr")

PrePostDf <- left_join(PreData,PostData,by="ID")
```

Can we do a t-test?

In order for our t-test to be valid, our data needs to be approximately normally distributed! We can check by quickly doing some histograms.

Check to see if data is normally distributed

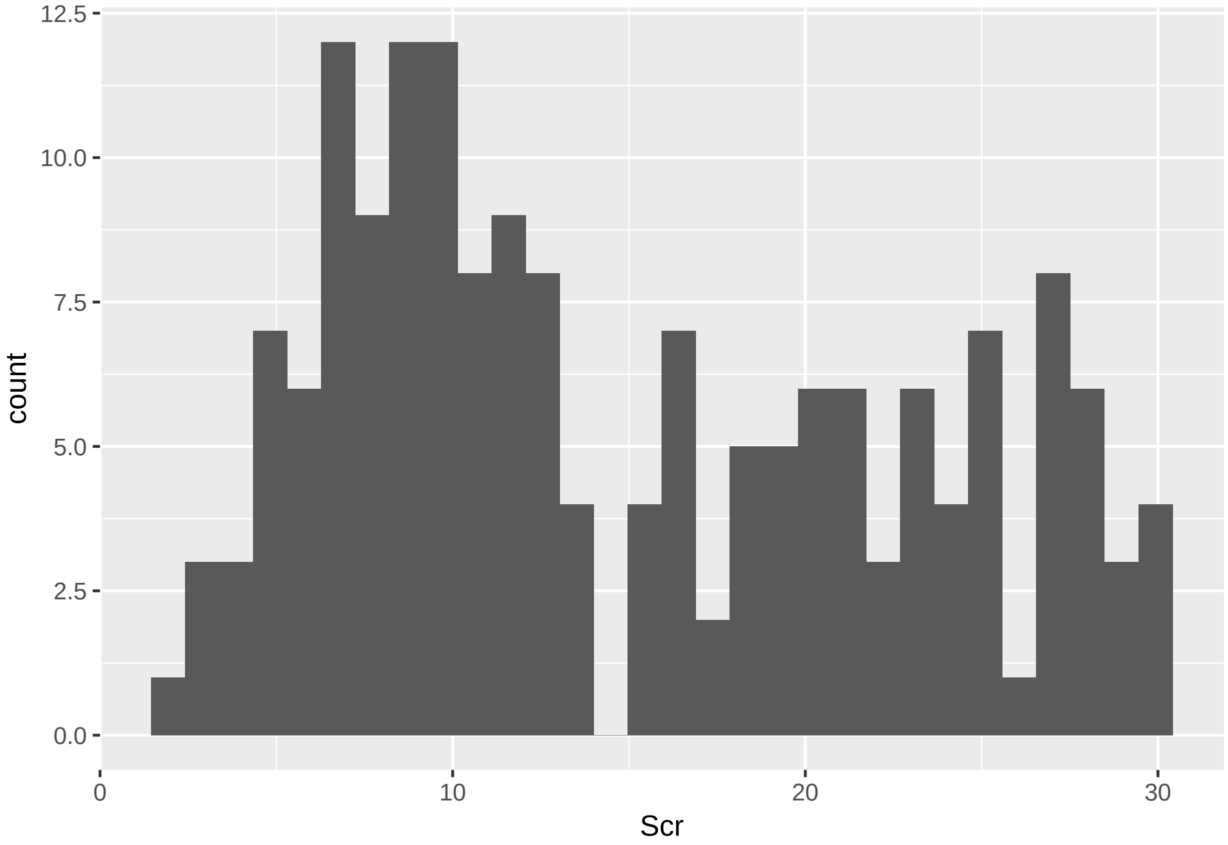
```
# all the pre/post scores
hist(TestData$Scr)
```



```r
# just the pre-test scores
TestData %>%
  filter(PrePost == "Pre") %>%
  ggplot(aes(x=Scr)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

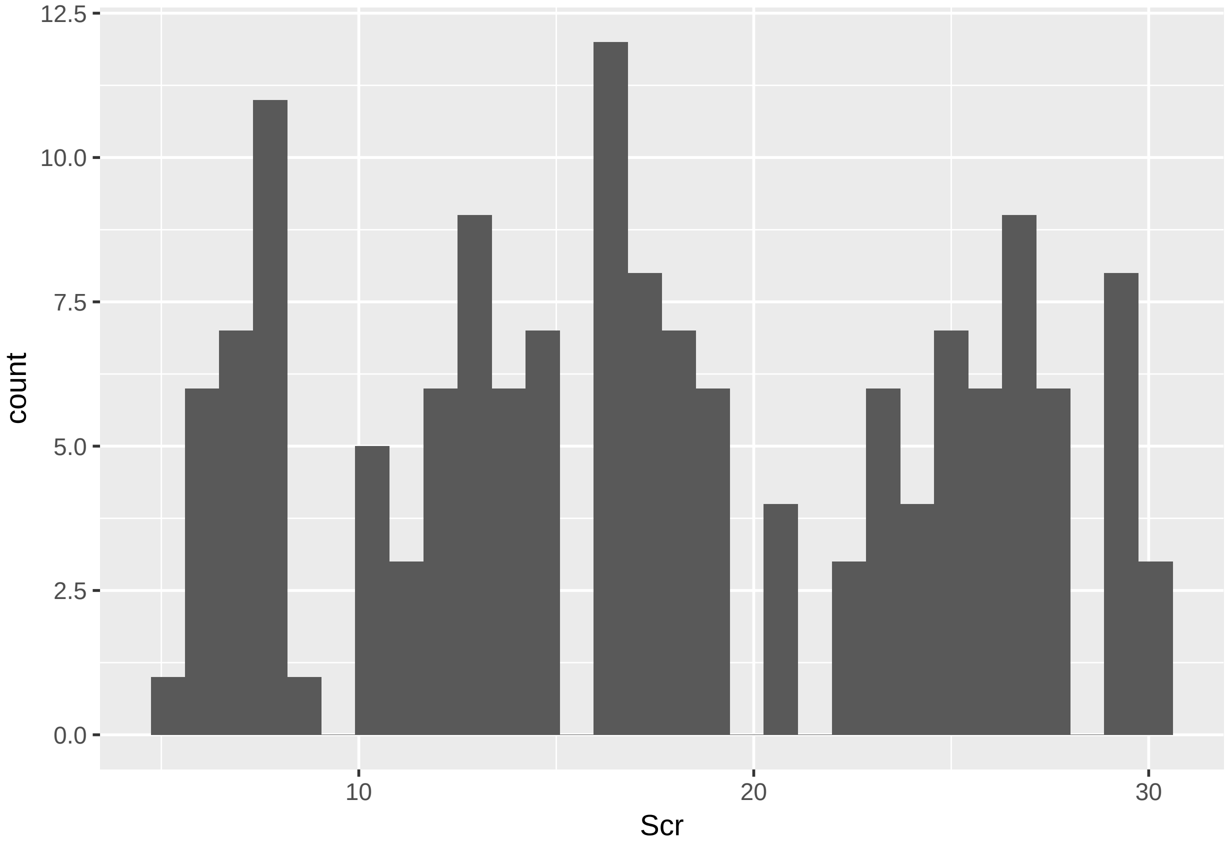
```
## Warning: Removed 7 rows containing non-finite values (stat_bin).
```



```
# just the post-test scores
TestData %>%
  filter(PrePost == "Post") %>%
  ggplot(aes(x=Scr)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



They're SUPER not normally distributed, dang. AH WELL, FUCK IT, WHO CARES.

(We do. We care. Otherwise you have bad science.)

Given that our data are not normally distributed, under what conditions can we assume our t-test is valid? Enter: the *Wilcoxon/Mann-Whitney* test!

```
wilcox.test(PrePostDf$Scr.Pre, PrePostDf$Scr.Post, paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  PrePostDf$Scr.Pre and PrePostDf$Scr.Post
## V = 1035, p-value = 1.676e-09
## alternative hypothesis: true location shift is not equal to 0
```

Since V is big ($V = 1035$, $p < 1.68 \times 10^{-9}$), it's reasonable to use a t-test. Because math.

So let's do it!

Run the t-test

```
t.test(PrePostDf$Scr.Pre, PrePostDf$Scr.Post,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  PrePostDf$Scr.Pre and PrePostDf$Scr.Post
## t = -5.7225, df = 126, p-value = 7.251e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.005679 -1.947077
## sample estimates:
## mean of the differences
##               -2.976378
```

I tried to use the t-test Navarro uses, but it keeps giving me an "oucome variable must be numeric" error

```
TD_tt <- TestData %>%
  filter(Scr != 'NA')

TD_tt$Scr <- as.numeric(TD_tt$Scr)

pairedSamplesTTest(
  formula = Scr ~ PrePost,
  data = TD_tt,
  id = "ID")
```

Wooo, looks like there is a difference. On average, students scored higher on the post-test. Yay, improvement!

Doing