Of course, there is a way to get RStudio to do it all for you. Just leave off the part at the end that requests the F-statistic:

```
anova( lm( Pulse ~ Award, data=StudentSurvey ) )
```

```
## Analysis of Variance Table
##
## Response: Pulse
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Award       2   2047 1023.62  7.1039 0.0009425 ***
## Residuals 359  51729  144.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Start here Fri., Nov. 19

Example: Textbook Costs

The Lock 5 dataset **TextbookCosts** contains cost information for textbooks coming from four different disciplies.

```
head(TextbookCosts)
```

Social Science denoted by 1
Nat. " " " 2
Humanities " " 3
Arts " " 4

```
##             Field Books Cost
## 1  SocialScience     3   77
## 2 NaturalScience     2  231
## 3 NaturalScience     1  189
## 4  SocialScience     6   85
## 5 NaturalScience     1  113
## 6     Humanities     9  132
```

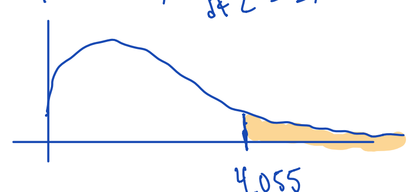
$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

In all 10 courses from each of the separate disciplines was sampled. In R we have seen the following command can be used to generate the ANOVA table.

```
anova( lm( Cost ~ Field, data=TextbookCosts ) )
```

```
## Analysis of Variance Table
##
## Response: Cost
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Field      3  30848 10282.6  4.0547 0.01397 *
## Residuals 36  91294  2535.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Theoretical F dist. w/ df1 = 3, df2 = 36



Since there are four different `Fields` represented, we should not be surprised that $\mathrm{df}_1 = 4 - 1 = 3$. And, with 10 courses per field, the full dataset has 40 cases, which explains why $\mathrm{df}_2 = 40 - 4 = 36$. Were we to do the other calculations, SSG, SSE, MSG, MSE and F by hand, they would match what appears in the output above. (Can you locate each of those?) The only number we might be more cautious to believe is the P-value. This R command always displays a P-value taken from a theoretical F-distribution, in this case the same result we would obtain using the command

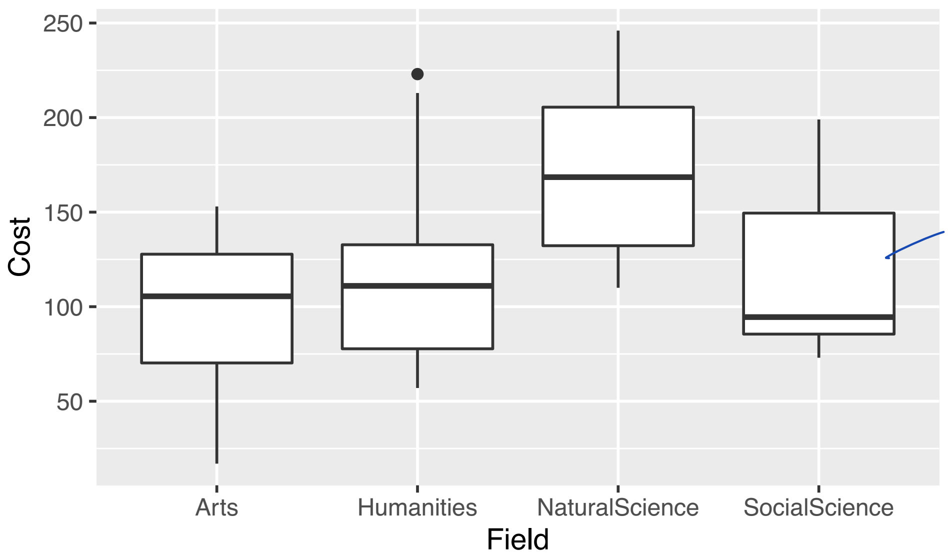
```
1 - pf(4.055, df1 = 3, df2 = 36)
```

```
## [1] 0.01396561
```

But is it reasonable to obtain our P-value this way? Look back at the conditions stated near the start of the last section. Are they met?

- The textbook (pp. 515-516) tells us the samples of courses were all taken at the same college, nothing more. So these probably cannot be considered random samples from the populations of all Arts (resp. Humanities, NaturalScience, SocialScience) courses throughout the country, but perhaps they can for the courses in those disciplines at this college. It is likely reasonable to assume that book prices and samples, within the more limited scope of the one college, are independent.
- We can look at plots of `Cost` broken down by `Field` in an attempt to verify normality, but there are so few data points, it is difficult to get any degree of surety from the data itself. (Perhaps from past experience?) In looking at side-by-side boxplots such as those displayed here, the textbook (p. 516) declares "All four samples are relatively symmetric, have no outliers, and appear to have about the same variability," words used to justify that we are "close enough" on this condition. Do you agree?

```
gf_boxplot( Cost ~ Field, data=TextbookCosts )
```



10 pieces of data go into this (and each) boxplot

- We look at the various sample standard deviations

```
sd( Cost ~ Field, data=TextbookCosts )
```

```
##          Arts    Humanities NaturalScience  SocialScience
##      44.94738      58.14551       48.49238       48.89910
```

largest

(Note that `favstats()` could also have been used here, but gives extra information we do not need right now.) The ratio of largest-sd-to-smallest is

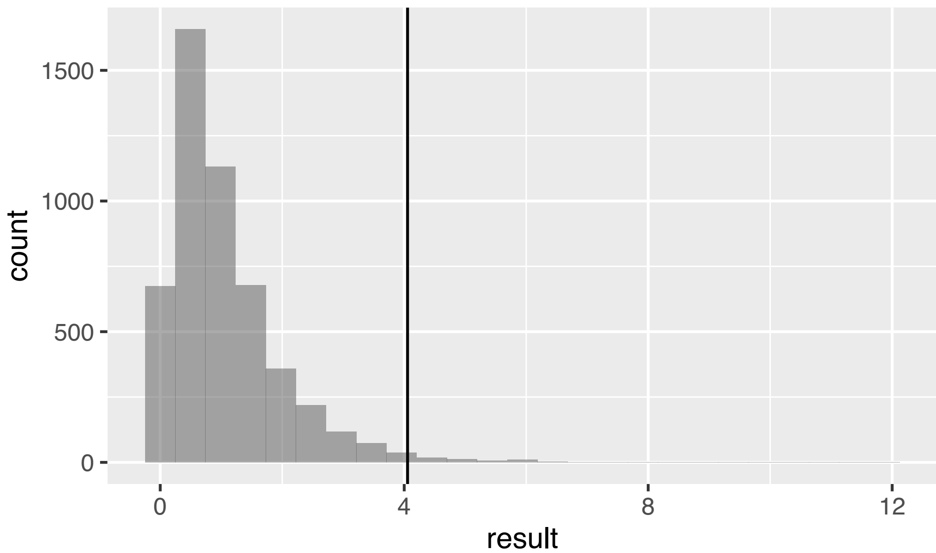
$$\frac{58.1455}{44.9474} \doteq 1.294,$$

Smallest group sd

well below 2.

Nevertheless, if we feel unsure, we can employ randomization to find an approximate P-value instead.

```
manyFs <- do(5000) * anova( lm( Cost ~ shuffle(Field), data=TextbookCosts ) )$F[1]
gf_histogram( ~result, data=manyFs ) %>% gf_vline( xintercept= ~ 4.055)
```




```
nrow( filter( manyFs, result > 4.054) ) / 5000
```

```
## [1] 0.0126
```

This P-value is quite similar to the one arising from the theoretical F distribution.

You rejected the null hypothesis, what now?

Recall that the null hypothesis is

$$\mathbf{H}_0: \ \mu_1 = \mu_2 = \cdots = \mu_k,$$

and if we rejected it, it is in favor of the alternative, that at least two population means are different. The natural follow-up queston is, "which ones?" The cautions discussed in Section 8.2, beginning with "Lots of Pairwise Comparisons", mirror those discussed in Section 4.5, p. 289, "The Problem of Multiple Testing." It is right for us to conduct the blanket test of 1-way ANOVA before charging into pairwise comparisons, but even after we have decided the null hypotheis above is to be rejected, we should proceed sensibly.

R offers a sensible approach to pairwise comparisons in the `TukeyHSD()` command. We apply it (note it uses another command, `aov()`, as an intermediary) to the textbook data above.

```
TukeyHSD( aov( Cost ~ Field, data=TextbookCosts ) )
```

Tukey Honest Significant Differences

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Cost ~ Field, data = TextbookCosts)
##
## $Field
##                                diff        lwr        upr     p adj
## Humanities-Arts                25.7  -34.95384  86.353844 0.6669143
## NaturalScience-Arts            76.2   15.54616 136.853844 0.0090147
## SocialScience-Arts             23.7  -36.95384  84.353844 0.7201024
## NaturalScience-Humanities      50.5  -10.15384 111.153844 0.1312366
## SocialScience-Humanities       -2.0  -62.65384  58.653844 0.9997441
## SocialScience-NaturalScience -52.5 -113.15384   8.153844 0.1097759
```

This line is about $\mu_3-\mu_4$

Note: Read Lock 8.2 as if the message is important, but the details are not.