Open up three browser tabs, one each pointing at

StatKey: `https://www.lock5stat.com/StatKey/`
ANOVA shiny app: `https://shiny.calvin.edu:3838/scofield/fstatRandomizationDist/`
RStudio: `https://rstudio.calvin.edu:8787/`

1. If you want to use a theoretical F-distribution to go from test statistic to P-value, there are two things you look to be true:
   - the sampling distributions for sample means are all normal, and
   - the population standard deviations are the same number for all populations from which your samples are drawn.

   These things are difficult to check. As before, the first of these might be true even without sample sizes all being at least 30, but that rule of thumb pretty well seals it. As for the second, we presume we are close enough if the ratio of largest *sample* standard deviation to smallest sample standard deviation is no bigger than 2.

   for 1st bullet

   for 2nd bullet

   In StatKey, select "ANOVA for Difference in Means". Then determine if both, only one, or neither of these rules of thumb are met for the following data sets available from the drop-down menu:

   (a) **SandwichAnts**
   (b) **StudentSurvey** (`Pulse` and `Award`)
   (c) **FishGills3** (`GillRate` and `Calcium`)

   Moving over to RStudio, make sure you could use `favstats()`, as in

   ```
   favstats(Ants ~ Filling, data=SandwichAnts)
   ```

   to decide about these same rules of thumb. Try it again with all three data sets.

2. In StatKey, generate a randomization distribution for the F-statistic working with the **StudentSurvey** (`Pulse` and `Award`) data. Determine the test statistitc and the approximate P-value using this randomization distribution. What hypotheses are we testing, and what conclusion are we led to?

3. In the ANOVA shiny app, load the same data as above—that is, **StudentSurvey** (`Pulse` and `Award`). You will have to type these variable names to indicate your selection of explanatory and response variables. Then navigate through the tabs ("Side-by-side plots", "Randomization dist", and "Summaries", particularly) in order to find

   (a) where information is reported that helps you decide about the rules of thumb.
   (b) where the F-statistic is reported.
   (c) where you can learn the P-value as estimated via randomization.
   (d) where you can learn the P-value as estimated from a theoretical F-distribution.

4. In StatKey, select the theoretical F-distribution with "Numerator df" (also known as "df groups", or $df1$) equal to 4 and "Denominator df" (also known as "df residuals/error", or $df2$) equal to 53.

(a) What is the rejection region that corresponds to $\alpha = 0.05$? Answer: $(2.546, \infty)$

(b) Corresponding to these degrees of freedom, from how many populations have we sampled? $4+1 = 5$

(c) Corresponding to these degrees of freedom, what is our *overall* sample size $n = n_1 + n_2 + \cdots + n_I$? $df_1 + df_2 + 1 = 4 + 53 + 1 = 58$

(d) In RStudio, the command one uses to answer the rejection-region question above would be

```
qf(0.95, df1=4, df2=53)
```

If, for these same degrees of freedom, you had the test statistic $F = 1.9735$, what R command would you use to find the corresponding P-value?

5. Using the shiny app, "import-from-url" the data set found at `https://www.zoology.ubc.ca/~whitlock/ABD/teaching/datasets/15/15q01PlantPopulationPersistence.csv`. Select reasonable options for the explanatory and response variables.

(a) What is the test statistic for this data?

(b) Does it appear the rules of thumb are met by this data so that the sampling distribution for the F-statistic is well modeled by a theoretical F-distribution?

(c) If we made use of a theoretical F-distribution, which one should it be? Can you tell the values of $df1$ and $df2$ simply viewing the "Data set" tab?

(d) What is the P-value arising from the randomization distribution? What is the P-value we would get from the theoretical F-distribution? Are the two numbers noticeably different?

6. Import the same data set as in the last problem into RStudio and then view it using commands

```
ppp <- read.csv(
  "https://www.zoology.ubc.ca/~whitlock/ABD/teaching/datasets/15/15q01PlantPopulationPersistence.c
head(ppp)
```

The command

```
anova( lm( generations ~ treatment, data=ppp ) )
```

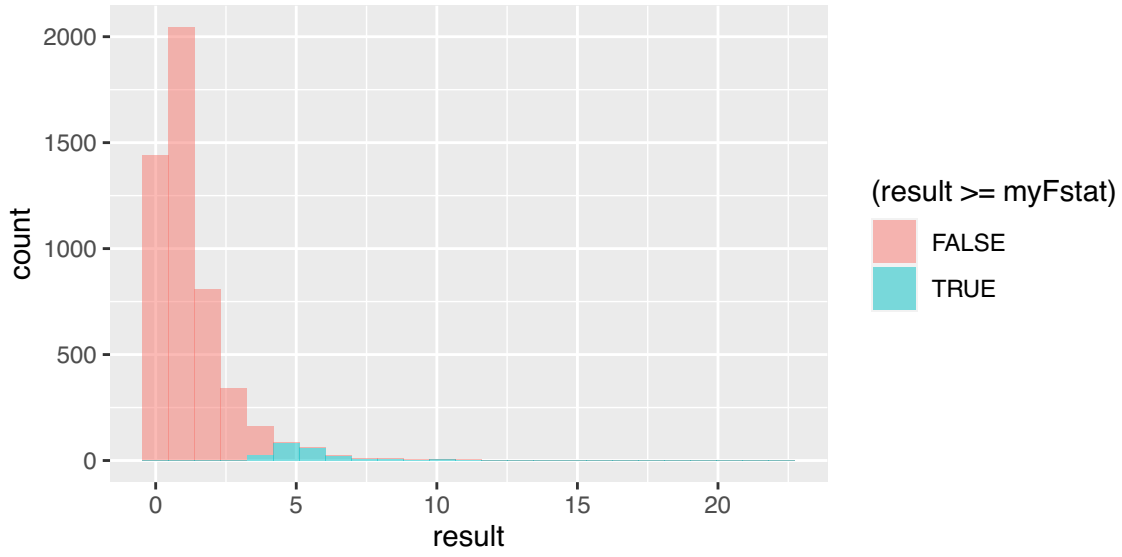
works a bit like the other all-in-one-step commands such as `prop.test()`, `t.test()`, and `chisq.test()`. It produces the test statistic, then computes a P-value based on a theoretical F-distribution model, *whether or not such a model is warranted*.

As we already decided above to be wary of the theoretical F-distribution model, because of small sample sizes, we may not want to be tempted into using information provided by the above command. We may prefer to use randomization. Doing so still requires us to learn the test statistic, achieved through the command (compare with the one given above)

```
myFstat <- anova(lm(generations ~ treatment, data=ppp))["F value"][1,1]
myFstat

[1] 3.996047
```

and then requires that we generate a bunch of randomization statistics by shuffling the values of one of the variables:

```
manyFs <- do(5000) *
  anova(lm(generations ~ shuffle(treatment), data=ppp))["F value"][1,1]
head(manyFs)

     result
1 0.2410901
2 0.6292906
3 0.4855876
4 0.1540041
5 0.4855876
6 2.7658863

gf_histogram(~result, data=manyFs, fill=~(result >= myFstat))
```



What R command could you use as follow-up to the above in order to ascertain an approximate *P*-value via randomization?