Test for Association Between a Categorical and Quantitative Variable

Thomas Scofield

November 16, 2021

Idea of ANOVA

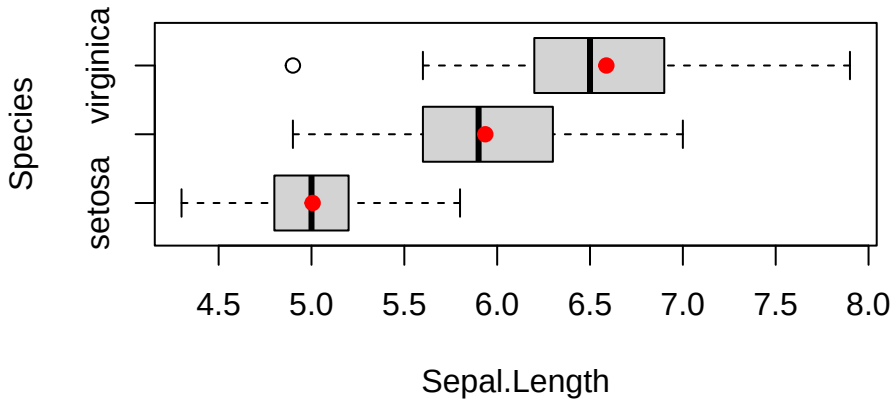
Consider the `iris` data set, for which a partial data table is given above. When we look `Sepal.Length` broken down by `Species`, we see that the samples have different means:

```
irisMeans <- mean(Sepal.Length ~ Species, data=iris)
irisMeans
```

```
##     setosa versicolor  virginica
##      5.006      5.936      6.588
```

Different sample means do not always lead to the conclusion that there are different population means. We might look at side-by-side boxplots to assist our intuition about whether the population means are different. What follows is a standard boxplot (the lines through the interior of the boxes give the locations of the three sample medians) enhanced by the incluion of a solid red dot at the locations of the group means.

```
boxplot(Sepal.Length ~ Species, data=iris, horizontal=TRUE)
points(c(1,2,3) ~ value(irisMeans), pch=19, col="red")
```

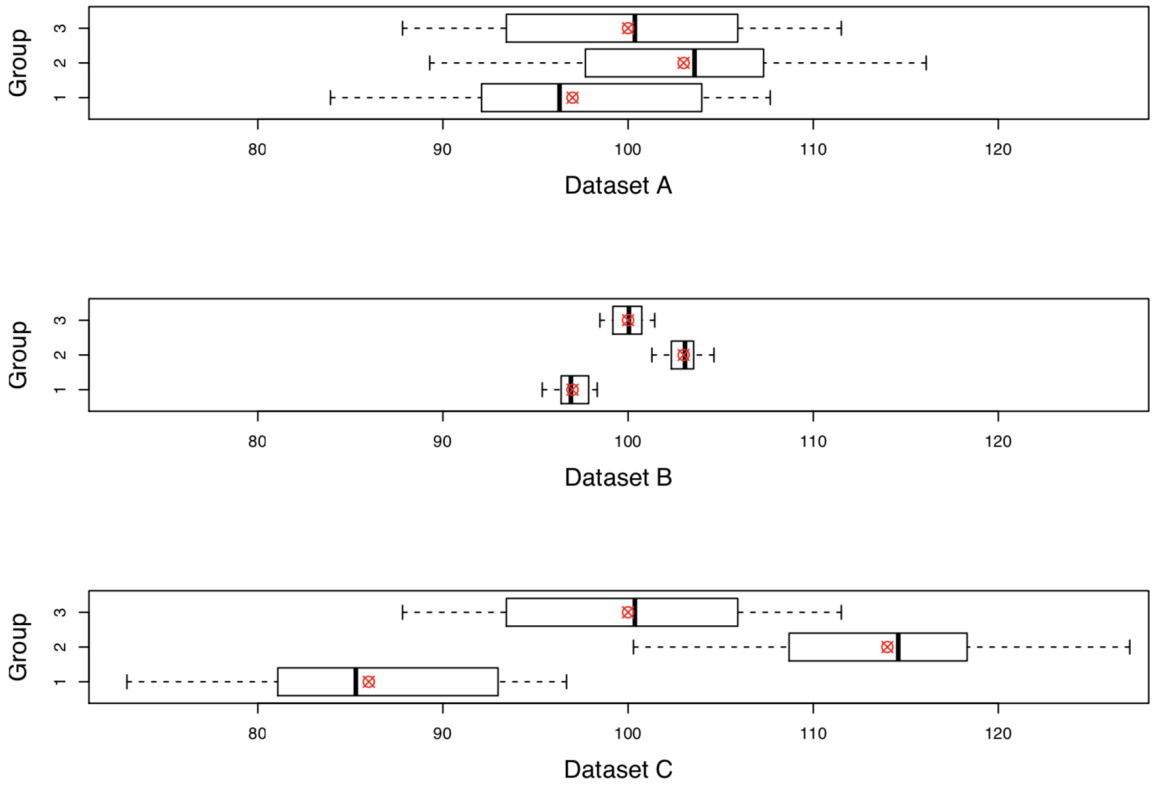


How does this plot add or detract from the evidence that the population means are not all the same across the three species? What is our basis for comparison? Consider this figure from the Lock 5 text. In particular, start by focusing on Dataset A, the first set of side-by-side boxplots. As with my boxplots above, there are boxplots for each of three "species", enhanced with a red mark is placed at the location of the sample means. In truth, this is made-up data to illustrate a point. We see the three sample means are not all the same (not vertically aligned), but are they far enough apart to provide convincing evidence that the corresponding population means are not all the same?

But compare with the three boxplots of Dataset B. Those have means in the exact same locations as the three means of Dataset A, yet somehow, it seems like the boxplots for Dataset B are more convincing of a difference in underlying population means than the boxplots for Dataset A. Why? Because the boxplots

rfor the three species in Dataset A are more spread out, overlapping more than occurs with the boxplots for Dataset B.

Now compare the three boxplots for Dataset A with those for Dataset C. The boxplot for Group 1 in Dataset C has exactly the same *range* (length between ends of the whiskers) and *IQR* (width of box) as for Group 1 in Dataset A. The same sort of consistency in range and IQR has been maintained for Groups 2 and 3 (only speaking about Datasets A and C here). Yet it seems that the boxplots for Dataset C are more convincing as evidence that group population means are different than the boxplots of A.



In this module we consider bivariate data where the explanatory variable is categorical with k distinct values (but, unlike the 2-sample mean tests of Chapter 6, we do not require $k = 2$), and the response variable is quantitative. For the iris data, the categorical variable is `Species` and has three values (corresponding to the three boxplots), and the response variable is `Sepal.Length`. Our main focus is a test with null hypothesis that all population means are equal

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

and the alternative hypothesis is that at least one of the population means is different from another.

Given the intuition arising from the side-by-side boxplots above, we must

- have a dataset which is comprised of independent random samples from each group,
- produce a test statistic that takes into account both
- how different the individual means are, and
- the spread of individual samples.

This is the idea of 1-way ANOVA (ANalysis Of VAriance).

The F-statistic

First, some symbol definitions. In our dataset we have k different groups/species/populations, and the categorical explanatory variable identifies the group to which each case belongs. We can take group sample means of the quatitative response fariable, means we refer to as $\overline{x}_i$. That is, $\overline{x}_1$ is the average response value for cases in Group 1, $\overline{x}_2$ is the average response value for cases in Group 2, and so on. In R, a command like

```
mean( responseVar ~ explanatoryVar, data=dataset )
```

would give us $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$ all at once.

We could also treat the dataset as a single whole, not distinguishing between individual groups/populations, and calculate the grand mean $\overline{x}$. The corresponding command might be something like

```
mean( ~ responseVar, data=dataset )
```

Along with symbols $\overline{x}, \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$ we use n_i to denote the size of the sample from Group i, and s_i to denote the standard deviation of the sample from Group i. (This sort of notation was employed in 2-sample mean problems found in Section 6.12 of the Lock 5 text.)

After viewing some side-by-side boxplots in the previous section, we stated some criteria for a test statistic. The quantities we consider next are for the purpose of constructing a test statistic that meets these criteria. First, we have the quantity SSG that helps to measure *between-group-variability*:

$$SSG \;=\; \sum n_i(\overline{x}_i - \overline{x})^2.$$

Here the SS stands for "sum of squares". Another quantity, SSE,

$$SSE \;=\; \sum (x - \overline{x}_i)^2,$$

helps to measure *within-group-variability*. Yet another sum of squares, SST,

$$SST \;=\; \sum (x - \overline{x})^2,$$

helps to measure *total variability*.

Example. Say the sampled values for Groups 1, 2 and 3 are as follows:

- Group 1: 15, 18, 17
- Group 2: 14, 11, 13
- Group 3: 16, 17, 19, 17

Starting with the grand mean, $\overline{x}$, inclusive of all response values in every group:

$$\overline{x} \;=\; \frac{1}{10}(15+18+17+14+11+13+16+17+19+17) \;=\; 15.7.$$

Next, the mean $\overline{x}_1$ for Group 1 is

$$\overline{x}_1 \;=\; \frac{1}{3}(15+18+17) \;=\; 16.667,$$

with the result that Group 1 contributes

$$(15-16.667)^2+(18-16.667)^2+(17-16.667)^2 = 4.667 \text{ to } SSE \qquad \text{and} \qquad 3(16.667-15.7)^2 \;=\; 2.803 \text{ to } SSG.$$

Similarly, for Group 2,

$$\overline{x}_2 \;=\; \frac{1}{3}(14+11+13) \;=\; 12.667,$$

so it contributes

$(14-12.667)^2+(11-12.667)^2+(13-12.667)^2 = 4.667$ to SSE and $3(12.667-15.7)^2 = 27.603$ to SSG.

Finally, for Group 3,

$$\overline{x}_3 = \frac{1}{4}(16+17+19+17) = 17.25.$$

so Group 3 contributes

$(16-17.25)^2+(17-17.25)^2+(19-17.25)^2+(17-17.25)^2 = 4.75$ to SSE and $4(17.25-15.7)^2 = 9.61$ to SSG.

We get SSE and SSG by summing the individual contributions to each:

$$SSE = 4.667+4.667+4.75 = 14.084, \quad \text{and}$$
$$SSG = 2.803+27.603+9.61 = 40.016.$$

The total sum-of-squares is

$$SST = (15-15.7)^2+(18-15.7)^2+(17-15.7)^2+(14-15.7)^2+(11-15.7)^2+(13-15.7)^2$$
$$+(16-15.7)^2+(17-15.7)^2+(19-15.7)^2+(17-15.7)^2 = 54.1.$$

Notice that

$$SST = 54.1 \quad \text{and} \quad SSG+SSE = 40.016+14.084 = 54.096$$

are nearly equal, and would be exactly the same but for rounding off during computation. This is an important fact:

$$SST = SSG + SSE.$$

Our test statistic will be a ratio of between-group-variability and within-group-variability. However, a simple ratio of SSG to SSE would not quite be comparing apples to apples, so to speak. When, in Chapter 2, the Locks introduced the formula for sample *variance*, it was

$$\frac{1}{n-1}\sum(x-\overline{x})^2.$$

The variance contains a sum-of-squares, tempered (divided) by its degrees of freedom. Similarly, we consider a truer measure of variability between groups to be

$$MSG = \frac{SSG}{k-1},$$

and the measure of variability within groups to be

$$MSE = \frac{SSE}{n-k}.$$

The test statistic, called F, is the ratio

$$F = \frac{MSG}{MSE}.$$

The results of the various calculations are usually arranged in an **ANOVA table**.

Source	df	SS	MS	F-stat	P-value
Groups/Factors	$\text{df}_1 = k-1$	$SSG = \sum n_i(\overline{x}_i-\overline{x})^2$	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	P
Residuals/Errors	$\text{df}_2 = n-k$	$SSE = \sum(x-\overline{x}_i)^2$	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	$SST = \sum(x-\overline{x})^2$			

Exercise

Try your hand at complete a partial 1-way ANOVA table at this richland.edu link

Computations in R

The computations above are tedious to do by hand, particularly when sample sizes are larger. R can be made to do them for you. We get an ANOVA table involving response variable `Sepal.Length` and explanatory variable `Species` in the `iris` dataset by entering

```
anova( lm( Sepal.Length ~ Species, data=iris ) )
```

```
## Analysis of Variance Table
##
## Response: Sepal.Length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Species     2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals 147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that it only gives the first two rows, not the last ("Total") row.

With a couple extra keystrokes at the end, we can home in on the entries in the `df` column

```
anova( lm( Sepal.Length ~ Species, data=iris ) )$D
```

```
## [1]   2 147
```

the `MS` column

```
anova( lm( Sepal.Length ~ Species, data=iris ) )$M
```

```
## [1] 31.6060667  0.2650082
```

or the `F` column (try it both with and without the "[1]")

```
anova( lm( Sepal.Length ~ Species, data=iris ) )$F[1]
```

```
## [1] 119.2645
```

Randomization

To select a randomization sample using physical objects is very similar to the it was described for 2-sample mean procedures. Our null hypothesis suggests we view the variables as independent. So, we might put slips in one bag containing all the response values, and slips indicating the categorical value, one per case, in another. Then we draw one slip out of each bag, and use the slip with the categorical value to assign a group to the reponse value. This sampling is done without replacement, and we have our full randomization sample when the bags are empty.

To illustrate, we look at the count of `Ants` attracted/counted on a sandwhich as it relates to the filling in the sandwich.

```
head(SandwichAnts)
```

```
##   Butter       Filling     Bread Ants Order
## 1     no      Vegemite       Rye   18    10
## 2     no Peanut Butter       Rye   43    26
```

```
## 3     no Ham & Pickles       Rye   44    39
## 4     no      Vegemite Wholemeal   29    25
## 5     no Peanut Butter Wholemeal   59    35
## 6     no Ham & Pickles Wholemeal   34     1
```

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

```
## Ham & Pickles Peanut Butter      Vegemite
##         49.25         34.00         30.75
```

We get our test statistic with

```
testStat <- anova( lm( Ants ~ Filling, data=SandwichAnts ) )$F[1]
testStat
```

```
## [1] 5.626674
```

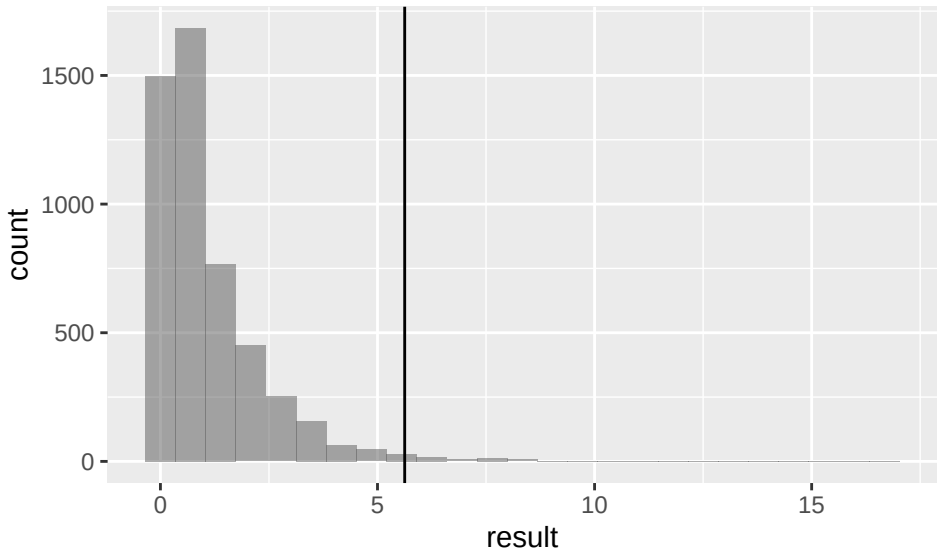
We will locate 5.627 on the approximate null distribution. To simulate it via randomization, we reuse the data, breaking breaking associations between variables (as asserted in the null hypothesis) with the use of `shuffle()`.

```
manyFs <- do(5000) * anova( lm( Ants ~ shuffle(Filling), data=SandwichAnts ) )$F[1]
head(manyFs)
```

```
##       result
## 1 0.29000861
## 2 0.60501743
## 3 3.37286822
## 4 0.01115959
## 5 0.35922330
## 6 0.63726885
```

Now display

```
gf_histogram( ~ result, data=manyFs ) %>% gf_vline(xintercept = ~5.627)
```



As with χ^2 tests, all F tests are 1-sided, right-tailed. The approximate P-value is

```
nrow( filter(manyFs, result >= 5.626 ) ) / 5000
```

```
## [1] 0.0124
```

Obtaining a P-value from a theoretical F distribution

With χ^2 tests (goodness-of-fit, test for association) there we learned conditions under which you can obtain obtain the approximate P-value from a theoretical chi-square distribution, thereby bypassing the need for a randomization distribution. The same can be said for obtaining a P-value in 1-way ANOVA: once you have your test F-statistic, df_1, and df_2 from the sampled data, you can obtain the approximate P-value using a theoretical F-distribution, under these (sufficient) conditions:

- Your data represents an amalgam of independent random samples from the various populations (one population for each value of the explanatory variable),
- Within each of these populations, the distribution of the response variable is normal (or the sample size n_i from Group i is large enough, say 30 or more, though fewer is allowable if Group i appears to be symmetric without outliers), and
- The variances in each population is the same. (In practice, we find the sample standard deviations broken down by group, and check that the ratio of largest to smallest is no bigger than 2.)

When these conditions are met, it is OK to use a theoretical F distribution. Compare randomization and F distributions using this app, and see how well these assumptions predict close alignment between the two.

Example 8.5 (p. 500 from the textbook): The example has us focus on the categorical variable `Award` and quantitative variable `Pulse`. Let's look at various statistics displayed by `favstats()` broken down by `Award`:

```
favstats(Pulse ~ Award, data=StudentSurvey)
```

```
##     Award min   Q1 median Q3 max     mean       sd   n missing
## 1 Academy  42 64.5     71 76  95 70.51613 12.35818  31       0
## 2   Nobel  40 65.0     72 80 130 72.21477 13.09093 149       0
## 3 Olympic  35 60.0     68 74  96 67.25275 10.97067 182       0
```

We see the variable `Award` has $k = 3$ values (so we are considering 3 populations). The null hypothesis, informally stated as "the two variables are independent", is

$$\mathbf{H}_0\colon\ \mu_A = \mu_N = \mu_O,$$

where μ_A represents the mean pulse rate among students who would prefer an Academy Award, μ_N represents the mean pulse rate among students who would prefer a Nobel Prize, and μ_o represents the mean pulse rate among students who would prefer an Olympic medal. Note the samples from these three populations are of size

$$n_A = 31, \quad n_N = 149, \quad \text{and} \quad n_O = 182,$$

all over 30, and the ratio of largest to smallest sample standard deviation

$$\frac{13.091}{10.971} \doteq 1.19$$

is safely smaller than 2. We can find the degrees of freedom:

$$\text{df}_1 = k - 1 = 2, \qquad \text{and} \qquad \text{df}_2 = n - k = (31 + 149 + 182) - 3 = 359.$$

And, we can use this command the calculate the F-statistic (our *test statistic*):

```
anova( lm( Pulse ~ Award, data=StudentSurvey ) )$F[1]
```

```
## [1] 7.103915
```

Finally, the P-value is the area in the F distribution above this test statistic using the `pf()` command with the inputs as just found:

```
1 - pf(7.1039, df1=2, df2=359)
```

```
## [1] 0.0009425421
```

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
```

Example: Textbook Costs

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

```
head(TextbookCosts)
```

```
##            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
```

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
```

Since there are four different `Fields` represented, we should not be surprised that $df_1 = 4 - 1 = 3. And, with 10 courses per field, the full data set has 40 cases, which explains why df\_2 = 40-4 = 36$. 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

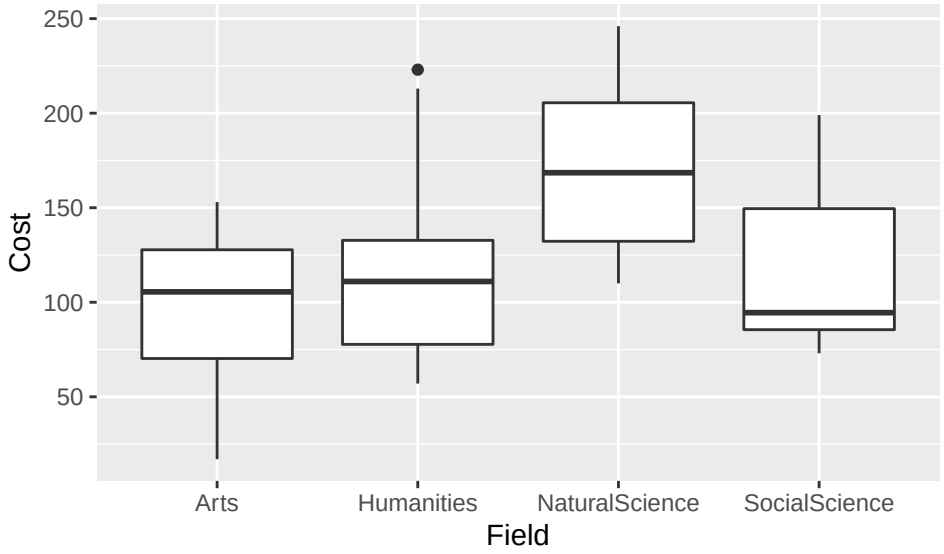
```
1 - pf(4.055, df1 = 3, df2 = 36)
```

```
## [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 )
```



- 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
```

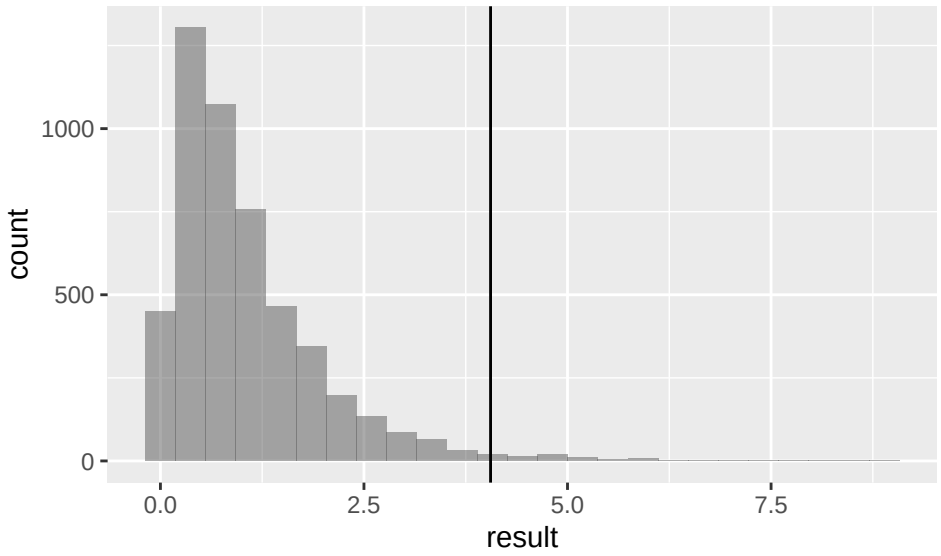
(Note that `favstats()` could also have been used here, but gives extra information we do not need right now.) The ratio of largest-sd-to-smallest is

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

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] 0.0172
```

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

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

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 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
```