

1-Way ANOVA Tests

T.Scofield

Nov. 18, 2016

The Setting

We have a quantitative variable of interest, and samples taken from at least 2 groups in which the (population) standard deviations are equal. If there are precisely 2 groups, we may use the bootstrapping/randomization approaches of Chapters 3 and 4, or (if appropriate) the 2-sample mean (t) procedures of Sections 6.10-6.12. One-way ANOVA is redundant in that case, but very useful for when there are more than 2 groups.

In this context we test these hypotheses:

$$\begin{aligned} \mathbf{H}_0: & \mu_1 = \mu_2 = \dots = \mu_k \quad (\text{the means of the } k \text{ groups are all equal}) \\ \mathbf{H}_a: & \text{for at least one pair of groups, } \mu_i \neq \mu_j \end{aligned}$$

Form of the Data

The form should be much the same as when we compare means of independent samples. For each *case* in the data set, we need (at least) one categorical variable that indicates what group the case is in, and one quantitative measurement (the response variable of interest).

In the iris data set, we have `Species` which can serve as the grouping variable, and we have many quantitative variables on which we could focus.

```
head(iris)
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1          3.5          1.4          0.2 setosa
## 2          4.9          3.0          1.4          0.2 setosa
## 3          4.7          3.2          1.3          0.2 setosa
## 4          4.6          3.1          1.5          0.2 setosa
## 5          5.0          3.6          1.4          0.2 setosa
## 6          5.4          3.9          1.7          0.4 setosa
```

In the job safety data frame (which must be loaded), there are variables which can serve in the two roles as well.

```
js = read.csv("http://www.calvin.edu/~scofield/data/csv/ips5e/jobSafety.csv")
head(js)
```

```
## X jobcat jobc SCI
## 1 1 unskill 1 76
## 2 2 unskill 1 61
## 3 3 unskill 1 56
## 4 4 unskill 1 92
## 5 5 unskill 1 41
## 6 6 unskill 1 51
```

In the `SandwichAnts` data frame, we have the required components, too.

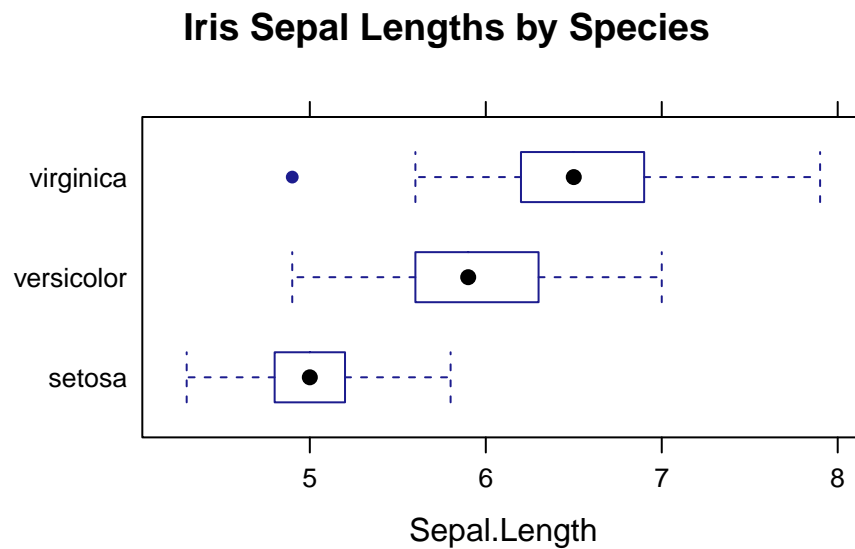
```
head(SandwichAnts)
```

```
##   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
```

Looking at the Data

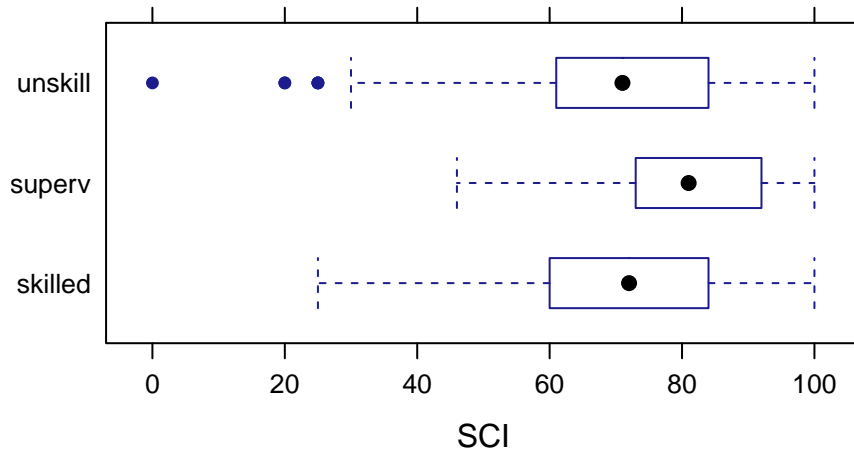
One might look at a breakdown of the response variable by group using, say, side-by-side boxplots. For the data sets mentioned above, here are some examples.

```
bwplot(Species ~ Sepal.Length, data=iris, main="Iris Sepal Lengths by Species")
```



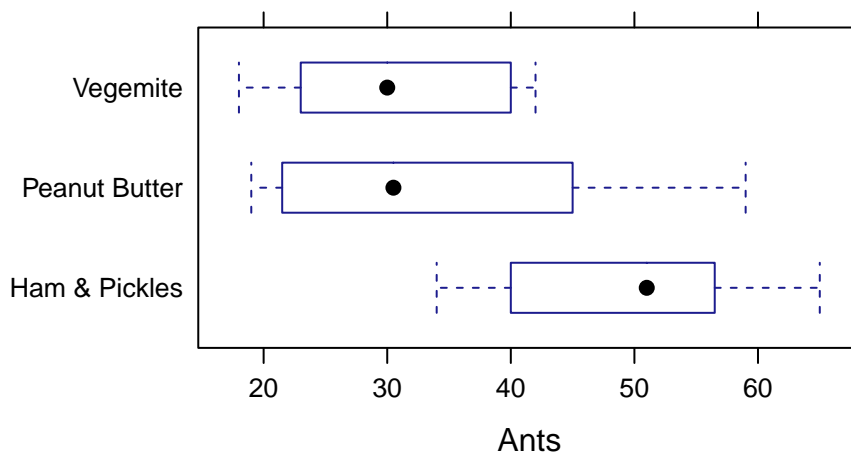
```
bwplot(jobcat ~ SCI, data=js, main="Job Safety Ratings (SCI) by Position Type")
```

Job Safety Ratings (SCI) by Position Type



```
bwplot(Filling ~ Ants, data=SandwichAnts, main="Counts of Ants on Sandwiches with Various Fillings")
```

Counts of Ants on Sandwiches with Various Fillings



Viewing these pictures is not a necessary step to a 1-way ANOVA test. But it is plots such as these which are used in Section 8.1 to build intuition about how a comparison between “variance in group means” and “variance within groups” leads to a testing procedure (in particular, a test statistic).

The ANOVA table and F -statistic

The ANOVA table—how it is constructed, what its various entries reveal—is described in the text. Suffice it to say that the calculations involved are among the most technical yet, though they bear a resemblance to the calculation of *standard deviation* (after all, it is analysis of *variance*). We will use a (mostly) all-in-one-step command to build the table. The root command is one we used earlier in the course when finding the least-squares regression line: `lm()`. (In the earlier usage, *both* our variables were quantitative.) To obtain an ANOVA table for the `iris` variables plotted above, we type

```
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
```

All the numbers in this table are calculated as explained in the text. The number the text refers to as MSG , which measures the variability *between* groups, 31.606. It is the ratio SSG/df_G :

$$MSG = \frac{SSG}{df_G} = \frac{SSG}{k-1} = \frac{63.212}{2} = 31.606,$$

where $k = 3$ represents the number of groups (in this case, the number of species of iris plants represented in the study). While the formula for MSG is not identical to that for the variance of a single quantitative variable, it is the result of adapting that formula to the present situation.

The MSE that appears in this table is the result of a similar calculation:

$$MSE = \frac{SSE}{df_E} = \frac{SSE}{n-k} = \frac{38.956}{147} = 0.265,$$

where the df_E is the difference $n - k$ (sample size minus the number of groups). This number measures the amount of variation *within* groups. If, at this point, it was up to us to compute the F statistic knowing only the MSG and MSE , we could divide them to obtain

$$F = \frac{MSG}{MSE} \doteq 119.26,$$

the ratio of *between*-group variation to *within*-group variation. When the null hypothesis is met, this number would be 1, or quite close to it. As the data becomes less consistent with the null hypothesis, this ratio grows, meaning that a corresponding P -value comes from the proportion of F -values to the *right* of our calculated one (i.e., our *test statistic*), just as when we calculated P -values from a sample's χ^2 test statistic in Chapter 7.

Obtaining a P -value Via Randomization

The command

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

gives us the ANOVA table (see above). The F value is in the first row, 4th column of that table, so this command

```
anova(lm(Sepal.Length ~ Species, data=iris))[1,4]
```

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

gives us just the F -value, our test statistic.

Generating a randomization distribution in an ANOVA setting is no more difficult than in a 2-sample mean setting. We use all the data (i.e., we sample without replacement), but randomly shuffle the values of the (categorical) explanatory variable first. Thus,

```
anova(lm(Sepal.Length ~ shuffle(Species), data=iris))[1,4]
```

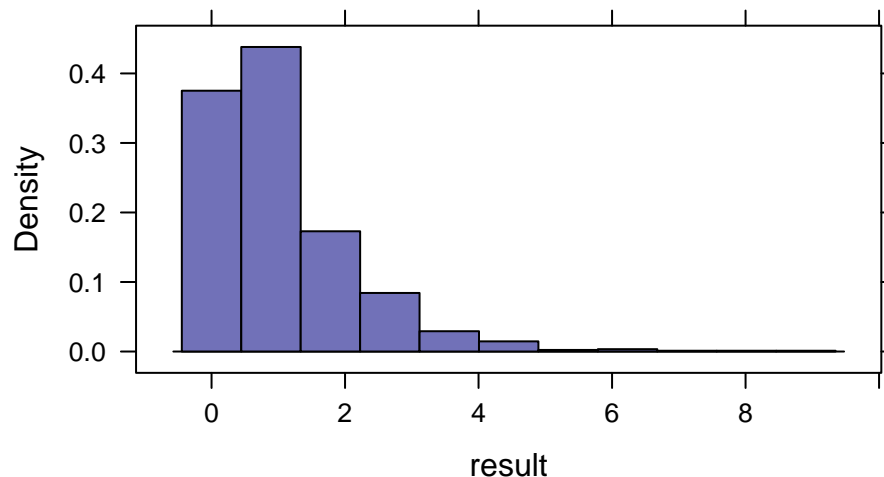
```
## [1] 0.9284832
```

gives us one randomization statistic, and we obtain a simulated distribution of such statistics by repeating this command many times:

```
manyFs = do(1000) * anova(lm(Sepal.Length ~ shuffle(Species), data=iris))[1,4]
head(manyFs)
```

```
##      result
## 1 0.9140273
## 2 1.3421058
## 3 0.2188681
## 4 0.1916673
## 5 1.1742053
## 6 0.6990331
```

```
histogram(~result, data=manyFs)
```



We obtain an approximate P -value by determining how often a randomized F -statistic exceeded ours.

```
nrow(subset(manyFs, result>119.26))
```

```
## [1] 0
```

While a P -value is never zero, this reported number is so because in our 1000 tries we never obtained a randomized F -statistic this large (i.e., it is quite rare when the null hypothesis is true).

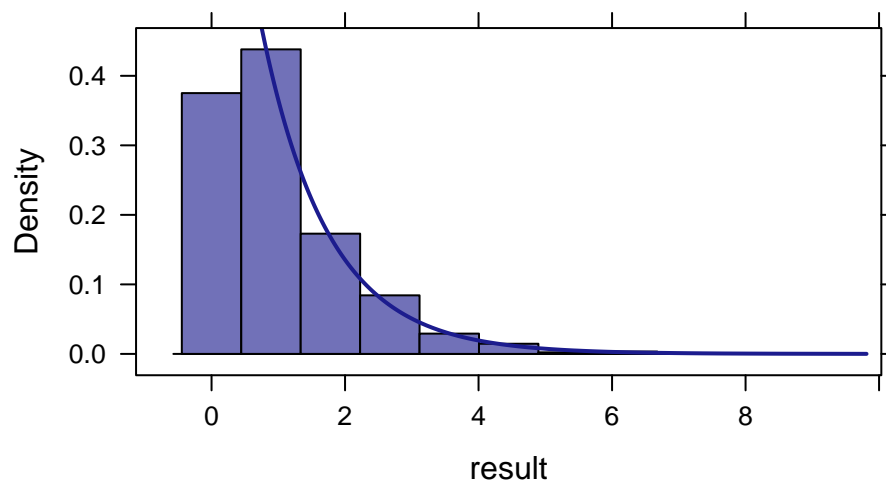
F distributions

The shape of the randomization distribution displayed above may make you wonder if there is a known family of distributions for which this one is a predictable special case. Indeed this is so, and will be when the following assumptions are met:

- Within each group, the response variable is normal. This is, of course, a difficult thing to check, unless within-group sample sizes are large enough to make it less important.
- The (population) standard deviations within each group are equal. This, of course, does not mean the group (sample) standard deviations are equal, but as a rule of thumb we generally ask that the ratio of largest sample standard deviation to smallest is no bigger than 2.

When these conditions are in place, the distribution of F -values approximately follows an F -distribution, of which there are many; the best-fitting one is the one with the two parameters, df_1 and df_2 , set to their corresponding values from the table (with df_G being the one used for df_1). We overlay such a distribution on the previous histogram:

```
histogram(~result, data=manyFs)
plotDist("f", df1=2, df2=147, add=TRUE)
```



As our null distribution is well-modeled using this particular F -distribution, we may compute the P -value using the `pf()` command:

```
pf(119.26, df1=2, df2=147)
```

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

Pairwise Comparisons

When, as above, the P -value leads to a rejection of the null hypothesis, we have only settled that at least one group (population) mean is different. One wants next to know what one (or ones) are different. This is not an issue to be settled “by eye”, but rather with a statistical procedure. The Lock text discusses one method of comparing means two at a time, but you will not be held responsible for knowing the approach they discuss. Instead, you will be expected to be able to use and interpret the results of the `TukeyHSD()` command, used below. (Note the use of `aov()` along with it.)

```
TukeyHSD( aov( Sepal.Length~Species, data=iris ) )
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Sepal.Length ~ Species, data = iris)
##
## $Species
##           diff          lwr          upr p adj
## versicolor-setosa  0.930 0.6862273 1.1737727  0
## virginica-setosa   1.582 1.3382273 1.8257727  0
## virginica-versicolor 0.652 0.4082273 0.8957727  0
```

For this data, we conclude there is a significant difference (at least at the 95% level, as each of the confidence intervals displayed are 95% ones) between each pair of means.