One-Way ANOVA

T.Scofield

You may click here to access the .qmd file.

Settings for 1-way ANOVA

One-way ANOVA generally arises when the research question asks, "Is the (mean) response different between groups?" We must primarily be thinking about a *quantitative* response variable in such settings, but have a categorical group-identifying variable (our explanatory variable). This should sound like 2-sample t, where we have independent sample (response) values from 2 different groups. One-way ANOVA was designed for independent samples taken from k groups (with $k \geq 2$).

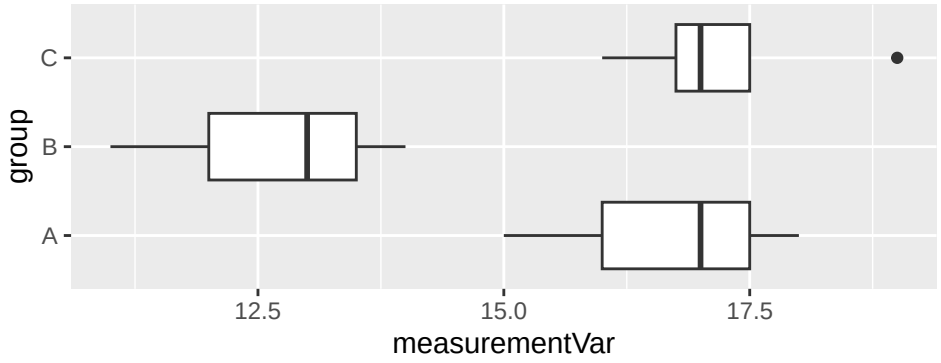
First let's see how raw data for such settings is arranged. I create a data frame with nonsense values to illustrate the typical layout.

```
playData <- data.frame(
  measurementVar = c(15,18,17, 14,11,13, 16,17,19,17),
  group = c(rep("A",3), rep("B",3), rep("C",4))
)
playData
```

```
   measurementVar group
1              15     A
2              18     A
3              17     A
4              14     B
5              11     B
6              13     B
7              16     C
8              17     C
9              19     C
10             17     C
```

If we wish to see side-by-side boxplots for data such as this, the command below can be tailored to the situation.

```
gf_boxplot(group~measurementVar, data=playData)
```



The Sum-of-Squares Calculations of 1-Way ANOVA

There are three sum-of-squares quantities of interest:

$$\text{SSG} = \sum_{\text{groups j}} n_j(\bar{x}_j - \bar{x})^2, \qquad \text{measures across-group variation}$$

$$\text{SSE} = \sum_{\text{observations i, groups j}} (x_{ij} - \bar{x}_i)^2, \qquad \text{measures within-group variation}$$

$$\text{SST} = \sum_{\text{observations i, groups j}} (x_{ij} - \bar{x})^2, \qquad \text{measures total variation in data}$$

To get started on these, we need to know the various means. The grand mean $\bar{x}$ comes from

```
mean(~measurementVar, data=playData)
```

```
[1] 15.7
```

The individual group means is a similar command, but conditioned on group:

```
mean(~measurementVar | group, data=playData)
```

```
       A        B        C
16.66667 12.66667 17.25000
```

Since the individual group sample sizes are $n_A = 3$, $n_B = 3$, $n_C = 4$,

$$\begin{aligned} \text{SSG} &= n_A(\bar{x}_A - \bar{x})^2 + n_B(\bar{x}_B - \bar{x})^2 + n_C(\bar{x}_C - \bar{x})^2 \\ &= 3(16.667 - 15.7)^2 + 3(12.667 - 15.7)^2 + 4(17.25 - 15.7)^2 \quad = \quad 40.013. \end{aligned}$$

For SSE, the calculation takes into account squared deviations of each observation from its own group mean:

$$\begin{aligned} \text{SSE} &= (15 - 16.667)^2 + (18 - 16.667)^2 + (17 - 16.667)^2 + \\ &\quad (14 - 12.667)^2 + (11 - 12.667)^2 + (13 - 12.667)^2 + \\ &\quad (16 - 17.25)^2 + (17 - 17.25)^2 + (19 - 17.25)^2 + (17 - 17.25)^2 \\ &= 14.083. \end{aligned}$$

For SST is similar to SSE, but the calculation takes into account squared deviations of each observation from the *grand mean* $\bar{x}$:

$$\begin{aligned} \text{SST} &= (15 - 15.7)^2 + (18 - 15.7)^2 + (17 - 15.7)^2 + \\ &\quad (14 - 15.7)^2 + (11 - 15.7)^2 + (13 - 15.7)^2 + \\ &\quad (16 - 15.7)^2 + (17 - 15.7)^2 + (19 - 15.7)^2 + (17 - 15.7)^2 \\ &= 54.1. \end{aligned}$$

Note that SSG + SSE = SST.

To make for a useful comparison of across-group variability to within-group variability, we create **mean-square** versions:

$$\begin{aligned} \text{MSG} &= \frac{SSG}{df_1}, \qquad \text{where} \quad df_1 = \#(\text{groups}) - 1, \\ \text{MSE} &= \frac{SSE}{df_2}, \qquad \text{where} \quad df_2 = \#(\text{cases}) - \#(\text{groups}), \end{aligned}$$

The test statistic, called an F-statistic, is the ratio

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

1-Way ANOVA Tables

These calculations are usually displayed together in an ANOVA table. The general layout of an ANOVA (though it may be a matter of taste whether the `df` column comes before the `SS` column) is as displayed in Figure 1 below. In Figure 2, you see a snapshot of an ANOVA table from p. 496 of our textbook. It displays the layout scheme described in Figure 1, but with numbers coming from a specific data set.

Source	df	SS	MS	F-stat	P-value
Groups/Factors	$df_1 = k-1$	$SSG = \sum n_i(\bar{x}_i - \bar{x})^2$	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	P
Residuals/Errors	$df_2 = n-k$	$SSE = \sum (x - \bar{x}_i)^2$	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	$SST = \sum (x - \bar{x})^2$			

Figure 1: ANOVA table layout

Source	DF	SS	MS	F
Filling	2	1561	781	5.63
Error	21	2913	139	
Total	23	4474		

Figure 2: ANOVA table from Example 8.3, p. 496

Not surprisingly, R can be used to generate ANOVA tables. The layout follows the scheme of Figure 1 with the one exception that R does not bother displaying a "Total" row. Here is the command and output table when dealing with my `playData` above. You might be surprised that the `lm()` command plays a role. Look over the computer output of this pair of commands (it follows Figure 2), and see if you recognize some of the numbers from earlier calculations.

```
lmModel <- lm(measurementVar ~ group, data=playData)
anova(lmModel)
```

```
Analysis of Variance Table

Response: measurementVar
          Df Sum Sq Mean Sq F value   Pr(>F)
group      2 40.017 20.0083   9.945 0.009001 **
Residuals  7 14.083  2.0119
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1-Way ANOVA is an Hypothesis Test

One has a quantitative variable in mind, and wants to compare means across various groups or populations.

Note: If the number of groups is only two, then 2-sample t procedures allow us to carry out such comparisons. 1-way ANOVA allows us to deal with more than 2 groups/populations. But,

like in the case of 2-sample t, 1-way ANOVA relies on having **independent samples** from our populations.

The hypotheses

Formal statement of our hypotheses: Given k groups/populations, our null hypothesis is

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

and the alternative is that at least one of these population means is different.

The test statistic: F

The calculations required to build an ANOVA table are routine, but tedious. Most of them are intermediate calculations required in order to get the test statistic, F, which compares (through a ratio) the across-group variation to the within-group variation:

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

Obtaining a P-value

As in other contexts, we place the test statistic on a reference (**null**) distribution, and determine (usually as an area, so employing a cdf function) how often test statistics at least as extreme arise in a world where the null hypothesis is true. It has often been the case that some family of theoretical distributions serves in this role of null distribution, but only if some criteria are met. That is option 1 described below:

1. We may use, as null distribution, a theoretical F-distribution with degrees of freedom df_1 and df_2 whenever these criteria are met

- samples are (near-enough) iid samples drawn independently from their respective groups
- the sampling distributions for each sample mean are approximately normal. As with t-tests, we can know this to be the case if either
  - the groups/populations from which we are drawing are normal, or
  - the sample sizes are at least 30
- the standard deviations in each group/population are the same. (Our rule of thumb here is that the ratio of largest-to-smallest sample standard deviation is no more than 2.)

2. When these criteria are not met, we might turn to constructing an approximate null distribution using simulation.

Example: Sepal Widths of Iris Plants (uses an F-distribution)

`Sepal.Width` is a quantitative variable available in the `iris` data set. We investigate summaries, broken down by `Species`:

```
favstats(Sepal.Width ~ Species, data=iris)
```

```
     Species min    Q1 median    Q3 max  mean        sd  n missing
1     setosa 2.3 3.200    3.4 3.675 4.4 3.428 0.3790644 50       0
2 versicolor 2.0 2.525    2.8 3.000 3.4 2.770 0.3137983 50       0
3  virginica 2.2 2.800    3.0 3.175 3.8 2.974 0.3224966 50       0
```

Here, the 2-to-1 ratio of sample standard deviations is met

$$\frac{s_{\max}}{s_{\min}} = \frac{0.379}{0.314} < 2,$$

and sample sizes are all 50. The ANOVA table arising from this data

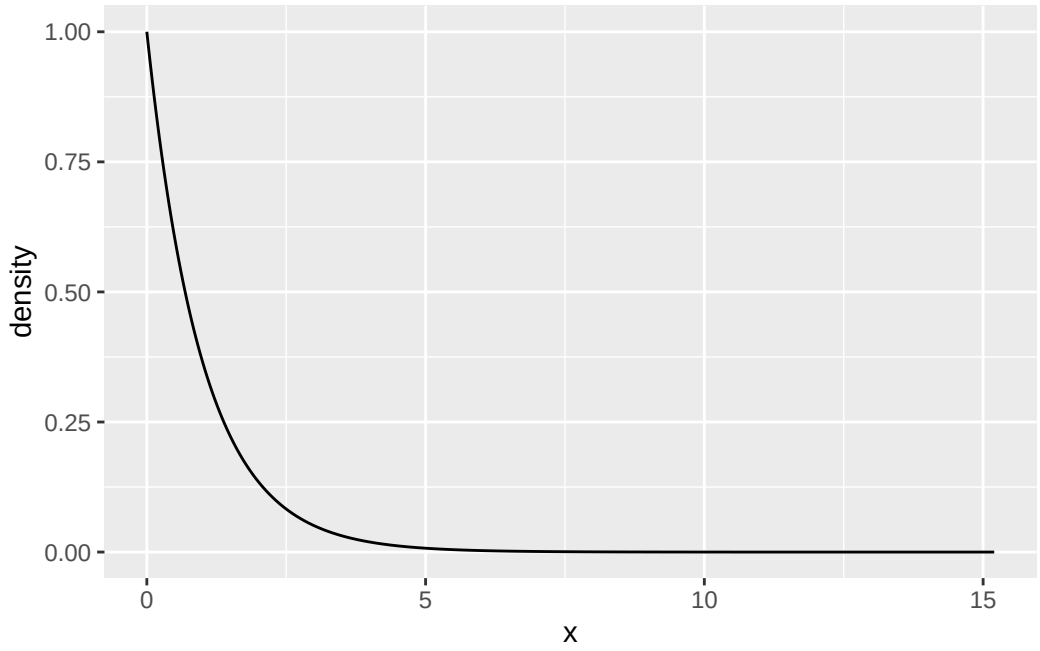
```
anova( lm(Sepal.Width ~ Species, data=iris) )
```

```
Analysis of Variance Table

Response: Sepal.Width
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 11.345  5.6725   49.16 < 2.2e-16 ***
Residuals 147 16.962  0.1154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not only is the F-statistic important, but so are $df_1 = 2$ and $df_2 = 147$. We will use, as null distribution, a theoretical $F(df1 = 2, df2 = 147)$ distribution:

```
gf_dist("f", params=c(2, 147))
```



It appears F-statistics as large as 49.16 are relatively rare. The cdf function `pf()` confirms this:

```
1-pf(49.16, df1=2, df2=147)
```

```
[1] 0
```

So, we would reject the null hypothesis, concluding that at least one of the species of iris plants represented here has a different mean sepal width from another.

Example: Sandwich Ants (uses simulation)

We look at the first few lines of the data set.

```
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
```

We can check assumptions using `favstats()`:

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

```
        Filling min    Q1 median    Q3 max  mean        sd n missing
1 Ham & Pickles  34 42.00   51.0 55.25  65 49.25 10.793517 8       0
2 Peanut Butter  19 21.75   30.5 44.00  59 34.00 14.628739 8       0
3      Vegemite  18 24.00   30.0 39.00  42 30.75  9.254343 8       0
```

The 2-to-1 ratio of sample standard deviations is met,

$$\frac{s_{\max}}{s_{\min}} = \frac{14.63}{9.25} < 2,$$

which is good. But it is anyone's guess whether the number of ants drawn to sandwiches of a particular type of filling follow a normal distribution! And sample sizes of 8 for each filling type do not let us off the hook with regards to normality. That is why a simulation may be more appropriate. The steps taken below can be greatly simplified by using the 1-way ANOVA app at https://connect.cs.calvin.edu/content/f0eac06e-6ec0-481e-91a4-d5dab9b6b966/.

We can obtain an ANOVA table using commands

```
myModel <- lm(Ants ~ Filling, data = SandwichAnts)
anova(myModel)
```

```
Analysis of Variance Table

Response: Ants
          Df Sum Sq Mean Sq F value  Pr(>F)
Filling    2   1561  780.50  5.6267 0.01105 *
Residuals 21   2913  138.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we want only the F-statistic, we can query that with syntax tailored to that aim:

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

```
[1] 5.626674
```

So, our data yields $F = 5.627$ as test statistic.

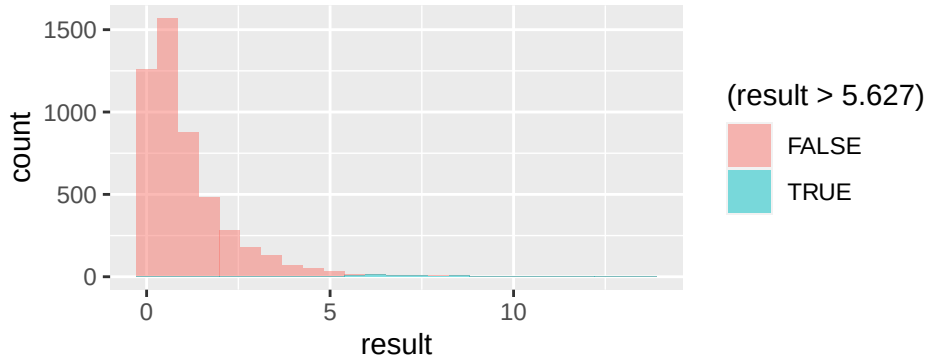
To simulate a null distribution, we will `shuffle()` values in the `Filling` column. (Note: It would work just as well to `shuffle()` the `Ants` column.)

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

```
    result
1 0.8450462
2 0.6983313
3 0.9550110
4 0.0536647
5 0.5113097
6 1.1186236
```

Each of the results in `manyFs` is an F-statistic from a simulated sample for which the null hypothesis—that ants don't pay any attention to `Filling`—holds. A picture of this simulated null distribution, with our test statistic thrown in comes from this command:

```
gf_histogram(~result, data=manyFs, fill= ~(result > 5.627))
```



Note that I am shading only values to the right of our test statistic. Like chi-square tests, 1-way ANOVA (sometimes called F-tests) are right-tailed tests. The proportion of times something as extreme as 5.627 arises in a world where the null hypothesis is true is approximately

```
prop(~ (result > 5.627), data=manyFs)
```

```
prop_TRUE
   0.0124
```

At the 5% level, we reject the null hypothesis, in favor of the alternative, which is that at least one sandwich filling (among vegemite, ham-and-pickles, and peanut butter) draws an average number of ants different from another.