One-Way Analysis of Variance

What it is

One-way analysis of variance, or ANOVA, compares means across two or more samples. The concept is similar to that of the independent samples t test, but generalizes that procedure to any number of groups. Though ANOVA can be used to compare means in two groups, we usually use the independent samples t test for this case and reserve ANOVA for comparing more than two groups.

When to use it

ANOVA can be used to compare means across two or more samples provided the following conditions are true:

- The samples are mutually independent (not paired or otherwise correlated with one another)
- Each sample is either:
  - Approximately normally distributed, OR
  - Large (using the rule of thumb $n \geq 30$)
- The population variances are equal across all groups[1]

How to use it

An ANOVA test begins with hypotheses. If K is the number of groups, the hypotheses are:

H_0: All population means are equal ($\mu_1 = \mu_2 = \ldots = \mu_K$).

H_A: Not all the population means are equal.

Note that the alternative hypothesis is quite vague. There are a lot of ways that "not all" the means can be equal.[2] In the event that the null hypothesis is eventually rejected, a further analysis should be done to determine where any differences exist.

Suppose we want to compare average BMI across metro groups. Then the hypotheses would be:

H_0: All metro groups have the same average BMI ($\mu_{Urban} = \mu_{Suburban} = \mu_{Rural}$).

H_A: Not all metro groups have the same average BMI.

Once the hypotheses are determined, there are a number of ways to run a one-way ANOVA in R[3]. For a quick-and-dirty ANOVA, use the `oneway.test()` function. Provide the function with a formula structured as `outcome ~ groups`, the data set, and set `var.equal=TRUE` (assuming the equal variance rule of thumb is met). The result will

[1] A simple rule of thumb is to check the ratio of the largest sample standard deviation to the smallest. If $\frac{s_{max}}{s_{min}} \leq 2$ then this condition can be said to be reasonably well met.

[2] For example, μ_1 could be different from all the other means, or μ_2 could be different from all others. Or maybe μ_1 and μ_2 are the same, but different from the other means. Or maybe all K means are unique. The possibilities here are numerous...

[3] Either way, it is good to first check some sample statistics. This can be done with `aggregate(bmi ~ metro, data=mydat, FUN=mean)` for means. You can replace `FUN=mean` with `FUN=` any other statistic you want; for example, `FUN=sd` will give the group-specific standard deviations and `FUN=length` will give the group specific sample sizes.

include an test statistic (F), numerator and denominator degrees of freedom, and a p-value:

```
oneway.test(bmi ~ metro, data=mydat, var.equal=TRUE)
```

```
##
##  One-way analysis of means
##
## data:  bmi and metro
## F = 0.02679, num df = 2, denom df = 729, p-value = 0.9736
```

For more detail, you first need to create a linear model with the `lm()` function, then run `anova()` on the result. Within the `lm()` function, You should use the same formula and data setup as in `oneway.test()`. The `anova()` results will include the same information as `oneway.test()`, but with a full ANOVA table, complete with sums of squares and mean squares:

```
bmi_model = lm(bmi ~ metro, data=mydat)
anova(bmi_model)
```

```
## Analysis of Variance Table
##
## Response: bmi
##            Df  Sum Sq Mean Sq F value Pr(>F)
## metro       2     2.2   1.085  0.0268 0.9736
## Residuals 729 29515.5  40.488
```

Why it works

The ANOVA procedure is highly computational, so we will stick with conceptual understanding here. One-way ANOVA compares the differences between groups (denoted by the `metro` row in the table above) to the differences within groups (denoted by the `Residuals` row).[4] If the differences between groups are small, then the F statistic will also be small; conversely, if the differences between groups are large, then the F statistic will be large. Therefore, large values of F provide strong evidence against the null hypothesis.

[4] Explicitly, the F statistic is $F = \frac{MSB}{MSE}$, or mean square between groups divided by the mean square within groups.