Anova and others

Resources:

- wikipedia one-way-anova
- Excellent: ONE-WAY ANOVA (PYTHON FOR DATA SCIENCE)

Notes:

ANOVA

- compare means of two (or usually more) groups using F-test.
- If only two groups, we can use t-test. (for two cases, F-test is equal to t-test).
- Usually three independent features are categorical and target y must be numerical.
- ANOVA is omnibus test, this means it only can say groups are different, but can not say which group is more important. For that, we need to use Post-hoc study.
- If data are ordinal, a non-parametric alternative to this test should be used such as Kruskal–Wallis one-way analysis of variance.
- If the variances are not known to be equal, a generalization of 2-sample Welch's t-test can be used.
- the two-way analysis of variance (ANOVA) is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. The two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

What is the difference between a one-way and a two-way ANOVA?

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

One-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Reebok,Asics) and race finish times in a marathon. Two-way ANOVA: Testing the relationship between shoe brand (Nike, Adidas, Reebok, Asics), runner age group (junior, senior, master’s), and race finishing times in a marathon. All ANOVAs are designed to test for differences among three or more groups. If you are only testing for a difference between two groups, use a t-test instead.

Two Way ANOVA

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable. With a Two Way ANOVA, there are two independents. Use a two way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called factors in a Two Way ANOVA.

The factors can be split into levels. In the above example, income level could be split into three levels: low, middle and high income. Gender could be split into three levels: male, female, and transgender. Treatment groups are all possible combinations of the factors. In this example there would be 3 x 3 = 9 treatment groups.

Assumptions of F-test (Anova test)

Parametric test assumptions

- Population distributions are normal
- Samples have equal variances
- Independence

Anova calculation

```
H0: x1bar = x2bar = x3bar = ... = xkbar
H1: At least one of the group has different mean

F = MS_between / MS_within

MSB = SSB / (k-1)
MSW = SSW / (n-k)   # n is total number and k is number of groups.


if F-stat > F_critical:
   we reject H0

F-critical is obtained from table using given degree of freedom and alpha.
```

Let's say we have 3 types of doses `placebo`, `low` and `high` with each having 5 values for libido. We want to test is high dose better than low dose.

Data Table

Drug Dose	Libido					Sample Size	Sample Means	Sample Variance
Placebo (k_1)	3	2	1	1	4	5 (n_1)	2.2 ($\bar{x}_1$)	1.7 (s_1^2)
Low (k_2)	5	2	4	2	3	5 (n_2)	3.2 ($\bar{x}_2$)	1.7 (s_2^2)
High (k_3)	7	4	5	3	6	5 (n_3)	5.0 ($\bar{x}_3$)	2.5 (s_3^2)
Total ($k = 3$)						15 (n_T)	3.5 ($\bar{x}$)	3.1 (s^2)

Now using the formulas from above, the ANOVA table can be filled in.

Between samples row

$$SS_B = \sum_k n_k(\bar{x}_k - \bar{x})^2 = 5(2.2 - 3.5)^2 + 5(3.2 - 3.5)^2 + 5(5.0 - 3.5)^2 = 20.15$$

$$\text{Degrees of Freedom} = k - 1 = 3 - 1 = 2$$

$$\text{Mean square } (MS_B) = \frac{SS_B}{k-1} = \frac{20.15}{2} = 10.07$$

Within samples row

$$SS_W = \sum_{i,k}(x_{i,k} - \bar{x}_i)^2 \text{ or } \sum_k (n_k - 1)s_k^2 = (5-1)1.7 + (5-1)1.7 + (5-1)2.5 = 23.6$$

$$\text{Degrees of Freedom} = n_T - k = 15 - 3 = 12$$

$$\text{Mean square } (MS_W) = \frac{SS_W}{n_T - k} = \frac{23.6}{12} = 1.97$$

Total row

$$TSS = SS_B + SS_W = 20.15 + 23.6 = 43.75$$

$$\text{Degrees of Freedom} = n_T - 1 = 15 - 1 = 14$$

F-statistic

$$\text{F-statistic} = \frac{MS_B}{MS_W} = \frac{10.07}{1.97} = 5.11$$