Mathematics for Supply Chain

Msc Supply Chain & Purchasing

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

1 Course summary: ANOVA

In the previoux part, we tried to compare means when we had two independent samples or groups. We saw that we could test whether or not the means of the measurement carried out on the two groups are significantly different, using a Student's t test.

In this section, we will try to generalize this to several groups, *i.e.* when we want to compare the means of more than two groups. This can be useful in certain marketing situations, where you need to determine which category of individuals to focus on in order to boost sales or turnover. It is not uncommon then to have to segment the population into disjointed groups in order to identify whether average spending varies from one group to another.

Let us take a look at how to do this, using Analysis of Variance (ANOVA) and which test to use.

1.1 Some recalls

Generalities There are a number of situations where we need to compare averages, for example when we want to study average spending according to vacation location or type of vacation (ski, sea, mountain, hotel club, etc.). The aim is to test whether average expenditure is independent of vacation type. This leads to the following hypothesis test:

Let us take a look at how to do this, using Analysis of Variance (ANOVA) and which test to use.

$$H_0 : \mu_i = \mu \; \forall i \quad \text{v.s.} \quad H_1 : \exists \, i, j \text{ tels que } \mu_i \neq \mu_j$$

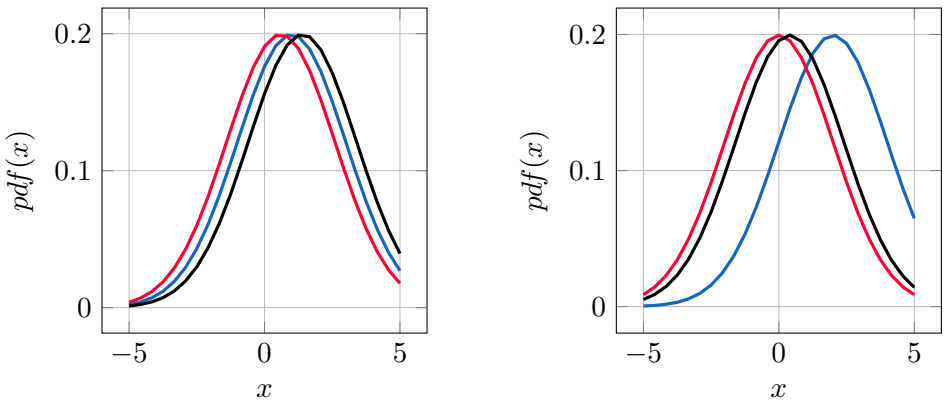

Figure 1: Graphical representation of expenditure as a function of three different groups. The figure on the left illustrates the case where average expenditure is identical from one group to the next, *i.e.* it illustrates the H_0 hypothesis. The figure on the right illustrates the case where the hypothesis H_1 is retained, i.e. the average of at least one of the groups is different from the others

.

This is illustrated on Figure 1 which provides an exempls where the assumption H_0 and H_1 holds respectively. As shown by the graphs, some other assumptions are required in order to perform our analysis of variance:

- having a sample size of at least size 30 in each studied group. This allow to consider their estimator of the mean, in each group, to be normally distributed in the situation the data are normally distributed. Otherwise, the sample size does not matter.
- we will also assume that the variance are equal among the different groups. This assumption is called the **homoscedasticity**.

In what follows, we denote by Y the random variable for which we are studying the mean value and see if its values depends on the studied group.

We are going to consider K different groups and we assume that we have samples of size n_k et we denote by N the total sample size, *i.e.*, $N = \sum_{k=1}^{K} n_k$. The observations will be denoted by $y_{i,k}$ in order to say that the i-th instance belongs to eh group k.

Study of the variance The question now is how to highlight whether or not there is a difference between the mean values of the different groups. The term "Analysis of variances" should give us a hint that we'll certainly have to study variances, but which ones?

So far, we've defined variance in relation to a random variable or a sample, but what if we have a sample for each i group? We can in fact study two "natures" of variance:

- the inter-class variance, denoted SS_{factor} (also known as the sum of squared errors due to the model factor, in another context).
  The variance **between the different classes** is defined by :

$$SS_{\text{factor}} = \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2,$$

  where $\bar{y}_k = \dfrac{1}{n_k} \sum_{i=1}^{n_k} y_{i,k}$ denotes the mean evaluated on the sample of group k and $\bar{y} = \dfrac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} y_{i,k}$ is the overall mean (*i.e.* on all samples).
  This term can be thought of as a weighted variance of means.

- the intra-class variance, denoted SS_{residual} (also known as the sum of squared errors due, which is not due to the factor in another context). This time it's a question of calculating variances within each group/sample. This intra-class variance is defined by

$$SS_{\text{residual}} = \sum_{k=1}^{K} (n_k - 1) s_k^2,$$

  where $s_k^2 = \dfrac{1}{n_k - 1} \sum_{i=1}^{n_k}$ is the **unbiased variance** of the k group/sample. This intra-class variance can therefore be seen as a weighted average of variances.

These two variances are linked by the following equality

$$SS_{\text{total}} = SS_{\text{factor}} + SS_{\text{residual}},$$

where $S_{\text{total}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{i,k} - \bar{y})^2$ is just the variance in the whole dataset.
The previous relation is just the decomposition of the variance.

Hypothesis testing. Let us come back to our test, We recall that we aim to test if the mean of the different groups are equal or not. To do so, we have to use the statistical quantity defined by:

	Mountain	Sea	Hotel-Club
Moyenne : $\bar{y}_k$	1 000	980	1050
Variance : s_k^2	100	150	75

Table 1: Table showing traveler spending by type of stay.

$$F = \frac{\frac{SS_{\text{factor}}}{K-1}}{\frac{SCE_{\text{residual}}}{n-K}} \sim F(K-1, N-K).$$

This test statistic is distributed according to a Fisher distribution with $K-1$ and $N-K$ degrees of freedom. Note that this probability distribution depends on two parameters.

To determine whether or not to reject the hypothesis H_0, we'll carry out a one-tailed test, i.e., we'll compare the value of the test statistic F with the quantile of order $1-\alpha$ of a Fisher distribution with $K-1$ and $N-K$ degrees of freedom.

- we reject H_0 if $f \geq F_{1-\alpha}(K-1, N-K)$,
- we do not reject H_0 otherwise.

We can also compute the p-value and compare it to α. In this case, the p-value is given by

$$\mathbb{P}\left[F(K-1, N-K) \geq F\right].$$

Exemple 1.1. *Let's take a small example to illustrate this notion by returning to our vacation example: we want to know if the type of vacation: mountain, sea or hotel club has an influence on holidaymakers' spending during their stay.*
To do this, a survey was carried out on a total of $N = 600$ *people:* $n_k = 200$ *per type of vacation. The data collected have been processed and are summarized in Table reftab:anova and the distributions are shown in Figure reffig:exanova on the left.*

From this data table, we need to calculate our two variance terms, which we have called the mean of the variances and the variance of the means. Here the calculation will be simpler, as the samples are all of the same size between the different groups.
So we need :

- *calculate the variance of the means, i.e. SS_{factor}, to do this we start by evaluating our overall mean $\bar{y}$ which is equal to*

$$\bar{y} = \frac{1}{K}\sum_{k=1}^{K}\bar{y}_k = \frac{1000 + 980 + 1050}{3} = 1010.$$

So we have

$$\begin{aligned} SS_{factor} &= \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2, \\ &= 200 \times \left((1000 - 1010)^2 + (980 - 1010)^2 + (1050 - 1010)^2\right), \\ &= 520000. \end{aligned}$$

- *calculate the mean of variances, i.e. $SS_{residual}$.*

$$\begin{aligned} SS_{residual} &= \sum_{k=1}^{K}(n_k - 1)s_k^2, \\ &= 199 \times \left(100^2 + 150^2 + 75^2\right), \\ &= 7586875. \end{aligned}$$

So the statistical test is equal

$$F = \frac{\frac{SS_{factor}}{K-1}}{\frac{SS_{residual}}{n-K}} = \frac{\frac{520000}{3-1}}{\frac{7586875}{600-3}} = 20.46.$$

And the quantile of order 0.95 *Of a Fisher distribution with* 2 *and* 597 *degrees of freedom is equal to* $F_{1-\alpha}(2, 597) = 3.01$.

We therefore reject the hypothesis H_0 and can conclude that spending is not the same according to the type of stay, i.e. that the nature of the stay has an influence on the amount spent by the persons in holidays.

To finish with. Note one very important thing: ANOVA won't tell you which group has a significantly different mean from the others, or even how many groups have means that are different from at least one other group. **You can only conclude that there is one group whose mean is different from at least one other group**. But that's enough to conclude that an external factor has an influence on the observed values, *i.e.* that the quantitative variable studied is dependent on the qualitative variable.

2 Exercises

Exercice 2.1 (Mail Services, part II)**.**
The computer operations department had a business objective of reducing the amount of time to fully update each subscriber's set of messages in a special secured email system. An experiment was conducted in which 24 subscribers were selected and three different messaging systems were used. Eight subscribers were assigned to each system, and the update times were measured and they are provided in the file Message.

1. *Analyze the data and write a report to the computer operations department that indicates your findings.*

Exercice 2.2 (Banks)**.**

A data scientist from an insurance company has been asked to study the impact of an advertising campaign carried out in 7 regions in which the company already operates. To this end, he has extracted from the database, for a certain number of general agents in each region, the number of new customers harvested.

Region	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Nb of general agent	*9*	*7*	*7*	*6*	*7*	*6*	*6*
Average nb of new customers	*26.88*	*22.34*	*19.54*	*18.95*	*27.17*	*25.87*	*25.72*
Nb variance of new customers	*13.54*	*12.59*	*12.87*	*13.42*	*13.17*	*12.56*	*12.64*

The statistical engineer then decides to carry out an analysis of variance to test whether the region factor has an influence on the number of new customers collected. Let X_k^i be the number of new customers of the $i-th$ general agent in region k. Let n_k be the

number of general agents in region k, and K the number of regions ($K = 7$). We assume that the random variables X_k^i are normal, with mean μ_k and variance σ.

In the following we set:

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_k^i \quad \textit{and} \quad \bar{X} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_k^i \quad \textit{and} \quad N = \sum_{k=1}^{K} n_k.$$

1. *Recall the assumption of the ANOVA, are they fullfilled here?*
2. *What are the assumptions H_0 and H_1.*
3. *Give the mean meaning of $\bar{X}_k$ et $\bar{X}$.*
4. *Try to make a decomposition of the variance as follows*

$$V_T = V_R + V_A,$$

Saying differently, the total variance is equal to the sum of the variance explained by the factor A, V_A and the residual variance V_R

   (a) *Recall the definition of the previous quantities and compute them.*
   (b) *Determine whether the advertising campaign had the same impact in all regions. Conclude at the first order risk $\alpha = 0.05$. For this question, we don't ask you to calculate the p-value, but rather to compare yourself with the critical value equal to 2.33: quantile of order $1-\alpha$ of a Fisher distribution with $K-1$ and $N-K$ degrees of freedom.*