Analysis of Variance(ANOVA)

S CAgarkar

V N BRIMS

Thane

Introduction

- Analysis of variance (abbreviated as ANOVA) enables us to test the significance of the difference between more than two sample means.
- ANOVA would be useful in situations like comparing the mileage achieved by five different brands of gasolene, testing which of four different testing methods produces the faster learning record or comparing the earnings of graduates of different business schools.

The situation

- The training director of a company is trying to evaluate three different methods of training new employees. The first method assigns each to an experienced employee for individual help in the factory, The second method puts all new employees in a training room separate from the factory and the third method uses training films and programmed learning materials. The training director chooses 16 new employees assigned at random to 3 training methods and records their daily production after they complete the programme. Based on the data (given below) help the director to find out if there are differences in effectiveness among these methods.

Method 1	**15**	**18**	**19**	**22**	**11**	**-**
Method 2	22	27	18	21	17	-
Method 3	18	24	19	16	22	15

Basic concepts in ANOVA

- Analysis of variance is based on a comparison of two different estimates of the variance (σ^2) of our overall population. We can calculate one of these estimates by examining the variance among the three sample means (Between-column variance).
- The other estimate of the population variance is determined by the variation within the three samples themselves (Within-column variance).
- Compare these two estimates of population variance. Since both are estimates of σ^2 they should be approximately equal in value when the null hypothesis is true. If the null hypothesis is not true these two estimates will differ considerably.

Steps in ANOVA

- Determine one estimate of the population variance from the variance among the sample means.
- Determine a second estimate of the population variance from the variance within the samples.
- Compare these two estimates. If they are approximately equal in value, accept the null hypothesis.
- Going back to the problem given earlier we see that $x_1 = 17$ $x_2 = 21$ $x_3 = 19$; $n_1 = 5$, $n_2 = 5$ and $n_3 = 6$
- The grand mean x

$= (5/16)*17+(5/16)*21+ (6/16)*19=304/16= 19$

Between column variance

n	X	x	X-x	$(X-x)^2$	$n(X-x)^2$
5	17	19	-2	4	20
5	21	19	2	4	20
6	19	19	0	0	00

$\sigma^2 = \sum n_j (X-x)^2 / k-1$

σ^2= estimate of the population variance based on the variance among sample means called between-column variance

n_j= the size of the jth sample, X = the sample mean of jth sample

x= the grand mean and K= the number of samples

Substituting the values in the above formula we get

$\sigma^2 = 40/ 3-1$

= 40/2 = 20

It means the value of between-column variance = 20

Within-column variance

- The sample variation s^2 is given by the formula
- $s^2 = \sum (X-x)^2/n-1$
- Hence $s_1^2 = 70/ 5-1 = 17.5$;
- Similarly, $s_2^2 = 15.5$ and $s_3^2 = 12.0$
- Within column variance σ^2 is given by the formula
- $\sigma^2 = \sum (n_j-1/n_T-k)s_j^2$
- $= 4/13\ (17.5) + (4/13)\ (15.5) + (5/13)\ (12.0)$
- $= 192/13$
- $= 14.769$

The F Statistic

- Once two estimates of population variance are calculated then their ratio (called F) is taken.

F = Between-column variance/Within-column variance

F = 20/14.769 = 1.354

- In the above formula the numerator is the variation among sample means of the three methods. It is a good estimator of population variance. The denominator is based on the variance within the samples. It is also a good estimator of population variance.
- The numerator and denominator should be about equal if the null hypothesis is true.

The F Distribution

- Like the t distribution, the F distribution is a whole family of distributions.
- Each distribution is identified by a pair of DF (degrees of freedom). Recall that for t distribution, there is only one DF.
- The first number refers to the number of DF in the numerator of F ratio while the second to DF in the denominator.
- F distribution has a single mode. They are usually skewed towards the right but become symmetrical as both the DFs increase.

Calculating DFs

- The numerator of the F ratio is the value of between column variance. In our calculation we used three values to compute it. Once we know two of them the third is automatically determined. Thus the no. of DF for numerator in F ratio is one less than the number of samples (s-1).
- For calculating the denominator (within column variance) of a F ratio we used all three samples. In each sample one DF is lost. Hence no. of DF for denominator is (n_T – k), that is 16-3=13

Using F table

- To test the hypothesis using F test we have to use values given in F tables. Separate tables exists for each level of significance. In these tables columns show the DFs for numerator while rows show DFs for denominator.
- In the table prepared for 0.05 level of significance we find that the value corresponding to DFs 2 and 13 is 3.81.

Testing the hypothesis

- Suppose the director of training wants to test at 0.05 level of significance the hypothesis that there is no difference among the three training methods.
- From the table we found that value of F ratio for DF 2 and 13 as 3.81. It sets the upper limit of the acceptance region.
- Through calculation we found the value of F ratio as 1.354. Since this value lies in the acceptance region, the director can accept the null hypothesis.
- The conclusion is according to the sample information there is no difference in the effects of three training methods on employee productivity.

Use of Computers for F test

- For convenience we had taken a smaller sample in our problem. In actual practice the size may be quite large and calculation would be tedious. In such cases use of computers is advocated.
- A software package for social sciences called SPSS (Statistical Packages for Social Sciences) is developed. Using SPSS one can undertake ANOVA easily.

Problem to solve

- The manager of an assembly line in a clock manufacturing plant decided to study how different speeds of the conveyor belt affect the rate of defective units produced in an 8 hour shift. To examine this, he ran the belt at 4 different speeds for 8 hour shifts each and measured the number of defective units found at the end of the shift. The result is given below.
  - Sp 1: 36,34,37,35,33; Sp 2: 29,34,34,36,32
  - Sp 3: 31,35,32,33,39; Sp 4: 36,28,34,32,30
- Calculate the mean number of defective units for each speed, determine the grand mean (33.5), estimate between column variance, (8.333), calculate within column variance, (7.375), calculate F ratio (1.130). At the 0.05 level of significance do four different conveyor belt speeds produce the same rate of defective clocks per shift? (Ans. Yes, the same).

One more problem to solve

- The study compared the effects of 4 one- month point-of-purchase promotions sales. Below are the unit sales for 5 stores using all 4 promotions in different months,

Free sample: 77 86 80 88 84

One-pack gift: 95 92 88 91 89

Cents off:72 77 68 82 75

Refund by mail: 80 84 79 70 82

- Calculate the mean unit sale for each promotion and then determine the grand mean(81.95), estimate between-column variance, (238.05), calculate within-column variance (21.05), calculate F ratio (11.31). At the 0.05 level of significance do the promotions produce different effects on sale? (Ans. Yes, different).

Thank you. Remember, there is so much variation in nature.