How can I determine the correct term in an ANOVA using Stata?

Authored by

stats writer

July 1, 2024

In order to determine the correct term in an ANOVA using Stata, one must carefully consider the variables and their levels of significance. ANOVA, or analysis of variance, is a statistical test used to compare means between two or more groups. In Stata, this can be done by first importing the data and then using the "anova" command to run the test. It is important to carefully select the appropriate variables and check for any potential confounding factors. Additionally, understanding the output and interpreting the p-values can help determine the most significant term in the ANOVA. This process requires attention to detail and a thorough understanding of the statistical concepts involved.

FAQ:How can I determine the correct term in an anova using Stata?

One method for determining correct denominators in analysis of variance is the Cornfield-Tukey method. To learn how to manually compute Cornfield-Tukey see our FAQ page. This page will demonstrate the use of the Stata program ctems to do the Cornfield-Tukey computation. ctems can be found and downloaded using the command, search ctems (see How can I use the search command to search for programs and get additional help? for more information about using search), and following the instructions. The ctems package comes with three datasets for two-factor, three-factor and split-plot factorial designs; just click on the prompt to get ancillary files. These datasets will be used in the following examples.

ctems applies the Cornfield-Tukey algorithm to a special dataset to compute the expected mean squares for an ANOVA model. The dataset has two string variables: effect and subscript. There are as many rows as there are terms in the ANOVA linear model. In each row, effect is an ASCII representation of each effect and subscript are the letters (and parentheses) of the effect subscript.

ctems is not a particularly easy program to use. The program knows nothing about ANOVA or your design other than what is entered in the dataset and the command options. It cannot detect inconsistencies or deficiencies in your model. Please use ctems carefully.

We will begin with a simple two-factor design with four level of A, 2 levels of B and 8 subjects within each cell.

```
use crf2design, clear

clist
```

effect subscript

1. A j

2. B k

3. A*B jk

4. e i(jk)

As you can see each row in the dataset contains one effect and the subscripts associated with the effect. To run the ctems command we need to inform the program about the individual subscripts, index(i j k l), the levels associated with each subscript, levels(8 4 2), and whether each effect is random or fixed, random(1 0 0). For this first example we will treat the within subjects as random and everything else as fixed. We can tell this because in the random option only the first value is a one, the rest are zero.

ctems, index(i j k) levels(8 4 2) random(1 0 0)

Linear Model:

Y_ijk = mu + A_j + B_k + A*B_jk + e_i(jk)

+------------------------------+

effect subscript ems
1.	A j e + 16A
2.	B k e + 32B
3.	A*B jk e + 8A*B
4.	e i(jk) e

We can see, in the output above, that there is only one error term, ei(jk), needed for each of the effects. A correctly-formed F-ratio will have the same terms in the numerator and denominator except for the effect of interest.

Next, we will rerun the command specifying that all of the effects are random.

```
ctems, i(i j k) l(8 4 2) r(1 1 1)
```

Linear Model:

$$Y_{ijk} = mu + A_j + B_k + A*B_{jk} + e_{i(jk)}$$

effect subscript ems

1. | A j e + 8A*B + 16A |
2. | B k e + 8A*B + 32B |
3. | A*B jk e + 8A*B |
4. | e i(jk) e |

+-----------------------------------+

The output above shows that A*B is the error term for both A and B while the residual error, e, is the error term for A*B.

Now, let's try a three-factor model. In this example there are five subjects per cell with four levels of A and two levels of both B and C. We will start with a fixed-effects model.

use crf3design, clear

clist

effect subscript
1. A j
2. B k
3. C l
4. A*B jk
5. A*C jl

6. B*C kl

7. A*B*C jkl

8. e i(jkl)

ctems, i(i j k l) l(8 4 2 2) r(1 0 0 0)

Linear Model:

Y_ijkl = mu + A_j + B_k + C_l + A*B_jk + A*C_jl + B*C_kl + A*B*C_jkl + e_i(jkl)

+---------------------------------+

| effect subscript ems |

|---------------------------------|

1. | A j e + 32A |

2. | B k e + 64B |

3. | C l e + 64C |

4. | A*B jk e + 16A*B |

5. | A*C jl e + 16A*C |

6. | B*C kl e + 32B*C |

7. | A*B*C jkl e + 8A*B*C |

8. | e i(jkl) e |

+---------------------------------+

In the fixed-effects factorial ANOVA the residual within cell variance, ei(jkl), serves as the error term for all of

the effects in the model.

In the next example, we will declare B to be random along with the residual error, ei(jkl) by changing the random option to r(1 0 1 0).

ctems, i(i j k l) l(8 4 2 2) r(1 0 1 0)

Linear Model:

Y_ijkl = mu + A_j + B_k + C_l + A*B_jk + A*C_jl + B*C_kl + A*B*C_jkl + e_i(jkl)

+--+
| effect subscript ems |
|--|
1. | A j e + 16A*B + 32A |
2. | B k e + 64B |
3. | C l e + 32B*C + 64C |
4. | A*B jk e + 16A*B |
5. | A*C jl e + 8A*B*C + 16A*C |
6. | B*C kl e + 32B*C |
7. | A*B*C jkl e + 8A*B*C |
8. | e i(jkl) e |
+--+

Now, A*B serves as the error term for A, B*C is the error term for C main effect, and ei(jkl) is the error term for B, A*B, B*C and A*B*C.

Next, we will look at a model in which all of the variables are random.

ctems, i(i j k l) l(8 4 2 2) r(1 1 1 1)

Linear Model:

Y_ijkl = mu + A_j + B_k + C_l + A*B_jk + A*C_jl + B*C_kl + A*B*C_jkl + e_i(jkl)

+--+

| effect subscript ems |

|--|

1. | A j e + 8A*B*C + 16A*C + 16A*B + 32A |

2. | B k e + 8A*B*C + 32B*C + 16A*B + 64B |

3. | C l e + 8A*B*C + 32B*C + 16A*C + 64C |

4. | A*B jk e + 8A*B*C + 16A*B |

5. | A*C jl e + 8A*B*C + 16A*C |

6. | B*C kl e + 8A*B*C + 32B*C |

7. | A*B*C jkl e + 8A*B*C |

8. | e i(jkl) e |

+--+

The output indicates that A*B*C will work as the error term for A*B, A*C and B*C. The residual within cell, ei(jkl), is the appropriate error term for A*B*C. However, there is no single effect that will serve as an effort term for any of the main-effects. Testing these effects will require some type of quasi-F-ratio (Kirk, 1998).

For our final example, we will demonstrate a split-plot factorial design. In this example, the between-subjects factor (A) has two levels and the within-subjects factor (B) has four levels. This split-plot factorial design has a term B*Si(jk) with one observation in each cell.

use spfdesign, clear

clist

effect subscript
1. A j
2. S i(j)
3. B k
4. A*B jk
5. B*S ki(j)
6. e i(jk)

ctems, i(i j k) l(1 2 4) r(1 0 0)

Linear Model:

Y_ijk = mu + A_j + S_i(j) + B_k + A*B_jk + B*S_ki(j) + e_i(jk)

+-------------------------------------+
| effect subscript ems |
|-------------------------------------|
1. | A j e + 4S + 4A |
2. | S i(j) e + 4S |
3. | B k e + 1B*S + 2B |
4. | A*B jk e + 1B*S + 1A*B |
5. | B*S ki(j) e + 1B*S |
6. | e i(jk) e |
+-------------------------------------+

As you can see, the error term for A should be Si(j) and the error term for both B and A*B is B*Ski(j).

Reference

Kirk, Roger E. (1998) Experimental Design: Procedures for the Behavioral Sciences,
Third Edition. Monterrey, California: Brooks/Cole

Publishing. ISBN 0-534-25092-0