ANOVA with Summary Statistics: A STATA Macro

Nadeem Shafique Butt
Department of Social and Preventive Pediatrics
King Edward Medical College, Lahore, Pakistan

Shahid Kamal
Institute of Statistics, University of the Punjab
Lahore, Pakistan

Muhammad Qaiser Shahbaz
GC University, Lahore, Pakistan

Abstract

Almost all available statistical packages are capable of performing Analysis of Variance (ANOVA) from raw data. Some of statistical packages have capability to perform independent sample t-test, and some other tests of significance on summary data, but seldom would you come across a software that has the capability to perform ANOVA directly on summary data. However some packages can perform one-way ANOVA after generating surrogate data from summary statistics. In this short note we have given STATA program to perform one-way ANOVA on summary data; in addition this program also performs Bartllet's tests of equality of variances. The idea can be extended to two-way and higher way ANOVA's. Examples have been given for illustration.

Key Words: ANOVA, Summary data.

1. Introduction

In some cases, where only summary data is available, a researcher may want to perform Analysis of Variance (ANOVA) from the available information. Most statistical programs are designed to perform ANOVA on raw data only. David A. Larson (1992) describes a method to generate surrogate data from the summary statistics that can be used to perform one way ANOVA. According to Larson one has to generate two new columns namely $X_j's$ and $X_n's$ as

$$X_j's = \bar{X}_j + \sqrt{\frac{S_j^2}{n_j}} \quad \text{and} \quad X_n's = n_j\bar{X}_j - (n_j - 1)X_j's$$

And after some data manipulation data is ready to perform ANOVA in usual way. We use an alternative method that can be used to perform one way, two way and higher way ANOVA by using summary measures $n_j, \bar{x}_j$ and s_j for $j = 1, 2, ..., k$, where $n_j, \bar{x}_j$ and s_j are, respectively, the size, mean and standard deviation of j-th treatment. In this note we have given a STATA program that can be used jointly to perform one way ANOVA and equality of variances if only summary measures are available. However, this program can be easily extended for two-way and higher way ANOVA.

2. Methodology

The one-way, two-way and higher-way ANOVA uses certain statistical models for operation. Specifically, the one-way fixed effect ANOVA model is given as:

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad \begin{matrix} j = 1,2,...,k \\ i = 1,2,...,n_j \end{matrix} \tag{2.1}$$

This model can be used to compare the significance of treatments or to test the equality of several means. The null hypothesis of interest in this model can be stated in any one of the following ways:

$$H_0 : \tau_j = 0 \qquad for \qquad j = 1,2,.....,k \tag{2.2}$$

or $\quad H_0 : \mu_1 = \mu_2 = ... = \mu_k$

This hypothesis can be tested by using the F – ratio given as:

$$F = \frac{SST/(k-1)}{SSE/(n-k)} \tag{2.3}$$

where $SST = \sum_{j=1}^{k} n_j \bar{x}_j^2 - n\bar{x}_{..}^2 \quad with \quad \bar{x}_{..} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j \;;\; n = \sum_{j=1}^{k} n_j$ (2.4)

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2 \tag{2.5}$$

The significance of the group means can be tested by using the p–value of computed F statistic. STATA program to perform these analyses is given in Appendix - A.

We have also given the STATA commands to perform Bartlett's Test of equality of variance on summary data. The test statistic to be used in this test is given as:

$$\chi^2 = (MC^{-1}) \tag{2.6}$$

with $\quad M = \sum_{i=1}^{k} (n_i - 1)\ln\hat{\sigma}^2 - \sum_{i=1}^{k} (n_i - 1)\hat{\sigma}_i^2 \;\; and \;\; C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k}\right]$

also $\quad \hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{k} (n_i - 1)\hat{\sigma}_i^2$

The two-way fixed effect ANOVA model is given as:

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \qquad \begin{matrix} i = 1,2,...,r \\ j = 1,2,...,c \end{matrix} \tag{2.7}$$

The null hypotheses of interest in two-way ANOVA are:

$$\begin{aligned} H_0^{/} &: \beta_i = 0 \;\; for \; i = 1,2,...,r \quad or \quad H_0^{/} : \mu_{1.} = \mu_{2.} = ... = \mu_{r.} \\ H_0^{//} &: \tau_j = 0 \;\; for \; j = 1,2,...,c \quad or \quad H_0^{//} : \mu_{.1} = \mu_{.2} = ... = \mu_{.c} \end{aligned} \tag{2.8}$$

but generally we are more concerned with the hypothesis of columns (treatments). The necessary sums of squares to test the above hypotheses in two-way ANOVA are given as:

$$SSTr = r\sum_{j=1}^{c}\overline{x}_{.j}^2 - n\overline{x}_{..}^2 \qquad with \quad \overline{x}_{..} = \frac{1}{c}\sum_{j=1}^{c}\overline{x}_{.j} = \frac{1}{r}\sum_{i=1}^{r}\overline{x}_{i.} \tag{2.9}$$

$$SSB = c\sum_{j=1}^{k}\overline{x}_{i.}^2 - n\overline{x}_{..}^2 \tag{2.10}$$

$$SSE = (r-1)\sum_{j=1}^{c}s_j^2 - SSB \qquad \text{or} \qquad SSE = (c-1)\sum_{j=1}^{r}s_i^2 - SST \tag{2.11}$$

where $\overline{x}_{i.}$ is mean of i-th row (block), $\overline{x}_{.j}$ is mean of j-th column (treatment), s_i^2 is variance of i-th row and s_j^2 is variance of j-th column (treatment). The F–ratio's to test the significance of blocks and treatments are given as:

$$F_1 = \frac{SSB/(r-1)}{SSE/(r-1)(c-1)} \qquad and \qquad F_2 = \frac{SSTr/(c-1)}{SSE/(r-1)(c-1)} \tag{2.12}$$

The idea can be easily extended to three-way ANOVA with model:

$$y_{ij(k)} = \mu + \alpha_i + \beta_j + \tau_{(k)} + \varepsilon_{ij(k)} \qquad i, j, k = 1, 2, \ldots p \tag{2.13}$$

The null hypotheses in three -way ANOVA are given as:

$$\begin{aligned} &H_0^{/}: \alpha_i = 0 \ \ for\ i = 1,2,...,p \quad or \quad H_0^{/}: \mu_{1..} = \mu_{2..} = ... = \mu_{p..} \\ &H_0^{//}: \beta_j = 0 \ \ for\ j = 1,2,...,p \quad or \quad H_0^{//}: \mu_{.1.} = \mu_{.2.} = ... = \mu_{.p.} \\ &H_0^{///}: \tau_{(k)} = 0 \ \ for\ k = 1,2,...,p \ or \quad H_0^{///}: \mu_{..(1)} = \mu_{..(2)} = ... = \mu_{..(p)} \end{aligned} \tag{2.14}$$

but again we are more concern with the last hypothesis of treatments. The sums of squares to test the significance of above hypotheses are given as:

$$SSB = p\sum_{i=1}^{p}\overline{x}_{i..}^2 - n\overline{x}_{...}^2 \quad with \quad \overline{x}_{...} = \frac{1}{p}\sum_{i=1}^{p}\overline{x}_{i..} = \frac{1}{p}\sum_{j=1}^{p}\overline{x}_{.j.} = \frac{1}{p}\sum_{k=1}^{p}\overline{x}_{..(k)};\ n = p^2 \tag{2.15}$$

$$SSC = p\sum_{j=1}^{p}\overline{x}_{.j.}^2 - n\overline{x}_{...}^2 \tag{2.16}$$

$$SSTr = p\sum_{k=1}^{p}\overline{x}_{..(k)}^2 - n\overline{x}_{...}^2 \tag{2.17}$$

$$SSE = (p-1)\sum_{k=1}^{p}s_k^2 - SSR - SSC = (p-1)\sum_{i=1}^{p}s_i^2 - SSC - SSTr = (p-1)\sum_{j=1}^{p}s_j^2 - SSTr - SSC \tag{2.18}$$

in the above sums of squares $\overline{x}_{i..}, \overline{x}_{j..}$ and $\overline{x}_{..k}$ are, respectively, the row, column and treatment means. Also s_i^2, s_j^2 and s_k^2 are the rows, columns and treatment variances, respectively.

The F–ratios to test the significance of hypotheses in three-way ANOVA are given as:

$$F_1 = \frac{SSR/(p-1)}{SSE/(p-1)(p-2)}, F_2 = \frac{SSC/(p-1)}{SSE/(p-1)(p-2)}, F_3 = \frac{SSTr/(p-1)}{SSE/(p-1)(p-2)} \qquad (2.19)$$

3. Numerical Example

In this section we have given two numerical examples to demonstrate the usefulness of the program to perform ANOVA on summary statistics.

Example 1: Data has been taken from Moore and McCabe (1998). The variable of interest is lead concentration recorded as milligrams per square meter and research question is to see significant mean difference over the years. Here are some summary data for five years:

Years	n_j	$\bar{x}_j$	s_j
1976	59	6.80	0.58
1977	58	6.75	0.68
1978	58	6.76	0.50
1982	68	6.50	0.55
1987	70	6.40	0.68

The results of the example by using the STATA program “anovai” are given below:

```
. anovai n mean sd

------------------------------------------------------------------------
SOV                df       SS        MS        F            P
------------------------------------------------------------------------
Between Groups      4      8.37      2.09     5.74       0.0002
Within Groups     308    112.29      0.36
------------------------------------------------------------------------
 Total            312    120.66
------------------------------------------------------------------------
Bartllet's test for equality of variances
Chi-Sq = 8.693    df = 4      P = 0.0693
```

Example 2: Data has been taken from Hanif et al (2004). Study is to compare three different drugs, grouping the subject into blocks on the basis of age (because it is known that age affects systolic blood pressure systematically). Summary data is given in the following table.

Drug	Mean	SD
Drug A	100.00	7.91
Drug B	83.00	6.71
Drug C	95.00	11.18
Age Group		
20 – 30	100.00	10.00
30 – 40	91.67	12.58
40 – 50	85.00	8.66
50 – 60	95.00	18.03
> 60	91.67	2.89

The results of the example by using the STATA program "anova2i" are given below:

```
. anova2i cn cmean rmean csd
------------------------------------------------------------------
 SOV               df       SS        MS        F           P
------------------------------------------------------------------
 Treatment          2    763.34    381.67     5.36      0.0334
 Block              4    360.00     90.00     1.26      0.3596
 Error              8    570.00     71.25
------------------------------------------------------------------
 Total             14   1693.34
------------------------------------------------------------------
```

Appendix – A: STATA program to perform to perform one-way ANOVA

In this section we have given a STATA program named "anovai.ado" that can be used to perform the ANOVA on summary data. The program "anovai.ado" requires three columns as input, column of sample sizes, column of mean and column of standard deviations respectively. This program can be easily modified for two-way and higher-way ANOVA's, further this methodology can be easily incorporated in any spread-sheet based softwares (e.g. MS excel, SPSS, Statistica etc).

---

anovai.ado

```
program define anovai
/* Require three cols as input sample size mean and sd */
args `1' `2' `3'
gen v1=`1'*`2'
gen v2=`1'*`2'^2
gen v3= (`1'-1)*`3'^2
egen sumn=sum(`1')
egen sumv1=sum(v1)
egen sumv2=sum(v2)
egen sumv3=sum(v3)
gen gmean=sumv1/sumn
gen ssb=sumv2-sumn*gmean^2
gen sse=sumv3
gen df=sumn-1
```

```
  gen df1=[_N]-1
  gen df2=df-df1
  gen f=(ssb/df1)/(sse/df2)
  gen p=1- F(df1,df2,f)
  /*calculation for Bartllet's test*/
  gen ninv=1/(`1'-1)
  gen nmins=`1'-1
  egen sumnmins=sum(nmins)
  egen sumninv=sum(ninv)
  gen c=1+((sumninv-(1/sumnmins)))/(3*(_N+1))
  gen ssq=sumv3/sumnmins
  gen t1=(`1'-1)*ln(ssq)
  gen t2= (`1'-1)*ln(`3'^2)
  egen sumt1=sum(t1)
  egen sumt2=sum(t2)
  gen b=(sumt1-sumt2)/c
  gen pchi= chi2tail(_N-1,b)
  /*command to control output for anova*/
  display in green ""
  display in green "-------------------------------------------------------------"
  display in green " SOV              df       SS        MS        F        P"
  display in green "-------------------------------------------------------------"
  display in yellow" Between Groups" %6.0f df1[1] %10.2f ssb[1] %10.2f
ssb[1]/df1[1]
  %9.2f f[1] %12.4f p[1]
  display in yellow" Within Groups"  %7.0f df2[1] %10.2f sse[1] %10.2f
sse[1]/df2[1]
  display in green "-------------------------------------------------------------"
  display in green " Total"          %15.0f df[1] %10.2f  ssb[1]+sse[1]
  display in green "-------------------------------------------------------------"
  display in yellow" "
  display in yellow"Bartllet's test for equality of variances"
  display in green "Chi-Sq = " %4.3f b[1] " df = " %2.0f df1[1] "  P = " %5.4f
  pchi[1]
  /*drop all generated cols*/
  drop v1 - p
  end
  ------------------------------------------------------------------------------------
```

anovai.ado

References

1. Bartlett, M. S., (1937), *Properties of sufficiency and statistical tests*, Proceeding of Royal Society, Series A 160: 268 – 282.
2. Hanif, M., Ahmad M. and Ahmad A. M., (2004), *Biostatistics for Health Students*, ISOSS Publication.
3. Larson, David A., (1992), *Analysis of Variance with just Summary Statistics as Input*, American Statistician, 46, 151-152.
4. Moore, D. S. and McCabe, G. P., (1998), *Introduction to the practice of Statistics*, Freeman Publisher.