Chapter 6

Analysis of Variance

Aim: Upon completing the chapter, the learner should be able to perform an analysis of variance to compare the expected values of a continuous random variable between different groups.

6.1 Introduction

An analysis of variance (ANOVA) can be performed to compare two or more parameters (expected values and variances). The possible objectives of this analysis might be to:

1. Compare the expected value of a continuous random variable in "*m*" different groups to assess the following hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$$

2. Determine which expected values are different among comparison groups to evaluate any of the following potential null hypotheses:

$$H_0 : \mu_i = \mu_j ; H_0 : \mu_i = \frac{\mu_i + \mu_k}{2} ; H_0 : \frac{\mu_1 + \mu_2 + \mu_3}{3} = \frac{\mu_4 + \mu_5}{2} \cdots$$

3. Determine if the variability of a random continuous variable is the same between different groups to evaluate the following hypothesis:

$$H_0 : \sigma_\alpha^2 = 0$$

6.2 Data Structure

Assuming that we have three groups or three selected groups, the database to compare the expected value of random variable Y would have the following structure:

	Group 1	*Group 2*	*Group 3*	
	$Y_{1,1}$	$Y_{2,1}$	$Y_{3,1}$	
	$Y_{1,2}$	$Y_{2,2}$	$Y_{3,2}$	
	:	:	:	
	:	:	:	
	$Y_{1,n1}$	$Y_{2,n2}$	$Y_{3,n3}$	Total
Total	$\sum_{j=1}^{n1} Y_{1,j}$	$\sum_{j=1}^{n2} Y_{2,j}$	$\sum_{j=1}^{n3} Y_{3,j}$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{i,j}$
Mean	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}$
Variance	s_1^2	s_2^2	s_2^2	s^2
Number of subjects	n_1	n_2	n_3	n
Expected value	μ_1	μ_2	μ_3	μ

The possible research questions for this study are:

1. Assuming that the information available is from all possible groups, the research question can be stated as follows: Does the expected value vary by group ($\mu_1 = \mu_2 = \mu_3$)? (Fixed effects model)
2. Assuming that the information available is from a random sample of all possible groups, the research question can be stated as follows: Is there any variation among all the groups ($\sigma_\alpha^2 \neq 0$)? (Random effects model)

6.3 Example for Fixed Effects

To exemplify the research question for fixed effects (above), we are going to use the information from the previous database. As the *bmig* consists of three groups (normal, overweight, and obese), the research question would be the following: Are there differences in the expected age, according to the *bmig* categories?

The basic statistics from the database, using the *table* command, are in the following output:

```
. table bmig, c(n age mean age sd age)

------------------------------------------------
      bmig|     N(age)   mean(age)       sd(age)
----------+-------------------------------------
    Normal|          4       32.75      8.539125
Overweight|          3          27             2
     Obese|          3    31.66667      3.785939
------------------------------------------------
```

The results from the above table show that subjects having a normal bmi are in the older group, and those categorized as being overweight are younger. But the variability in these groups seems to be very different, based on the comparison of standard deviations. To assess whether these differences are statistically significant, either a linear model or an analysis of variance can be used (both of which are described in the following sections).

6.4 Linear Model with Fixed Effects

For a comparison of the expected values of a continuous random variable Y (e.g., age in years) for the groups of interest, we can establish the following model, using the bmig categories*:*

$$y_{ij} = \mu_1 + (\mu_2 - \mu_1)^* \mathrm{BMIG}_2 + (\mu_3 - \mu_1)^* \mathrm{BMIG}_3 + e_{ij}$$

and

$$y_{ij} = \mu_1 + \alpha_2{}^* \mathrm{BMIG}_2 + \alpha_3{}^* \mathrm{BMIG}_3 + e_{ij}$$

where:

- y_{ij} indicates the value of the continuous random variable Y in the ith subject that belongs to the jth bmig category
- μ_j indicates the expected value of Y in the jth *bmig* category, $E(y_{ij}) = \mu_j$
- e_{ij} indicates the difference between the observed value of y_{ij} and the expected value of the random variable Y in the jth bmig category (μ_j) (it is assumed that the errors, e_{ij}, are independent and follow an $N(0, \sigma^2)$ distribution)
- α_j indicates the effect of the jth *bmig* category with respect to the first bmig category (normal), subject to the restriction $\alpha_1 = 0$
- BMIG_j is a dummy variable whose value is 1 if the subject belongs to the jth bmig category; its value is 0 if the subject belongs to another bmig group

When the groups being compared correspond to all of the possible groups or when they represent a select group of interest, the α_is are constants and are defined as fixed effects. If the effects are fixed, then it is initially assumed that variances within groups are equal, $\text{Var}(Y_{ij}) = \sigma^2$.

6.5 Analysis of Variance with Fixed Effects

For analyzing the expected values of a continuous random variable between different groups using the statistical method of analysis of variance, we start by decomposing the numerator of the variance of Y, as can be accomplished using the following:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2=\sum_{i=1}^{k}n_i\left(\bar{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2$$

where:

$\sum_{i=1}^{k} n_i\left(\bar{Y}_i-\bar{Y}\right)^2$ indicates the variation between groups (*between sum of squares*)

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2$ indicates the overall variation within each group (*within sum of squares*)

The null hypothesis in ANOVA with fixed effects determines that the expected values of the random variable of interest, Y, in all groups are the same, $H_0: \mu_1 = \cdots = \mu_k$; thus, $\alpha_i = 0$ for all groups. To assess the null hypothesis, the estimated expected values of SS between and SS within are compared, considering their respective degrees of freedom, as follows:

Source of Variation	*SS*	*Df*	*MS*	*E*[*MS*]
Between	$\sum_{i=1}^{k} n_i\left(\bar{Y}_i-\bar{Y}\right)^2$	$k - 1$	$\sum_{i=1}^{k} n_i\left(\bar{Y}_i-\bar{Y}\right)^2 / (k-1)$	$\sigma^2+\phi$
Within	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2$	$n - k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2 \big/ (n-k)$	σ^2
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2$	$n - 1$		

Note: $\phi = \dfrac{1}{k-1}\sum_{i=1}^{k} n_i\alpha_i^2$

Under the null hypotheses, the ratio $\left(\sigma^2 + \frac{1}{k-1}\sum n_i \alpha_i^2 / \sigma^2\right) = 1$. To determine how far this ratio should be away from 1, once a dataset is collected and the parameters of the linear model $\left(\sigma^2, \alpha_i\right)$ are estimated, a P-value is computed using the F-Fisher probability distribution with 1 and n – 2 degrees of freedom.

6.6 Programming for ANOVA

To program a linear model to determine which fixed effect is different from zero, we can use the *regress* (or *reg*) command with a categorical predictor. For example, to compare age between the categories of the bmig, using the first category as the reference group, go to the *Statistics* dropdown menu and click on "Linear regression"; next, write *age* in the Dependent variable box and *bmig* (as a categorical variable) in the Independent variables box, remembering to precede "bmig" with the symbol "*i.*", as is illustrated in Figure 6.1.

Once the previous table is submitted, the following output will be displayed:

```
. reg  age i.bmig

     Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(2, 7)         =      0.83
      Model |  60.6833333         2  30.3416667   Prob > F        =    0.4742
   Residual |  255.416667         7  36.4880952   R-squared       =    0.1920
-------------+----------------------------------   Adj R-squared   =   -0.0389
      Total |       316.1         9  35.1222222   Root MSE        =    6.0405
```

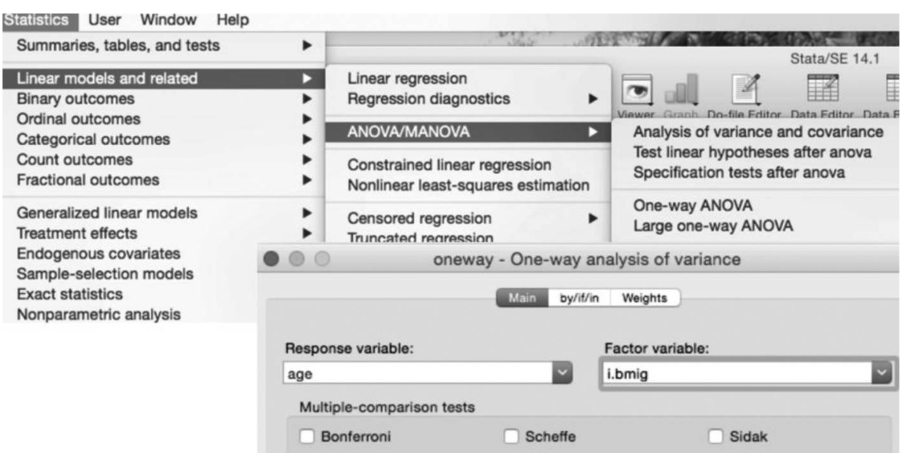

Figure 6.1 Linear regression model with categorical predictor.

```
------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bmig |
Overweight   |      -5.75   4.613537    -1.25   0.253    -16.65928    5.159281
     Obese   |  -1.083333   4.613537    -0.23   0.821    -11.99261    9.825948
             |
       _cons |      32.75   3.020269    10.84   0.000      25.6082     39.8918
------------------------------------------------------------------------------
```

The results show that there is no evidence of significant differences in the mean age across bmig categories using normal subjects as the reference group (P-value > .1). If the user wants to change the reference group (e.g., use the second category of bmig), the following command syntax should be used:

```
reg  age b2.bmig
```

Output

```
. reg  age b2.bmig

    Source |       SS       df       MS              Number of obs =      10
-----------+------------------------------           F(2, 7)       =    0.83
     Model |  60.6833333     2  30.3416667           Prob > F      =  0.4742
  Residual |  255.416667     7  36.4880952           R-squared     =  0.1920
-----------+------------------------------           Adj R-squared = -0.0389
     Total |       316.1     9  35.1222222           Root MSE      =  6.0405

------------------------------------------------------------------------------
       age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      bmig |
   Normal  |       5.75   4.613537     1.25   0.253    -5.159281    16.65928
    Obese  |   4.666667   4.932078     0.95   0.376    -6.995845    16.32918
           |
     _cons |         27   3.487506     7.74   0.000     18.75336    35.24664
------------------------------------------------------------------------------
```

The commands *oneway* and *anova* can be used in the assessment of the null hypothesis, $H_0 : \mu_1 = \cdots = \mu_k$ (using fixed-effect ANOVA). The *oneway* command includes Bartlett's test for equal variances, a condition needed in the F-test for comparing expected values. The *anova* command expands the sum of squares if more than one source of variation is used. For example, to compare age between the bmig categories, the following command is used:

```
oneway age bmig
```

Output

```
. oneway age bmig

                       Analysis of Variance
    Source              SS         df       MS         F     Prob > F
------------------------------------------------------------------------
Between groups   60.6833333      2    30.3416667   0.83     0.4742
 Within groups   255.416667      7    36.4880952
------------------------------------------------------------------------
   Total              316.1      9    35.1222222

Bartlett's test for equal variances: chi2(2) = 3.5156 Prob>chi2 = 0.172
```

The output of the *oneway* command provides evidence in favor of the null hypothesis, $H_0 : \mu_{\text{normal}} = \mu_{\text{overweight}} = \mu_{\text{obese}}$, and evidence in favor of equal variances, via the Bartlett's test (P-value > 0.1).

If the user includes the variable *sex* as a second source of variation and the interaction of *age* and *sex* to explore how the mean of *age* changes by bmig category and *sex*, the command line (making use of the *anova* command) is as follows:

```
anova age bmig sex bmig#sex
```

Output

```
         Number of obs =         10     R-squared     =  0.8439
         Root MSE      =    3.14113     Adj R-squared =  0.7191

   Source | Partial SS         df          MS          F     Prob>F
---------+----------------------------------------------------------
    Model |  266.76667          4    66.691667       6.76    0.0299
          |
     bmig |  157.11373          2    78.556863       7.96    0.0279
      sex |  131.92157          1    131.92157      13.37    0.0146
 bmig#sex |  62.745098          1    62.745098       6.36    0.0531
          |
 Residual |  49.333333          5    9.8666667
---------+----------------------------------------------------------
    Total |      316.1          9    35.122222
```

The output suggests that the mean of *age* changes according to the *bmig* and *sex* categories due to the fact that the interaction term bmig#sex is marginally significant (P-value = .053); however, caution should be taken with this interpretation because the sample size is very small.

6.7 Planned Comparisons (before Observing the Data)

Having determined that there is evidence of a difference between the expected values, the next step is to determine if the difference of interest is significant. This difference of interest can be defined with two expected values (pairs of means) or with a combination of expected values (linear contrasts).

6.7.1 Comparison of Two Expected Values

Continuing with the same example as before, suppose the user's main purpose is to compare the expected age of subjects having a normal bmi with the expected age of subjects who are categorized as being overweight. To do this, the null hypothesis is formulated in the following way:

$$H_0 : \mu_1 = \mu_2 \quad \text{or} \quad H_0 : \mu_1 - \mu_2 = 0$$

which is equivalent to

$$H_0 : \alpha_2 = 0$$

The test statistic will be

$$F = t^2 = \left(\frac{\left| \bar{Y}_i - \bar{Y}_j \right|}{\sqrt{s^2 \left[\left(1/n_i \right) + \left(1/n_j \right) \right]}} \right)^2 \sim F_{1, n-k|H_0}$$

where s^2 is MSE (within MS).

To evaluate the null hypothesis in Stata, using this statistic, we can use the *test* command after the *anova* command, as can be seen in the following:

```
anova age bmig
```

```
test 1.bmig=2.bmig
```

Output

```
. anova age bmig

            Number of obs =        10     R-squared     =  0.1920
            Root MSE      = 6.04054     Adj R-squared = -0.0389

     Source | Partial SS  df          MS       F          Prob>F
 -----------+---------------------------------------------------
      Model |  60.683333   2    30.341667    0.83         0.4742
            |
       bmig |  60.683333   2    30.341667    0.83         0.4742
            |
   Residual |  255.41667   7    36.488095
 -----------+---------------------------------------------------
      Total |      316.1   9    35.122222
```

```
.  test 1.bmig=2.bmig

 ( 1)  1b.bmig - 2.bmig = 0

       F(  1,      7) =    1.55
            Prob > F =    0.2527
```

The results suggest that there is no difference in the expected age between normal bmi subjects and those whose bmi indicates that they are overweight (P-value > .1).

6.7.2 Linear Contrast

A linear contrast is a combination of expected values, as is illustrated in the following:

$$L = \sum_{i=1}^{k} c_i \mu_i$$

where:

$$\sum_{i=1}^{k} c_i = 0$$

The definition for a linear contrast depends on the null hypothesis under evaluation. For example,

1. When $H_0 : \mu_1 - \mu_2 = 0$ then,

   $$L = (1)\mu_1 + (-1)\mu_2$$

   where $c_1 = 1$ and $c_2 = -1$

2. When $H_0 : \mu_1 = (\mu_2 + \mu_3/2) \Rightarrow H_0 : (1)\mu_1 + (-0.5)\mu_2 + (-0.5)\mu_3 = 0$ then,

   $$L = (1)\mu_1 + (-0.5)\mu_2 + (-0.5)\mu_3$$

   where $c_1 = 1$, $c_2 = -0.5$, $c_3 = -0.5$

Once the linear contrast is defined, the test statistic is

$$t = \frac{\hat{L}}{\sqrt{s^2 \sum_{i=1}^{k} \left(c_i^2 / n_i\right)}} \sim t_{\left(n-k, 1-\frac{\alpha}{2}\right)}$$

where s^2 = MSE (within MS) and $\hat{L}$ is determined by the sample means.

To evaluate linear contrasts in Stata, we can use the *test* command after the *anova* command. For example, assuming the user wants to compare the expected age between the subjects categorized as having a normal bmi against the average of the expected age in subjects categorized as being overweight or obese, the null hypothesis is formulated as H_0: $\mu_1 = (\mu_2 + \mu_3)/2$. To assess this hypothesis with the *test* command, the specified command line, after running the *anova* command, is as follows:

```
test 1.bmig=(2.bmig+3.bmig)/2
```

Output

```
. test 1.bmig=(2.bmig+3.bmig)/2
 ( 1)  1b.bmig - .5*2.bmig - .5*3.bmig = 0
       F(  1,     7) =    0.77
            Prob > F =    0.4099
```

The results suggest that the expected age does not change between the groups under comparison (P-value > .1).

6.8 Multiple Comparisons: Unplanned Comparisons

Having determined that there is evidence of a difference between the expected values, the next step is to determine which of the differences is or are significant. There are several methods, called post hoc tests, which have been developed to answer this type of question; two of the most commonly used methods are *Bonferroni's method* and *Scheffé's method.*

Bonferroni's method compares pairs of groups by adjusting the significance level of each pair of averages compared to the total possible number of paired comparisons. For example, if the level of significance is 5% and there are three possible pairwise comparisons, then the level of significance for evaluating one particular pair of means is divided by 3; that is, the level of significance of one pair would be 0.05/3 = 0.0167. Another alternative is to multiply the P-value of each comparison by 3. This method ensures that the overall significance level defined in ANOVA is maintained when it is conducted simultaneously on all possible pairwise comparisons. This method can be programmed through the *oneway* command. For example, assuming the user wants to compare the expected age in the following three bmig categories (normal, overweight, and obese), the command syntax with *oneway* command will be the following:

```
oneway age bmig, bon tab
```

Output

```
           |             Summary of age
      bmig |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
    Normal |       32.75   8.5391256           4
 Overweigh |          27           2           3
     Obese |   31.666667   3.7859389           3
-----------+------------------------------------
     Total |        30.7   5.9264004          10
```

```
                  Analysis of Variance
   Source             SS          df       MS          F     Prob > F
------------------------------------------------------------------------
Between groups    60.6833333      2    30.3416667   0.83     0.4742
 Within groups    255.416667      7    36.4880952
------------------------------------------------------------------------
   Total               316.1      9    35.1222222

Bartlett's test for equal variances: chi2(2) = 3.5156 Prob>chi2 = 0.172

                         Comparison of age by bmig
                               (Bonferroni)
Row Mean-|
Col Mean |      Normal   Overweig
---------+---------------------
Overweig |       -5.75
         |       0.758
         |
   Obese |    -1.08333    4.66667
         |       1.000      1.000
```

Note: The option Bon displays the Bonferroni multiple-comparison test. The option tab produces a summary of age at each category of bmig.

At the bottom of the output table, all the pairwise comparisons between the samples means can be seen; for example, the difference between the mean age of the obese group and that of the overweight group is 4.67, which can be confirmed with the first table requested in this output, in which $\bar{Y}_{\text{normal}} = 32.8$, $\bar{Y}_{\text{overweight}} = 27.0$, and $\bar{Y}_{\text{obese}} = 31.7$. Below the pairwise differences is the *P*-value for the *F*-statistics of one pair, which was computed by multiplying by the total number of possible mean pairs to be compared. The results show that there are no evidences of significant differences (*P*-values > .1) in any of the pairwise comparisons.

Scheffé's method performs multiple comparisons through linear contrasts, but the significance level of each comparison is not adjusted. In Scheffé's method, the test statistic is calculated using the following formula:

$$t^2/(k-1) = \left(\frac{\hat{L}}{\sqrt{s^2 \sum_{i=1}^{k} \left(c_i^2/n_i\right)}}\right)^2 / (k-1)$$

The null hypothesis is rejected with a certain significance level, α, when

$$t > \sqrt{(k-1)F_{k-1,n-k,1-\alpha}} \quad \text{or} \quad t < -\sqrt{(k-1)F_{k-1,n-k,1-\alpha}}$$

The P-value is calculated with the F-Fisher probability distribution with k – 1 and n – k degrees of freedom. The Stata command line for performing multiple comparisons by Scheffé's method is as follows:

```
oneway age bmig, sch
```

Output

```
**The ANOVA table is omitted

Comparison of age by bmig
                                   (Scheffe)
Row Mean-|
Col Mean |      Normal   Overweig
---------+---------------------
Overweig |       -5.75
         |       0.496
         |
   Obese |    -1.08333    4.66667
         |       0.973      0.656
```

The results displayed in the above output table show that the P-values that result when using Scheffé's method are different from those that result when Bonferroni's method is used; however, the statistical evidence confirms that there are no significant differences (P-values > .1) in any of the pairwise comparisons.

6.9 Random Effects

When the statistical information is collected only from a random sample of groups of subjects, which are part of all possible study groups (Figure 6.2), we can define a linear model with random effects.

The linear model with random effects is represented as follows:

```
Yij|µi = µi + eij,
```

where:

$Y_{ij}|\mu_i \sim$ i.i.d $N\left(\mu_i, \sigma^2\right)$

$\mu_i \sim$ i.i.d $N\left(\mu_i, \sigma_\mu^2\right)$

i.i.d = independently and identically distributed

The main outcome, Y, is a continuous random variable distributed as a normal distribution with the following parameters: μ_i,σ^2. These parameters could be also random variables; however, in the random effect model considered, only μ_i is assumed to be a random variable with normal distribution and parameters: μ, σ_μ^2.

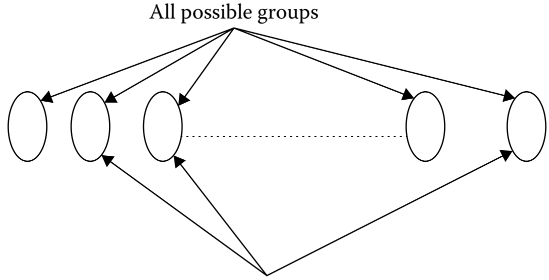

Figure 6.2 Random effects scheme.

Following is an alternative expression of the linear model with random effects:

$$E\left(Y_{ij}|\alpha_i\right) = \mu + \alpha_i$$

$$Y_{ij}|\alpha_i \sim \text{i.i.d}\, N\left(\mu_i, \sigma^2\right)$$

$$\alpha_i \sim \text{i.i.d}\, N\left(0, \sigma_\alpha^2\right)$$

The procedure to estimate the parameters of this model is similar to the process used in Bayesian data analysis; however, in ANOVA with random effects, we are only assuming randomness in μ_i. The null hypothesis of the ANOVA with random effects is formulated as follows:

$$H_0 : \sigma_\alpha^2 = 0$$

Having $\sigma_\alpha^2 = 0$ implies that the expected values of Y for all groups, including the groups in the study sample and the groups of subjects who were not included in the study sample, are equal $\left(\mu_1 = \mu_2 = \cdots = \mu_m\right)$.

Under the assumption of random effects, the expected values for the sum of squares in the ANOVA can be seen below in the following:

Source of Variation	*SS*	*Df*	*MS*	*E[MS]*
Between	$\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{Y}\right)^2$	$k - 1$	$\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{Y}\right)^2 / (k-1)$	$\sigma^2 + n_0\sigma_\alpha^2$
Within	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$	$n - k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 / (n-k)$	σ^2
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2$	$n - 1$		

where:

$$n_0 = \frac{\sum_{i=1}^{k} n_i - \left[\left(\sum_{i=1}^{k} n_i^2 \bigg/ \sum_{i=1}^{k} n_i\right)\right]}{k-1}$$

According to the null hypothesis $\left(H_0 : \sigma_\alpha^2 = 0\right)$,

$$\frac{E\left(\text{MS}_{\text{Between}}\right)}{E\left(\text{MS}_{\text{Within}}\right)} = 1$$

To evaluate the null hypothesis with the data collected in the sample, the *F*-statistic is obtained using the following equation:

$$F = \frac{\sum_{i=1}^{k} n_i\left(\bar{Y}_i - \bar{Y}\right)^2 / (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 / (n-k)} \sim F_{(k-1,n-k,1-\alpha)}$$

The estimates of σ_α^2 and σ^2 are calculated by the following expressions:

$$\widehat{\sigma_\alpha^2} = \max\left(\frac{\text{MS}_{\text{Between}} - \text{MS}_{\text{Within}}}{n_0}, 0\right)$$

$$\hat{\sigma}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\left(Y_{ij} - \bar{Y}_i\right)^2}{(n-k)}$$

6.10 Other Measures Related to the Random Effects Model

6.10.1 Covariance

In the previous example, there is a possible relationship between the subjects of the same group. To assess this, the covariance statistic is defined using the following equation:

$$\mathrm{Cov}\left(Y_{im},Y_{in}\right)=\mathrm{Cov}\{E(Y_{im}|\alpha_i,E\left(Y_{in}|\alpha_i\right)\}+E\left\{\mathrm{Cov}\left(Y_{im},Y_{in}|\alpha_i\right)\right\}$$
$$=\mathrm{Cov}\left(\mu+\alpha_i,\mu+\alpha_i\right)+0=\mathrm{Cov}\left(\alpha_i,\alpha_i\right)=\sigma_\alpha^2$$

6.10.2 Variance and Its Components

Another expression for the variance of Y_{ij} with its components is as follows:

$$\mathrm{Var}\left(Y_{ij}\right)=\mathrm{Var}\left\{E\left(Y_{ij}|\alpha_i\right)\right\}+E\left\{\mathrm{Var}\left(Y_{ij}|\alpha_i\right)\right\}$$
$$=\mathrm{Var}\left(\mu+\alpha_i\right)+E\left(\sigma^2\right)=\sigma_\alpha^2+\sigma^2$$

where σ_α^2 and σ^2 are defined as components of the variance.

6.10.3 Intraclass Correlation Coefficient

A useful measure of association within groups is the intraclass correlation coefficient defined as follows:

$$\mathrm{Corr}\left(Y_{ij},Y_{il}\right)=\frac{\mathrm{Cov}\left(Y_{ij},Y_{il}\right)}{\sqrt{\mathrm{Var}\left(Y_{ij}\right)\mathrm{Var}\left(Y_{il}\right)}}=\frac{\sigma_\alpha^2}{\sqrt{\left(\sigma_\alpha^2+\sigma^2\right)\left(\sigma_\alpha^2+\sigma^2\right)}}=\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma^2}=\rho_I$$

This index can be interpreted as a measure of reproducibility or consistency (reliability coefficient) between the measurements of a group. Rosner (2010) suggests using the following interpretations:

$\rho_I < 0.4$	Poor reproducibility
$0.4 \le \rho_I < 0.75$	Moderate reproducibility
$\rho_I \ge 0.75$	Excellent reproducibility

6.11 Example of a Random Effects Model

Let us assume that the user is interested in assessing the systolic blood pressure readings (mm Hg) from a portable machine. In addition, we must take into consideration the fact that the experimental design was defined to measure these readings from a subject twice a day, for 10 consecutive days, as a random sample of days in 1 month. Finally, let us assume that the data in Stata conform to the following format:

```
     +-----------------+
     | day   sb1   sb2 |
     |-----------------|
  1. |   1    98    99 |
  2. |   2   102    93 |
  3. |   3   100    98 |
  4. |   4    99   100 |
  5. |   5    96   100 |
  6. |   6    95   100 |
  7. |   7    90    98 |
  8. |   8   102    93 |
  9. |   9    91    92 |
 10. |  10    90    94 |
     +-----------------+
     sb1 indicates the first measure of systolic blood pressure.
     sb2 indicates the second measure of systolic blood pressure.
```

To perform an ANOVA, the user has to modify the previous database structure. The actual format is called *wide* format, where every row in the dataset contains all the information of one subject. To run an ANOVA the database structure must be in the *long* format, where every row contains the information of each subject's visit. The ***reshape*** command can be used to change the database structure from *wide* to *long* format, as follows:

```
reshape long sb, i(day)
```

Output

```
. reshape long sb, i(day)
(note: j = 1 2)

Data                               wide   ->   long
-----------------------------------------------------------------
Number of obs.                       10   ->      20
Number of variables                   3   ->       3
j variable (2 values)                     ->   _j
xij variables:
                                sb1 sb2   ->   sb
-----------------------------------------------------------------
```

After the *reshape* command, use the *list* command to see the current data structure, as is demonstrated in the following table:

```
list
      +----------------+
  i   | day   _j     sb |
      |----------------|
  1.  |   1    1     98 |
  2.  |   1    2     99 |
  3.  |   2    1    102 |
  4.  |   2    2     93 |
  5.  |   3    1    100 |
  6.  |   3    2     98 |
  7.  |   4    1     99 |
  8.  |   4    2    100 |
  9.  |   5    1     96 |
 10.  |   5    2    100 |
 11.  |   6    1     95 |
 12.  |   6    2    100 |
 13.  |   7    1     90 |
 14.  |   7    2     98 |
 15.  |   8    1    102 |
 16.  |   8    2     93 |
 17.  |   9    1     91 |
 18.  |   9    2     92 |
 19.  |  10    1     90 |
 20.  |  10    2     94 |
      +----------------+
```

To run the linear model with random effects, use the *loneway* command, as follows:

```
loneway sb day
```

Output

```
       One-way Analysis of Variance for sb:

                                          Number of obs =          20
                                              R-squared =      0.5118

  Source                  SS          df       MS            F      Prob > F
-------------------------------------------------------------------------
Between day              152          9    16.888889       1.16      0.4050
Within day               145         10         14.5
-------------------------------------------------------------------------
Total                    297         19    15.631579

       Intraclass        Asy.
       correlation       S.E.       [95% Conf. Interval]
       ------------------------------------------------
          0.07611      0.32301       0.00000      0.70920

       Estimated SD of day effect                1.092906
       Estimated SD within day                   3.807887
       Est. reliability of a day mean             0.14145
            (evaluated at n=2.00)
```

The results indicate that there is evidence in favor of the null hypothesis $\left(H_0 : \sigma_\alpha^2 = 0\right)$. Therefore, the expected values of the systolic blood pressure readings from the portable machine are equal for the subject under study for 1 month (P-value > .1). The estimated intraclass correlation coefficient is as follows:

$$\widehat{\rho_I} = \frac{1.19}{1.19 + 14.5} = 0.076$$

where $\hat{\sigma}_\alpha^2 = \left(16.9 - 14.5/2\right) = 1.19$. Therefore, based on the intraclass correlation coefficient estimation, there is a poor reproducibility index between the two measurements taken in a single day.

6.12 Sample Size and Statistical Power

To determine the minimum sample size for comparing means with enough statistical power (i.e., at least $1 - \beta = 0.8$) and a minimum significance level (i.e., $\alpha = 0.05$), Stata provides the option for comparing the means of independent samples using ANOVA with the *Power and sample size analysis* window in the *Statistics* menu. For example, assuming we want to determine the minimum sample size to compare their mean age across the three bmig categories, the following information must be provided in the option *One-way analysis of variance*: significance level, statistical power, the group of means under the alternative hypothesis, and the error (within-group) variance (Figure 6.3).

Once the previous table is submitted, the following output will be displayed:

```
. power oneway 32.8  27 31.7, varerror(36.5)

Performing iteration ...

Estimated sample size for one-way ANOVA
F test for group effect
Ho: delta = 0  versus  Ha: delta != 0

Study parameters:

       alpha =    0.0500
       power =    0.8000
       delta =    0.4163
         N_g =         3
          m1 =   32.8000
          m2 =   27.0000
          m3 =   31.7000
       Var_m =    6.3267
       Var_e =   36.5000
```

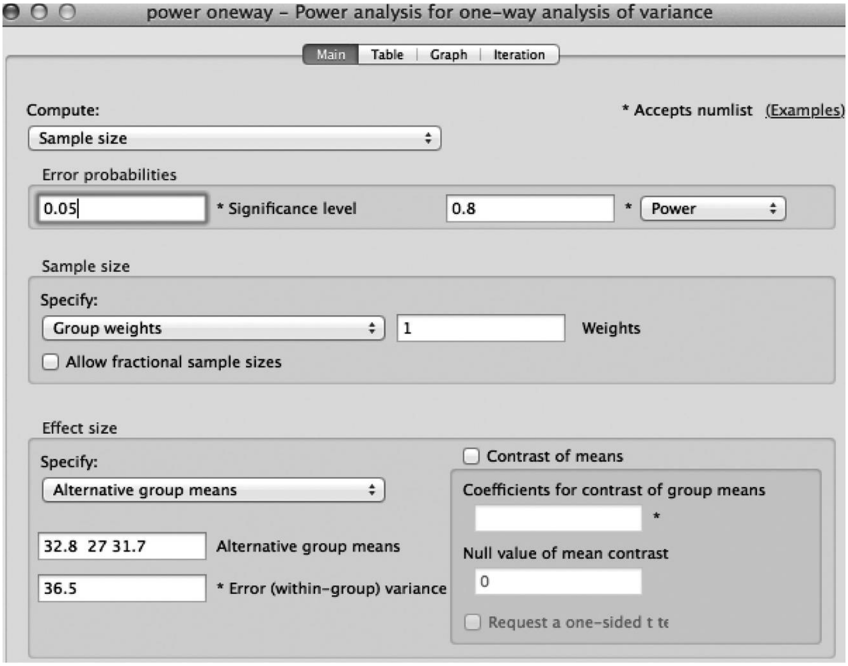

Figure 6.3 Sample size for one-way ANOVA.

```
Estimated sample sizes:

         N =          60
N per group =          20
```

Both delta (effect design, $\sqrt{\text{Var\_m}}/\sqrt{\text{Var\_e}}$) and Var_m (variance between groups, based on the means of each group and the grand mean, $\sum_{i=1}^{3}\left[\left(\bar{Y}_i-\bar{Y}\right)^2/3\right]$) are computed automatically by Stata.

Therefore, the minimum sample size to compare the means across the three bmig categories is 20 subjects per group, assuming the following conditions: an 80% statistical power, a 5% significance level, and an error (within-group) variance of 36.5.