The text describes the analysis of variance in terms of the "extra sum of squares" principle. This is a useful concept with many applications, and will be demonstrated on two applications to the Spock trial example. This application is described below.

We define a FULL model – a model that. in the first case, fits a separate mean to each Judge in the District court. This model would require the estimate of 7 parameters (one mean for each judge) before the estimate of variance would be done. This requires a total of 7 degrees of freedom.

The REDUCED model would fit all judges to the same mean. That is, only a single mean would be fitted. This would require only one degree of freedom to be fitted prior to estimating the variance. The variance estimated for this reduced model is simply the usual variance for a single set of data. We can estimate this with PROC UNIVARIATE.

```
Chapter 5: Spock Conspiracy Trial
Proc univariate for all data combined

The UNIVARIATE Procedure
Variable:  Percent

Moments
N                          46    Sum Weights                46
Mean                26.5826087    Sum Observations         1222.8
Std Deviation       9.17911408    Variance             84.2561353
Skewness            0.03841523    Kurtosis             -0.0346302
Uncorrected SS       36296.74    Corrected SS          3791.52609
Coeff Variation      34.530524    Std Error Mean       1.35338654
```

In order to do the analysis of extra sum of squares we need the corrected SS from this model (corrected for a single mean fitted to all Judges). This is provided by the value "Corrected SS" from proc univariate. This value is the "unexplained" variance for a model with a single mean. The model is $Y_i=\mu+\varepsilon_i$. This model will require only 1 d.f. where n-1=46 d.f. for error.

The full model will require the corrected SS adjusted for the mean of each individual judge. Since Analysis of Variance fits a mean to each judge we can use the error SS from ANOVA. The model is $Y_{ij}=\mu_i+\varepsilon_{ij}$ where i=1 to 7. This model will require only 7 d.f. where n-7=39 d.f. for error. From the ANOVA we need the corrected SS for the error.

```
Chapter 5 : Spock Conspiracy Trial
Analysis of variance with PROC GLM

The GLM Procedure

Dependent Variable: Percent
                              Sum of
Source                DF      Squares    Mean Square   F Value    Pr > F
Model                  6  1927.080865     321.180144      6.72    <.0001
Error                 39  1864.445222      47.806288
Corrected Total       45  3791.526087   (Var=3791.526087/45=84.25613527)

R-Square     Coeff Var      Root MSE    Percent Mean
0.508260      26.01027      6.914209        26.58261
```

To test for a difference we use the General Linear Hypothesis Test. This test uses the calculation of the "extra sum of squares" and can be used to test for differences between any two linear models where the terms in one model are a subset, or reduced version, of the terms in the other model.

General Linear Hypothesis test (GLHT)

Source	**d.f.**	**SS**	**MS**	**F**
Reduced model error	**n–1 = 45**	**3791.52609**		
Full model error	**n–1– (t–1) = 39**	**1864.44522**		
Difference	**t–1 = 6**	**1927.080865**	**321.180144**	**6.72**
Full model error	**39**		**47.806288**	

From Excel, P>F = 0.000060825

This is a general and useful way to test between full and reduced models in numerous situations. However, in this case we see that all of these components were already given in the ANOVA source table, making this particular calculation unnecessary. Essentially what we have done is the usual ANOVA test expressed in the form of the GLHT.

The following is a more interesting application of Extra SS. We have seen that there are differences among the Judges. We ask, are the differences due just to the Dr. Spock judge being different from the others, or are there differences among the others as well?

To do this test we need again need two models. The full model fits a separate mean to all judges. This model is the ANOVA that we already have, and it had 39 d.f.. The reduced model will distinguish only the Spock judge from all others. Note that this is essentially a t-test. We will fit 2 means, one for the Spock judge and one for all others, this will require 2 d.f. where n-2=44 d.f. for error.

To fit this model I created a new variable called JUSTSPOCK.

```
data Jury; set jury;
  if judge = 'SPOCK' then JustSpock = 'SPOCK' ; else JustSpock = 'Other';
run;
```

The analysis of variance for this model is given below (GLM only).

```
Chapter 5 : Spock Conspiracy Trial
Analysis of SPOCK versus others
Analysis of variance with PROC GLM

The GLM Procedure

Dependent Variable: Percent
                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        1     1600.622964     1600.622964      32.15    <.0001
Error                       44     2190.903123       49.793253
Corrected Total             45     3791.526087
```

General Linear Hypothesis test

Source	**d.f.**	**SS**	**MS**	**F**
Reduced model error	**44**	**2190.903123**		
Full model error	**39**	**1864.445222**		
Difference	**5**	**326.457901**	**65.291580**	**1.365753**
Full model error	**39**	**1864.445222**	**47.806288**	

From Excel, P>F = 0.258179356

The so called "extra SS" is the SS difference between a reduced model and a full model. In the first example the extra SS was **3791.52609 – 1864.44522 = 1927.080865** where this difference had 6 d.f. This test is the usual ANOVA, testing the hypothesis that H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ versus the alternative that some μ_i is different.

The second test was one we have not seen before, testing just for differences between the "other judges". The reduced model fits two means, giving the difference between the Spock judge and the other judges. The full model fits 7 means, so it adds the difference among the other judges to the model and the extra SS fits the difference, which is the variability only among the other judges (since the Spock judge difference is already fitted).

There are two categories of sums of squares that will be of interest for ANOVA and regression. One SS type is called the TYPE I SS or the sequentially adjusted SS. The other is the TYPE III SS or the fully adjusted SS. A more thorough discussion of these two types of SS will be given in the discussion of regression. For the moment we will simply describe the two types of sums of squares as TYPE I, which is fitted with each variable entering the model in a predetermined order. Variables entering the model first are unadjusted for variables entering the model later, while the variables entered late are adjusted for variables entered previously. The TYPE III or partial SS is the SS that each variable would have if it entered the model last and was fully adjusted for all other variables in the model.

These SS can be used to test full and reduced models by putting both the "reduced" model and the "full" model in the same model. A sum of squares Type I will fit the variables in order, so the difference can be observed. Compare the two types of SS below to the results of the previous GLHT.

```
The GLM Procedure

           Class Level Information
Class          Levels    Values
JustSpock           2    Other SPOCK
Judge               7    A B C D E F SPOCK'S
Number of Observations Read          46
Number of Observations Used          46

Dependent Variable: Percent

                                    Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        6     1927.080772      321.180129       6.72    <.0001
Error                       39     1864.445255       47.806289
Corrected Total             45     3791.526027

R-Square     Coeff Var      Root MSE    Percent Mean
0.508260      26.01027      6.914209        26.58261

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
JustSpock                    1     1600.622903     1600.622903      33.48    <.0001
Judge                        5      326.457869       65.291574       1.37    0.2582

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
JustSpock                    0       0.0000000               .          .         .
Judge                        5     326.4578692      65.2915738       1.37    0.2582
```

Additional notes on the GLHT.

This is an intuitive test because we wish to test for a difference in two models and we literally calculate a difference between those models and test that difference. The test requires the degrees of freedom and calculation of a mean square. The mean square is a variance estimate, and tests of variances are done as an F test.

General Linear Hypothesis test

Source	d.f.	SS	MS	F
Reduced model error	**44**	**2190.903123**		
Full model error	**39**	**1864.445222**		
Difference	**5**	**326.457901**	**65.291580**	**1.365753**
Full model error	**39**	**1864.445222**	**47.806288**	

From Excel, P>F = 0.258179356

Once we determine the MS difference we need an error term to test with. We actually have two error terms, one for the full model and one for the reduced model. Why do I choose the full model error over the reduced model error? Consider two questions. Which error is better if there is no difference in the models and which error is better if there is a difference in the models?

The General Linear Hypothesis Test is a very general and flexible application. It can be used in both Analysis of Variance and Regression to test for differences in any two models where one model is a subset of the other. The hypotheses tested above represent the following set of models.

All judges equal	$Y_i=\mu+\varepsilon_i$ where and i = 1 to n
Full Spock different from others	$Y_{ij}=\mu_i+\varepsilon_{ij}$ where i=1 and 2 and j = 1 to n_i
All judges different	$Y_{ij}=\mu_i+\varepsilon_{ij}$ where i=1 to 7 and j = 1 to n_i