

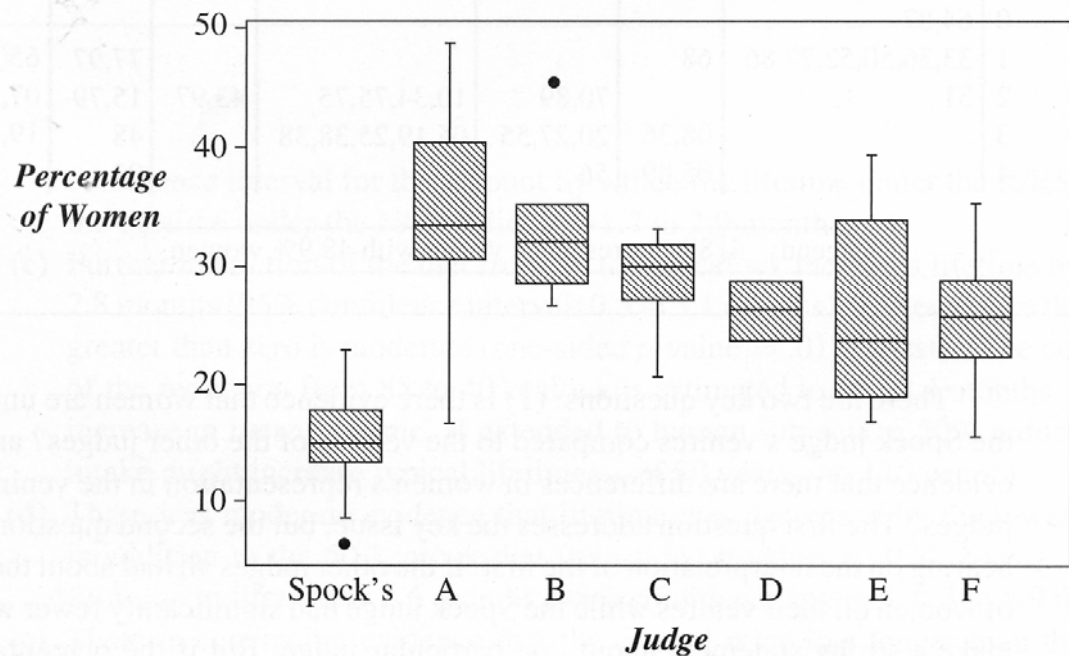
Analysis of Variance [Chapter 5, part 2]

A second case of analysis of variance, the Dr. Spock conspiracy trial.

This case is an observational study, so the data does not come from a planned experiment, conducted under controlled conditions.

The claim by the defense is that the number of women (who might favor Dr. Spock) was underrepresented. In fact, his jury had no women. There were 7 U. S. District Court judges in the Boston area, including Dr. Spock's judge. The null hypothesis is that the mean number of women for the 7 judges is equal ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$) versus the alternative (some μ_i is different).

Display 5.5 Percentages of women on venires of the seven Boston judges



Although an observational study, the analysis is the same. Using PROC MIXED and PROC GLM to do our Analysis of Variance we get the following.

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

The GLM Procedure

```

Class Level Information
Class          Levels  Values
Judge          7      A B C D E F SPOCK

Number of Observations Read          46
Number of Observations Used          46

```

Dependent Variable: Percent

Source	DF	Sum of 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			

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
Judge	6	1927.080865	321.180144	6.72	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Judge	6	1927.080865	321.180144	6.72	<.0001

What conclusion can be made from these results? Clearly the F value of 6.72 would be unusual under the null hypothesis, and would occur by random chance with a probability of less than one in 10,000. The null hypothesis ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$) would be rejected for the alternative (some μ_i is different).

For these relatively simple problems both PROC MIXED and PROC GLM should give the same results.

Chapter 5 : Spock Conspiracy Trial
Analysis of variance with PROC MIXED

The Mixed Procedure

Model Information	
Data Set	WORK.JURY
Dependent Variable	Percent
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	Levels	Values
Judge	7	A B C D E F SPOCK

Dimensions	
Covariance Parameters	1
Columns in X	8
Columns in Z	0
Subjects	1
Max Obs Per Subject	46

Number of Observations	
Number of Observations Read	46
Number of Observations Used	46
Number of Observations Not Used	0

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	47.8063

Fit Statistics

-2 Res Log Likelihood	274.0
AIC (smaller is better)	276.0
AICC (smaller is better)	276.1
BIC (smaller is better)	277.6

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Judge	6	39	6.72	<.0001

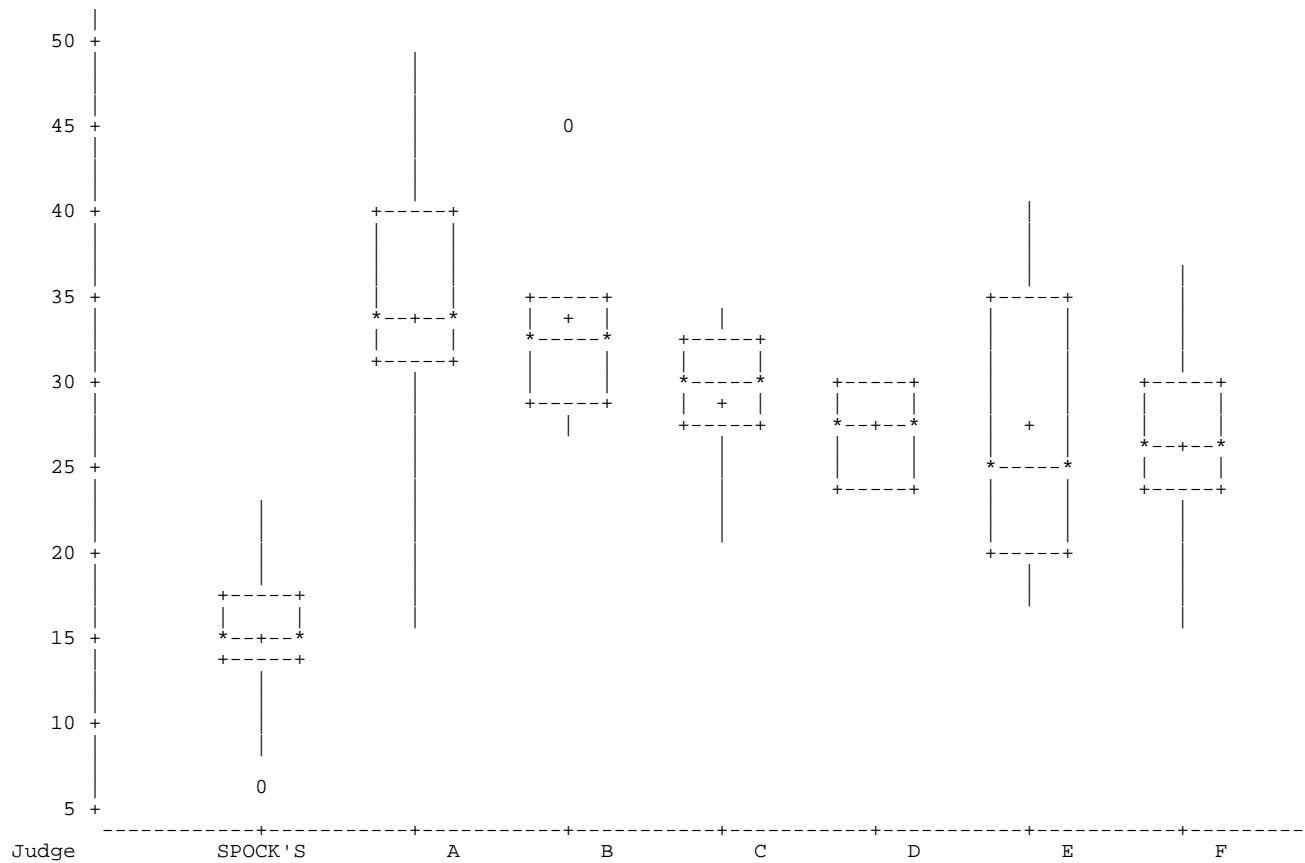
Note that these results match those of the GLM. For some more complicated models and some other types of problems, this will not be true.

Checking the assumptions: Here, as with the first example we will examine the residuals for normality and homogeneity of variance.

First, the PROC UNIVARIATE with the BY JUDGE statement provides a plot similar to that provided by the book.

The UNIVARIATE Procedure
Variable: Percent

Schematic Plots



From this plot we can see box plots of the individual group members (judges). Note that there are few potential outliers and no consistent indication of skewness (mean < or > the median). Some judges seem to have relative large variability while other are smaller. This may indicate nonhomogenous variance.

The dataset below is the output from the PROC GLM OUTPUT statement. Many SAS procs have facilities for outputting results from the procedure. We had previously seen the “OUTP=somename” option on the PROC MIXED MODEL statement. The style “OUTPUT” below is more common in SAS.

```

44      proc glm data=Jury;
45          class judge;
46          model percent = judge;
47          output out=next1 r=e p=yhat;
48      run;

```

With this statement it is possible to specify names for key variables (keyvariable=somename). The key variable names include the following: P or PREDICTED, R or RESIDUAL, RSTUDENT, STUDENT, L95, L95M, U95 and U95M.

Obs	Percent	Judge	e	yhat					
1	16.8000	A	-17.3200	34.1200	21	24.3000	D	-2.7000	27.0000
2	30.8000	A	-3.3200	34.1200	22	29.7000	D	2.7000	27.0000
3	33.6000	A	-0.5200	34.1200	23	17.7000	E	-9.2667	26.9667
4	40.5000	A	6.3800	34.1200	24	19.7000	E	-7.2667	26.9667
5	48.9000	A	14.7800	34.1200	25	21.5000	E	-5.4667	26.9667
6	27.0000	B	-6.6167	33.6167	26	27.9000	E	0.9333	26.9667
7	28.9000	B	-4.7167	33.6167	27	34.8000	E	7.8333	26.9667
8	32.0000	B	-1.6167	33.6167	28	40.2000	E	13.2333	26.9667
9	32.7000	B	-0.9167	33.6167	29	16.5000	F	-10.3000	26.8000
10	35.5000	B	1.8833	33.6167	30	20.7000	F	-6.1000	26.8000
11	45.6000	B	11.9833	33.6167	31	23.5000	F	-3.3000	26.8000
12	21.0000	C	-8.1000	29.1000	32	26.4000	F	-0.4000	26.8000
13	23.4000	C	-5.7000	29.1000	33	26.7000	F	-0.1000	26.8000
14	27.5000	C	-1.6000	29.1000	34	29.5000	F	2.7000	26.8000
15	27.5000	C	-1.6000	29.1000	35	29.8000	F	3.0000	26.8000
16	30.5000	C	1.4000	29.1000	36	31.9000	F	5.1000	26.8000
17	31.9000	C	2.8000	29.1000	37	36.2000	F	9.4000	26.8000
18	32.5000	C	3.4000	29.1000	38	6.4000	SPOCK'S	-8.2222	14.6222
19	33.8000	C	4.7000	29.1000	39	8.7000	SPOCK'S	-5.9222	14.6222
20	33.8000	C	4.7000	29.1000	40	13.3000	SPOCK'S	-1.3222	14.6222
					41	13.6000	SPOCK'S	-1.0222	14.6222
					42	15.0000	SPOCK'S	0.3778	14.6222
					43	15.2000	SPOCK'S	0.5778	14.6222
					44	17.7000	SPOCK'S	3.0778	14.6222
					45	18.6000	SPOCK'S	3.9778	14.6222
					46	23.1000	SPOCK'S	8.4778	14.6222

Refer to the SAS output for evaluation of the assumptions. In particular, note the following concerning assumptions.

- 1) Is the assumption of normality met?
- 2) Are there any outliers?
- 3) Is there a suggestion of non-homogeneous variance in the residuals?