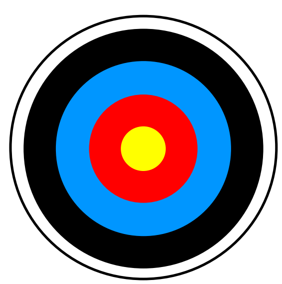
Statistical Techniques II

EXST7015

Post-ANOVA
More LSMeans



LSMeans

- There is something else about the SAS LSMeans statement you should know.
- There are actually several "unusual" or unexpected behaviors of this statement. One we will discuss in connection with AnCova.
- However, there is another general behavior that we should see first.

LSMeans *(continued)*

- **What is the overall mean?**

	Rep					Tmt
Tmt	1	2	3	4	5	Mean
1	2	4	6			4
2	2	6				4
3	3	3	7	8	9	6
4	4					4
5	3	4	6	7		5
6	5	6	7			6
7	3	5				4
				Sum	100	
				n	20	

LSMeans *(continued)*

	Rep					Tmt
Tmt	1	2	3	4	5	Mean
1	2	4	6			4
2	2	6				4
3	3	3	7	8	9	6
4	4					4
5	3	4	6	7		5
6	5	6	7			6
7	3	5				4
Mean	5			Sum	100	33
LSMean	4.71			n	20	7

LSMeans *(continued)*

- **LSMeans calculates means as the mean of means, not the raw mean of all observations.**
- **This is particularly important in unbalanced factorial designs.**
- **For one unbalanced 4 by 5 factorial the means and lsmeans are given below.**
-

LSMeans *(continued)*

- **Raw data**

	Tmt2				
Tmt1	1	2	3	4	5
1	2	3	1	2	2
	3	4	2	4	3
	4	5	3		3
					4
2	5	6	4	8	3
	9	6	5		3
			6		
3	4	6	4	3	8
	5	8	6	7	
	7				
	8				
4	7	6	5	4	5
	8	9	7	7	6
	8	9		7	7
	9				8

LSMeans *(continued)*

- **Comparison of Means & LSMeans.**

	Tmt2						
Tmt1	1	2	3	4	5	LSMean	Raw Mean
1	3	4	2	3	3	3.00	3.00
2	7	6	5	8	3	5.80	5.50
3	6	7	5	5	8	6.20	6.00
4	8	8	6	6	6.5	6.90	7.00
LSMean	6.00	6.25	4.50	5.50	5.13	5.48	5.35
Mean	6.08	6.20	4.30	5.25	4.73		

LSMeans *(continued)*

- **Which is better?**
- **This depends on the situation. Suppose we caught fish in the summer and in the winter, and wanted to express the average temperature at which fish were caught.**
- **The winter mean is 15c and the summer mean is 25c. What is the mean.**

LSMeans *(continued)*

- We do the calculations on the individual catches and find the mean is equal to 24.
- How can that be?
- Well we did 180 samples in the summer and only 20 samples in the winter. So the summer temperatures dominate our samples.

LSMeans *(continued)*

- Perhaps the average temperature would be better expressed as 20, the mean of the means. That is LSMeans
- I generally use LSMeans.
- When testing hypotheses such as $H_0: \mu_1 = \mu_2 = \mu_3$ it is best that the overall mean not be dominated by some cell that has an unusually high number of observations.

LSMeans *(continued)*

- On the other hand, cells with more observations are better estimates of the mean than cells with fewer estimates.
- If the null hypothesis is true, why loose power by treating the cells equally?
- Traditional ANOVA will use RAW means in it's calculation.
- The choice is yours, except that PROC MIXED has only the LSMeans.

Testing for differences between models

- **PROC MIXED provides several tools for comparing models**
- **The intent is to compare between full and reduced models. The statistics used differ from those used in regression.**
  - **Reduced models may be models with some terms omitted, or**
  - **Reduced models may be models with a simpler variance or covariance structure**

Testing for differences between models *(continued)*

- **The test is called a likelihood ratio test.**
  - **It produces a Chi square statistic.**
  - **The degrees of freedom are the d.f. difference between the two models.**
-

Testing for differences between models *(continued)*

- Homogeneous variance is tested automatically with some simple models
- Recall our Typhoid strain example, we requested separate variances for each group with the statement
  - `repeated / group=strain;`
- The resulting output was

```
  Null Model Likelihood Ratio Test
     DF      Chi-Square        Pr > ChiSq
      2           14.56            0.0007
```

Testing for differences between models *(continued)*

- **Note that fitting 3 variances requires 3 d.f., while fitting a homogeneous variance model requires only 1 d.f.**
- **The 2 d.f. difference are the reason the test on the preceding page is a 2 d.f. model.**
- **This test is very similar to Bartlett's test of homogeneity of variance.**

Testing for differences between models *(continued)*

- Suppose that for the baseball example you were told that the salaries of the some positions were highly variable, while others were more stable.
- Perhaps we should have tested for nonhomogeneous for this example.
- So we add the statement,
  - `repeated / group=strain;`

Testing for differences between models *(continued)*

- **SAS fits the different variances for the positions, but does not provide a test.**
- **For some cases we will not get this test automatically. In that case we can calculate it ourselves.**
- **For the original fit we got the results,**

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
team	3466.41	30458	0.11	0.4547	0.05	513.45	3.81E125
Residual	1924296	145057	13.27	<.0001	0.05	1668871	2243534

Testing for differences between models *(continued)*

- **When separate variances are requested we get the following results,**
  - Covariance Parameter Estimates

Cov Parm	Group	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
team		25008	35506	0.70	0.2406	0.05	4960.25	26828515
Residual	Position 1b	3126672	0	.	.	.	.	.
Residual	Position 2b	2276275	902599	2.52	0.0058	0.05	1189304	5985011
Residual	Position 3b	1512066	600277	2.52	0.0059	0.05	789517	3981295
Residual	Position c	759251	201637	3.77	<.0001	0.05	479387	1382686
Residual	Position if	626561	240028	2.61	0.0045	0.05	333467	1582294
Residual	Position of	2558744	407215	6.28	<.0001	0.05	1916409	3590143
Residual	Position p	1875902	208345	9.00	<.0001	0.05	1526216	2361923
Residual	Position ss	1384956	364052	3.80	<.0001	0.05	878092	2504484

- **The first model estimated 2 parameters, while this model fits 9, a difference of 7.**

Testing for differences between models *(continued)*

- **SAS reports the number of parameters fitted in the "Dimensions" section.**
  - **In order to do this 7 d.f. test we take the difference in the "-2 Res Log Likelihood" reported in the "Fit Statistics".**
  - **This value was 6346.8 for the reduced model and 6323.1 for the full model.**
  - **The difference is 23.7, a chi square value with 7 d.f.**

Testing for differences between models *(continued)*

- The probability of a greater chi square value is 0.001286226, a significant result.
- As with regression, when there is a difference in two models the larger model is better, since it presumably provides some information that the smaller model does not.
- If there is no significant difference we decide in favor of the simpler model.
- We just tested homogeneity of variance.

Other between model comparisons

- **SAS also provides some other statistics to compare between models. Also under the "Fit statistics" you will find**
  - `AIC (smaller is better)          6341.1`
  - `AICC (smaller is better)         6341.6`
  - `BIC (smaller is better)          6346.8`
- **And for the smaller model**
  - `AIC (smaller is better)          6350.8`
  - `AICC (smaller is better)         6350.8`
  - `BIC (smaller is better)          6352.1`

Other between model comparisons *(continued)*

- **These are all penalized index values called "Information Criteria". As the note says, smaller is better for all 3.**
- **AIC is the A**kaike **I**nformation **C**riteria
- **AICC is the "C**orrected **AIC "**
- **BIC is the B**ayesian **I**nformation **C**riterion
-
- **and there are others.**

Other between model comparisons *(continued)*

- These all work in a similar fashion. They provide an adjusted measure of goodness of fit.
- These are similar in concept to the "adjusted R^2", so they do not necessarily get smaller when the model gets larger.
- These results also indicate that the full model is better, but they do not provide a test with a probability value.