Example of why you should use the "Lsmeans" (Least squares means) statement instead of the "means" statement in Proc GLM:

Consider the unbalanced two-way ANOVA with factors "Power" and "Humidity." I've shown the y-responses in each respective cell.

This is simulated data, and in truth, there is only a Power effect and no humidity effect.

8 values from Power=20
have $\mu_{20} = 100$, $\sigma_0 = 10$

8 values from Power=30
have $\mu_{30} = 125$, $\sigma_0 = 10$

Consider how Lsmeans & means handles a
Comparison between Humidity=10 & Humidity=20...
Two-Factor unbalanced experiment with factors of Power and Humidity
Example compares LSmeans and means statement for unbalanced data

Power (levels 20, 30)
Humidity (levels 10, 20)
Response: strength of bond

There are a total 16 observations, but the design is unbalanced.
This is a completely randomized design (CRD).

data unb_data;
input power hum y;
cards;
20 10 107
20 10 98
30 10 134
30 10 133
30 10 120
30 10 125
30 10 139
30 10 116
20 20 94
20 20 92
20 20 101
20 20 94
20 20 106
20 20 103
30 20 117
30 20 130
;

/* Get a frequency table for power and humidity */
proc freq data=unb_data;
table power*hum/norow nocol nocum;
run;

/* Plot Y vs. power */
symbol1 value=diamond color=black;
symbol2 value=star color=blue;

proc gplot data=unb_data;
   plot y*power=hum/haxis=15 to 35;
run;
This is an unbalanced design.

The FREQ Procedure
Table of power by hum

| power | hum

**Frequency**
| Percent |   | 10 | 20 | Total *
|---------|---|----|----|--------*
|         | 20 | 2  | 6  | 8      *
|         | 12.50 | 37.50 | 50.00 |
|         | 30 | 6  | 2  | 8      *
|         | 37.50 | 12.50 | 50.00 |
| **Total** | 8 | 8  | 16 | 50.00 50.00 100.00 |

Simulated data truth:
Large power effect;
No humidity effect.

SAS Program LSmeans:

```sas
/*Fit the 2-way ANOVA (additive) model and compare humidity levels with the LSmeans statement*/
proc glm data=unb_data data=diagnostics;
  class power hum;
  model y*power hum;
  lsmeans hum/adjust=bon pdiff;
run;
```
The GLM Procedure
Class Level Information

Class   Levels  Values
power   2        20  30
hum     2        10  20

Number of Observations Read  16
Number of Observations Used   16

The GLM Procedure
Dependent Variable: y

Source     DF  Type I SS  Mean Square  F Value  Pr > F
Model      2  3051.750000  1525.875000  28.78  <.0001
Error     13  689.187500   53.014423
Corrected Total  15  3740.937500

Source     DF  Type III SS  Mean Square  F Value  Pr > F
power      1  2997.562500  2997.562500  56.54  <.0001
hum        1  54.187500    54.187500   1.02  0.3305

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

H0:LSMean1=LSMean2
Pr > |t|

hum   y LSMEAN
10    115.187500    0.3305
20    110.937500

1+test, no adjustment

appropriate way to "aggregate" information from the cells.
SAS Program means:
/*Fit the 2-way ANOVA model and compare humidity levels with the means statement (incorrect option)*/
proc glm data=unb_data;
   class power hum;
   model y=power hum;
   means hum/bon;
run;

The GLM Procedure
Class Level Information

Class   Levels  Values
power   2      20 30
hum     2      10 20

Number of Observations Read         16
Number of Observations Used          16

The GLM Procedure
Dependent Variable: y

Source                  DF      Sum of        Mean Square    F Value  Pr > F
                        DF      Squares      Mean Square
Model                   2   3051.750000  1525.875000    28.78   <.0001
Error                   13  689.187500   53.014423
Corrected Total         15  3740.937500

Source                  DF    Type I SS   Mean Square   F Value  Pr > F
power                   1   2997.562500 2997.562500    56.54   <.0001
hum                     1   54.187500   54.187500     1.02    0.3305

Source                  DF    Type III SS  Mean Square  F Value  Pr > F
power                   1   1912.687500 1912.687500   36.08   <.0001
hum                     1   54.187500   54.187500     1.02    0.3305

The GLM Procedure
Bonferroni (Dunn) t Tests for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha             0.06
Error Degrees of Freedom  13
Error Mean Square    53.014424
Critical Value of t  2.16037
Minimum Significant Difference  7.8643
The GLM Procedure
Bonferroni (Dunn) t Tests for y
Means with the same letter are not significantly different.

Bonferroni Grouping

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>N</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>121.500</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>104.625</td>
<td>8</td>
<td>20</td>
</tr>
</tbody>
</table>

These ‘main effects’ means for humidity are different than those found with the LSmeans statement (these are further apart). The means statement is not finding the average of the two fitted values for each cell within a humidity level (e.g. \( \mu_{11} \) and \( \mu_{21} \) when Humidity=10). It is simply taking the average of all values at Humidity=10 and all values at Humidity=20.

Using the “means” statement SAS declares Humidity=10 significantly different than Humidity=20 (used wrong averages)

\[
\frac{\bar{M}_{11} + \bar{M}_{21}}{2} \quad \text{vs.} \quad \frac{\bar{M}_{12} + \bar{M}_{22}}{2}
\]
We can show this by using the Proc Sort and Proc Means statement:

``` SAS
proc sort data=unb_data;
  by hum;
run;

proc means data=unb_data;
  by hum;
  var y;
run;
```

Sort data by humidity levels.

Take mean of y's for each humidity level.

The MEANS Procedure

hum=10

Analysis Variable : y

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>121.5000000</td>
<td>14.1522335</td>
<td>98.0000000</td>
<td>139.0000000</td>
</tr>
</tbody>
</table>

hum=20

Analysis Variable : y

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>104.6250000</td>
<td>13.0923915</td>
<td>92.0000000</td>
<td>130.0000000</td>
</tr>
</tbody>
</table>

seen in "means" statement output from PROC GLM.

If you just take the overall mean of all values at Humidity=10, this average will look higher than the Humidity=20 average because many of the Humidity=10 observations were from a Power=30, and many of the Humidity=20 observations were from a Power=20. This is why unbalanced data can be a little tricky. LSmeans will do the appropriate thing, but the means statement will not.
Request the appropriate 'contrast' associated with the main effects for Humidity (i.e. linear combination of parameters in the additive model).

```
proc glm data=unb_data;
  class power hum;
  model y=power hum CLparm;
  lsmeans hum/pdiff(CL);
  estimate "main effects humidity" hum 1 -1;
run;
```

The GLM Procedure
Least Squares Means

| hum | y LSMEAN       | Pr > |t| |
|-----|----------------|------|---|
| 10  | 115.187500     | 0.3305 |
| 20  | 110.937500     |       |

<table>
<thead>
<tr>
<th>hum</th>
<th>y LSMEAN</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>115.187500</td>
<td>109.180560 121.194440</td>
</tr>
<tr>
<td>20</td>
<td>110.937500</td>
<td>104.930560 116.944440</td>
</tr>
</tbody>
</table>

Least Squares Means for Effect hum

<table>
<thead>
<tr>
<th>Difference Between Means</th>
<th>95% Confidence Limits for LSMean(i)-LSMean(j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2</td>
<td>-4.831640 13.33164</td>
</tr>
</tbody>
</table>

Dependent Variable: y

| Parameter                  | Estimate | Standard Error | t Value | Pr > |t| |
|---------------------------|----------|----------------|---------|------|---|
| main effects humidity     | 4.2500000 | 4.20374528     | 1.01    | 0.3305 |
| Parameter                 | 95% Confidence Limits |
| main effects humidity     | -4.83163954 13.33163954 |