UNEQUAL CELL SIZES DO MATTER

David C. Howell

Most textbooks dealing with factorial analysis of variance will tell you that unequal cell sizes alter the analysis in some way. I recently came across an excellent example that illustrates this point, and its elaboration may be helpful to people who have to work in this environment. This example is valuable for several reasons. First of all, the pattern of sample size inequality is not dramatic, at least in the sense that within every ethnic group the sexes have roughly similar sized samples, and there are approximately as many males as females. Second, the effect is quite dramatic. Finally, my initial dramatic explanation, while still correct, works even in a relatively undramatic case.

I received the following example from Jo Sullivan-Lyons, who is a research psychologist at the University of Greenwich in London. She was kind enough to share her data with me. In her dissertation, she was asking how men and women differ in their reports of depression on the HADS (Hospital Anxiety and Depression Scale), and whether this difference depends on ethnicity. So we have 2 independent variables—Gender (Male/Female) and Ethnicity (White/Black/Other), and one dependent variable—HADS score.

I have created data which exactly reflect the cell means and standard deviations that the author obtained, and these are available at JSLdep.dat as a tab-delimited ASCII file with the variable names in the first line. An SPSS data file is available at JSLdep.sav .The cell means and standard deviations are given in the table that follows. (Please recall that the data, though fictitious, do reflect her results, and the results are proprietary.)

Report

HADS

Gender of subject	Ethnicity	Mean	N	Std. Deviation
Male	White	1.4800	133	1.6300
	Black	6.6000	10	1.7800
	Other	12.5600	9	2.7400
	Total	2.4729	152	3.3121
Female	White	2.7100	114	1.9600
	Black	6.2600	19	1.2400
	Other	11.9300	28	4.1100
	Total	4.7324	161	4.2419
Total	White	2.0477	247	1.8889
	Black	6.3772	29	1.4262
	Other	12.0832	37	3.7964
	Total	3.6351	313	3.9770

*All output here was created by SPSS 10.

You might suspect that the original data are somewhat skewed given the fact that the standard deviation for males is larger than the mean, and you don't expect negative depression scores. However the data that I am using were drawn from random normal distributions, so that is not a concern to me, though it should be to her.

The author's first question concerned whether males and females differ in their level of reported depression. From the table you can see that the mean for males is approximately 2.47, while the mean for females is 4.73, for a difference of 2.26 units. This difference appears substantial, especially given the sample sizes and the standard deviations.

A quick *t* test comparing males to females would seem to support this conclusion. The *t* test is given below. Here we lumped all males together, regardless of Depression category and compared the mean with the mean of all females.

Group Statistics

	Gender of subject	N	Mean	Std. Deviation	Std. Error Mean
HADS	Male	152	2.4729	3.3121	.2686
	Female	161	4.7324	4.2419	.3343

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
HADS	Equal variances assumed	8.570	.004	-5.232	311	.000	-2.2595	.4319
	Equal variances not assumed			-5.268	300.6	.000	-2.2595	.4289

Whether or not we are willing to assume equal variances, the *t* value is clearly significant. Males report less depression (as measured by the HADS) than females. (Note that although the standard deviations are very similar, with these sample sizes a test on the assumption of homogeneity is significant. The fact that the difference is small is apparent in the relatively minor adjustment to the degrees of freedom in the unequal variance case.)

But, the original question implied that we should take the actual level of ethnicity into account. This would suggest a two-way analysis of variance, with Sex and ethnicity as independent variables, and HADS as the dependent variable. This analysis follows, with some surprising results.

Tests of Between-Subjects Effects

Dependent Variable: HADS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3577.528[a]	5	715.506	161.854	.000
Intercept	5465.033	1	5465.033	1236.240	.000
SEX	.214	1	.214	.048	.826
ETHNICIT	2790.110	2	1395.055	315.574	.000
SEX * ETHNICIT	32.663	2	16.331	3.694	.026
Error	1357.151	307	4.421		
Total	9070.746	313			
Corrected Total	4934.680	312			

a. R Squared = .725 (Adjusted R Squared = .720)

1. Gender of subject

Dependent Variable: HADS

1. Gender of subject

Dependent Variable: HADS

Gender of subject	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Male	6.880	.328	6.235	7.525
Female	6.967	.218	6.537	7.396

2. Ethnicity

Dependent Variable: HADS

Ethnicity	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
White	2.095	.134	1.831	2.359
Black	6.430	.411	5.622	7.238
Other	12.245	.403	11.452	13.038

Notice that there is a significant effect due to ethnicity, and there is an interaction of ethnicity by sex, but there is no sex effect. There isn't even an "almost" sex effect. The *F* and *p* values for Sex are 0.048 and 0.826, respectively. **What Happened!!!!**

Well, first of all look at the means for HADS in the table above. They are 6.880 and 6.967, which are awfully close. But why aren't they 2.47 and 4.73, as they were a minute ago? You should recall that when we have unequal sample sizes in a factorial analysis of variance, and we run a standard (default) analysis of variance using most statistical software, the solution we get is basically comparing unweighted means. Thus the mean for males in this analysis is not 2.47, but rather (1.48 + 6.60 + 12.56)/3 = 6.88. It is much different from a mean of 2.47 because the 2.47 was based on the fact that there were 133 subjects in the White cell (with a mean of only 1.48), but only 10 in the Black cell and 9 in the Other cell, with means of 6.60 and 12.56, respectively. Averaging the 6.6 and 12.56 requires that they carry exactly as much weight as the 1.48, whereas the original male mean of 2.47 was very heavily weighted by all of the White cases with a low mean.

As a short digression to satisfy my pedantic nature, you might think, as I did, that if you take these adjusted means (6.880 and 6.967), grab the error term from the overall Anova, and take the 2 sample sizes and run a *t* test, you could reconstruct the $F = t^2$ from the Anova. In other words, my first thought

was that I have just run a *t* test on these means rather than the weighted means. Well, not quite. When the answer didn't come out exactly the way I expected, I did a bit of computing backward and realized that I would need to run the *t* test with about 57 males and 57 females to get my desired *F*. And where did this come from, you ask? Well, if you look in Chapter 13 where I discuss unequal cell sizes, you will recall that I did that analysis with harmonic means of sample sizes. The harmonic mean of these 6 cells is 19.0025, and 3*19 = 57. What this means is that the difference between the standard *t* test and the analysis of variance here is primarily that the Anova has run a *t* test on unweighted means—with appropriate adjustments to cell sizes. I bet you're glad I cleared that up!

That is part of the answer, but it isn't all of it. Females also showed a similar pattern of cell sizes, with a predominance of women falling in the White category, so they would have similar weighting effects. But, the large group of women in the White cell differed noticeably from the males in that cell, even though males and females don't differ all that much in other cells. So what is happening with the test is that the cells where there is a difference (row 1) swamp the cells where there isn't much of a difference, and give a significant effect. In fact, in the Black and Other cells, women actually score slightly *lower* than men.

The differences described above, along with the significant interaction, should lead you to an interest in simple effects. Specifically, is there a significant sex effect at each level of ethnicity. You could run the analysis separately for each row, and then go back and recalculate the *F* for sex by substituting the error term from the overall analysis of variance. Alternatively, you could accomplish the same thing by using SPSS syntax commands. (If there is a way to do this from the menus, I haven't figured it out.) The relevant commands are

```
Manova hads by sex(1,2) Ethnicit(1,3)
/Design = Ethnicit, Sex within Ethnicit(1), sex within Ethnicit(2),
sex within Ethnicit(3).
Execute.
```

The printout follows. It is clear that the sex difference is restricted to the case of White subjects, though I don't know what to make of that in clinical terms..

```
The default error term in MANOVA has been changed from WITHIN CELLS to
WITHIN+RESIDUAL.  Note that these are the same for all full factorial
designs.



* * * * * * A n a l y s i s   o f   V a r i a n c e — design   1 * * * * * *

 Tests of Significance for HADS using UNIQUE sums of squares
 Source of Variation          SS      DF        MS         F  Sig of F

 WITHIN+RESIDUAL         1357.15     307      4.42
 ETHNICIT                2790.11       2   1395.05    315.57      .000
 SEX WITHIN ETHNICIT(      92.87       1     92.87     21.01      .000
 1)
 SEX WITHIN ETHNICIT(        .76       1       .76       .17      .679
 2)
```

```
SEX WITHIN ETHNICIT(       2.70        1       2.70        .61       .435
3)

(Model)                  3577.53        5     715.51     161.85       .000
(Total)                  4934.68      312      15.82

R-Squared =            .725
Adjusted R-Squared =   .720

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

*These are exactly the *F*s that you would obtain if you ran separate analyses of variance at each level of Ethnicity, but then replaced the individual error terms with the error term from the overall analysis (i.e. 4.42) and recalculated the *F*s.

So, What Does This All Mean?

That question really has several answers. The first answer is that White males may be less forthcoming in reporting depression than females. On the other hand, males and females in the Black and Other groups do not differ in their reporting.

But that begs the second question, which is why did the two analyses show such different results? I hope that I have explained that already. It is due to the fact that the two analyses are really comparing different means. The *t* test is comparing the weighted means of males and females. The *t* test just lumps all males together and calculates their mean. The same with females. The analysis of variance compares unweighted means; it takes the means of the three cell means for males, and then for the three cell means for females. These turn out to be quite different things when the sample sizes are quite unequal.

A third answer is contained in a more extreme situation, which I outlined in my initial response to Jo Sullivan-Lyons. That message follows—with some minor cleaning up to save me embarrassment.

> Jo,
>
> I do have one hypothesis that would explain this, but that is such a large discrepancy in *p* values that I'm not sure that it would explain all of it. *(It did.)*
>
> IF you ignored ethnicity completely, the square root of the *F* from a one-way Anova on gender would be exactly the same as the *t*. But when you move to a 2-way, things fall apart. Suppose that I have the following completely fictitious means:

	Black	White
Male	10	40
Female	10	40

> IF I had equal numbers in each cell, the male and female means would both be 25, and there would be no significant effect. That would be true whether I ran a *t* test or a one- or two-way Anova. Notice also that there is a huge difference between blacks and whites, whether we

> care about it or not.
>
> Suppose that I had 1 black male and 14 white males. Then the mean of the males would be (1*10 + 14*40)/15 = 38. And suppose that I had 14 black females and 1 white female. That mean would be (14*10 + 1*40)/15 = 12.
>
> A *t* test on the two groups of 15, with these peculiar sample sizes, would test the difference between 12 and 38, and it would very likely be significant.
>
> What you have here mirrors what you found, although without knowing your data I can't say whether it explains your data. What you really have is a difference due to ethnicity, but it *looks like* a difference due to sex because of the pattern of sample sizes.
>
> I can also make up a similar example with means of [10 40] and [40 10], just by adjusting my samples sizes differently.
>
> So, my suggestion to you is that you look closely at your sample sizes, and see if you are unintentionally doing something similar to what I did here.
>
> Good luck,
>
> Dave Howell

Finally, I will end with what is not really an answer at all, but a question. The question is "Which analysis is correct?" Would you rather go with the *t* test or the factorial analysis of variance? My first answer would be "Of course, go with the Anova." But I'm not sure that is always the best answer. IF you really want to know if men (in general) report fewer symptoms of depression than women, then I suppose you should go with the *t* test. My reasoning is that it asks a question about "men in general." It is true that in the population of all men and all women, men report fewer symptoms, even if you know that this really holds only for the White men and women.

Let me elaborate on that a bit. You might ask why I would want to ask such a stupid question when the "interaction question" is really more interesting. But suppose we were talking about salaries. Women earn less than men. Sure, you can explain a lot of that by grouping subjects by years in the work force, kind of job, etc., but the fact remains, *and the fact is important*, that women earn less than men. The *t* test would tell you that, the Anova wouldn't. (I know that I have grossly oversimplified the situation, but the point should be clear. Sometimes we don't want to adjust)

If these were your data, and if you really want to understand what is going on, you should probably just ignore the *F* value for Sex that you get in the Anova, and concentrate on the interaction and then on the simple effects. Here, the *F* for sex is probably answering a question you don't even want to ask. It is telling you how things average out across ethnic groups, when you probably don't care how they average out. You care if there are sex differences within some ethnic groups and not others. That's a different question.

I can modify Jo's data slightly to remove the interaction, and I still find a significant *t* but not a significant *F*. So the interaction isn't always the issue. All I did was to swap the male and female means in the 2nd and 3rd rows, and I got the following table, which has no interaction or sex effect. (The *t* test on Sex hardly changes.)

Tests of Between-Subjects Effects

Dependent Variable: DEP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3800.031[a]	5	760.006	171.920	.000
Intercept	5465.033	1	5465.033	1236.240	.000
SEX	15.329	1	15.329	3.467	.064
Ethnicity	2790.110	2	1395.055	315.574	.000
SEX * Ethnicity	6.326	2	3.163	.715	.490
Error	1357.151	307	4.421		
Total	9403.243	313			
Corrected Total	5157.182	312			

a. R Squared = .737 (Adjusted R Squared = .733)


Return to Dave Howell's Statistical Home Page

Send mail to: David.Howell@uvm.edu)

Created 1/2/2006
Last revised: 3/8/2009