

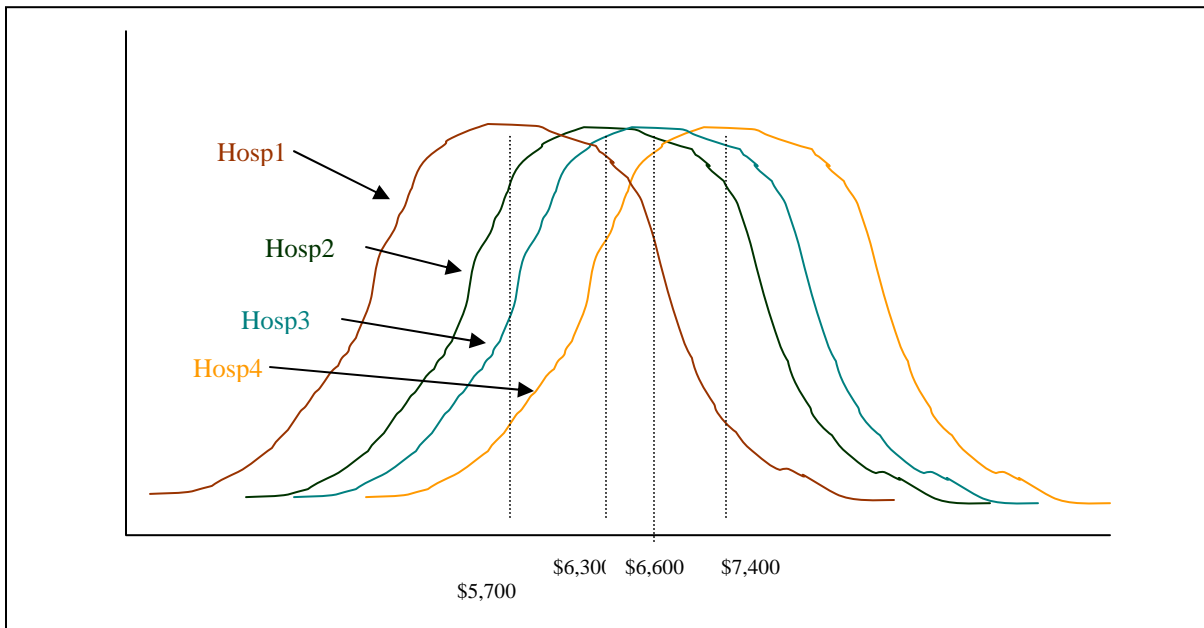
Analysis of Variance – ANOVA

Introduction

Suppose an alliance of four hospitals is interested in understanding the rise in costs and what can be done to control these costs. Three measures of hospital performance that might be considered are length of stay, readmission rates, and costs per case. One initial question might be whether there are real differences between the eight hospitals in these three –cost-related measures of performance.

Note that if we were interested if two of the hospitals differed from each other, we could do a simple F (or t) test as we discussed in the last section. But here things are a bit more complicated since we want to know if there are differences across the eight hospitals.

Suppose we have data on the cost of stay for the four hospitals and the distributions look as follows:



Note that they all have slightly different means. So the question is do these 4 hospitals have different distributions or is it just sampling error that gives the above differences?

II. One way Analysis of Variance

Analysis of variance depends on two pieces of information:

Between Group Variance – variation associated with the differences between the groups. This is also referred to as the explained variance

Within Group Variance – measure of the difference between the observations within groups. This is unexplained variance.

Between Group Variance is calculated by:

$$SS_B = \sum_{j=1}^m n_j (\bar{x}_j - \bar{x})^2$$

where m is the number of groups (hospitals in this case m=4). n_j is the sample size for group j, \bar{x}_j is the sample mean for group j, and \bar{x} is the mean for all the observations across all groups (a common sample mean).

So note that this is basically the difference between each group's mean and the mean for the pooled sample, squared and added up.

If there is a lot of differences between the groups note that this will be a relatively large number, whereas if the groups are largely the same then SS_B will be a small number.

Within Group Variance is calculated by:

$$SS_W = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

Where here we are looking at deviations from the mean within each group. So for each group take the difference from the mean, square it and add it up. Then add up the sum of squares for each group.

Another measure of variation is the total variance:

$$SS_T = \sum_{j=1}^m \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

This is just like SS_W except we are looking at the pooled mean instead of each group's mean.

It turns out that $SS_T = SS_B + SS_W$

1. The hard way:

Now we can do the F test. We basically want to know if the between group variance is large relative to the within group variance. Since the sum of the two is the total variance

we are asking the question do the data vary because of things going on across the groups or within the groups?

The null hypothesis is:

H_0 there are no differences in the groups – or in this case costs and hospital are independent of each other

H_a : There are some differences attributable to specific hospitals – or hospitals do differ in their costs.

The formula for the test is:

$$F = \frac{SS_B / df_{SSB}}{SS_W / df_{SSW}} = \frac{MS_B}{MS_W}$$

If there are m groups (hospitals) then there are m-1 degrees of freedom in SS_B . The degrees of freedom in MS_W is: $n_1 - 1 + n_2 - 1 + \dots + n_j - 1$

Suppose we have the following data for the cost per stay for four hospitals:

See the accompanying Excel file: AnovaExample.XLS for these examples

This is under the COSTS worksheet

Albemarle	Beaufort	Catawba	Dare
Cost	Cost	Cost	Cost
\$5,342.61	\$6,844.89	\$5,455.73	\$1,035.75
\$7,237.59	\$6,283.80	\$4,448.48	\$6,130.58
\$11,801.95	\$5,206.09	\$8,996.80	\$2,584.28
\$1,082.36	\$22,030.51	\$4,734.98	\$2,635.85
\$2,379.59	\$5,372.86	\$13,055.10	\$5,371.03
\$8,469.39	\$2,197.14	\$10,194.42	\$14,433.99
\$4,453.55	\$13,472.69	\$16,229.83	\$2,283.88
\$4,745.26	\$8,486.18	\$7,099.69	\$3,630.39
\$11,963.80	\$2,447.01	\$5,948.36	\$4,775.36
\$5,294.96	\$19,182.24	\$4,155.38	\$6,084.66
\$2,965.55	\$10,021.34	\$14,470.90	\$1,414.80
\$3,896.16	\$1,594.57	\$4,957.09	\$6,630.33
\$6,074.21	\$4,412.80	\$5,040.75	\$5,139.11
\$7,217.18	\$7,000.21	\$4,113.04	\$16,346.71
\$3,730.40	\$1,221.58	\$18,527.78	\$5,068.54
\$1,248.14	\$6,956.35	\$7,272.11	\$1,810.36
\$14,258.95	\$2,568.13	\$10,481.06	\$5,881.90
\$19,475.78	\$2,257.24	\$3,156.14	\$4,149.18
\$3,384.80	\$3,946.68	\$14,447.60	\$3,597.68
\$6,680.14	\$9,546.52	\$6,412.51	\$10,679.39
\$1,871.78	\$9,321.12	\$13,480.78	\$5,249.46
\$4,655.22	\$4,278.73	\$18,847.52	\$10,377.23
\$7,899.41	\$3,720.09	\$15,593.65	\$3,063.96
\$12,005.57	\$2,314.85	\$10,049.19	\$3,580.55

	\$3,117.39	\$2,741.32	\$9,632.61	\$10,516.46
	\$10,750.39	\$3,489.61	\$18,571.33	\$6,082.93
	\$5,097.72	\$9,916.19	\$4,513.67	\$960.28
	\$2,280.30	\$2,793.26	\$6,049.84	\$3,118.49
	\$16,611.95	\$4,394.15	\$6,716.69	\$12,919.48
	\$3,237.47	\$3,598.72	\$8,768.23	\$5,933.54
Group Mean	\$6,640.99	\$6,253.90	\$9,380.71	\$5,716.21
sample sizes	30	30	30	30
Pooled Mean	\$6,997.95			

To get SS_w we take each observation and subtract it from its group mean and square it and then add them up:

	Albemarle	Beaufort	Catawba	Dare
	$(x-\bar{x})^2$	$(x-\bar{x})^2$	$(x-\bar{x})^2$	$(x-\bar{x})^2$
	1685779	349274	15405458	21906659
	355937	894	24326880	171707
	26635553	1097897	147386	9808954
	30898319	248901560	21582795	9488587
	18159493	776224	13501152	119146
	3343062	16457267	662126	75999775
	4784875	52110992	46910463	11780855
	3593776	4983093	5203046	4350624
	28332352	14492378	11781017	885189
	1811785	167142087	27304060	135759
	13508827	14193637	25910048	18502085
	7534068	21709316	19568402	835625
	321235	3389633	18835241	333039
	332000	556985	27748333	113007637
	8471509	25324201	83668914	419470
	29082784	493442	4446188	15255625
	58033381	13584869	1210773	27455
	164731946	15973257	38745255	2455567
	10602745	5323244	25673388	4488148
	1533	10841375	8810203	24633205
	22745323	9407865	16810585	217851
	3943265	3901279	89620517	21725154
	1583632	6420171	38600640	7034404
	28778765	15516081	446867	4561022
	12415726	12338188	63454	23042448
	16887204	7641275	84467520	134487
	2381669	13412400	23688065	22618823
	19015579	11975999	11094686	6748123
	99420130	3458654	7096995	51887171
	11583919	7049958	375130	47235
sum =	630976172	708823494	693705589	452621828
SSW=	2486127083			

Then SS_W is the sum of each sum

To get SS_B We take each mean from the pooled mean and square it and then multiply it by its sample size:

Group Mean	\$6,640.99	\$6,253.90	\$9,380.71	\$5,716.21
sample sizes	30	30	30	30
Pooled Mean=average(\$6,997.95			
	3822679.29	16608449.72	170326344.61	49286011.22
SSB=	240043485			

Then add each up

So now the df for SS_B will be 3 since there are 4 groups. DF for SS_W will be $(30-1)+(30-1)+(30-1)+(30-1) = 116$

And so $F = (240043485/3)/(2486127083/116) = 3.7334$

Then using $FDIST(3.7334,3,116)$ we get a pvalue of .01323

Or if the null is true (there are no differences in costs) then there is only a 1.3 percent change we could get the sample we got. And so therefore we reject the null and conclude that there is evidence that hospital and cost are dependent. In other words there are hospital-specific costs differences in your system.

2. The easy way

There is an easier way to do this using the excel Add-in for one-way ANOVA

Tools – Data Analysis – ANOVA Single Factor – Fill in input range (each group must have its own column).

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Albemarle	30	199229.57	6640.9857	21757799
Beaufort	30	187616.87	6253.8957	24442189
Catawba	30	281421.26	9380.7087	23920882
Dare	30	171486.15	5716.205	15607649

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	240043484.8	3	80014495	3.73339	0.013233	2.68281

Within Groups	2486127083	116	21432130
Total	2726170568	119	

We get the same numbers.

3. Where are the differences?

Note that all this tells you is that there are some differences somewhere between the groups, but it does not tell you where. We can use the information above, to do some additional tests to decide where the groups differ from each other. One way would be to do tests for differences in the means as we did previously, or we could use the following formula:

$$F = \frac{(\bar{X}_1 - \bar{X}_2)^2 / \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{MS_w}$$

The numerator takes the squared difference of the two means from the groups we're interested in, then divides by the sample sizes. The denominator is the mean square within groups from the original F test.

Doing this for the above data gives:

	Albemarle	Beaufort	Catawba	Dare
	6640.98567	6253.895667	9380.7087	5716.205
Albemarle	6640.986	0.104869652	5.2533851	0.5985541
Beaufort	6253.896		6.8427353	0.2023443
Catawba	9380.709			9.3984502
Dare	5716.205			

And the associated Pvalues are:

	Albemarle	Beaufort	Catawba	Dare
	6640.98567	6253.895667	9380.7087	5716.205
Albemarle	6640.986	0.74664546	0.0237078	0.4407041
Beaufort	6253.896		0.0100837	0.6536747
Catawba	9380.709			0.002702
Dare	5716.205			

So note using an $\alpha=.05$ the significant differences are between Catawba and Albemarle, Catawba and Beaufort, and Dare and Catawba

But note that there is a slight problem with all this. Since we are doing 6 separate tests, there is a slight problem with our α -- the probability of a type I error. It turns out that if we conduct six tests, the overall probability of alpha for all six tests will be

$1-(1-\alpha)^6$ So if $\alpha=.05$ then there is about a .26 probability. So that means that between these 6 tests there is a 26% chance of making a type I error.

To adjust for this we can adjust our critical value with the following:

$$\alpha_{Adj} = 1 - \sqrt[c]{1 - \alpha}$$

Where c is the number of comparisons being made.

$=1-((1-0.05)^{(1/6)}) = .0085$, so this now becomes our new critical probability limit.

So now note that the only significant difference in our table is that between Catawba and Dare.

Later when we do the multiple regression material, we'll see that there is a much easier way to do all of this.

III. ANOVA for Repeated Measures

What we just looked at was the case where we had a number of groups and we wonder if they are different or not. Often we will have one group that is measured on some variable more than two times, and so we might ask "are there any differences between the several measurements?"

Suppose we are interested in the cost of readmissions. We are interested if the cost of later admissions in a string of readmissions for the same person are likely to be higher than earlier admissions.

So do costs differ from admission to readmission?

Suppose we have a sample of 12 people who have three admissions each. Note that if we just had two admissions, we could just do a t test on the mean of the first vs. the second admission. But here we have a more complicated scene.

Here the "group" is the person and so the total variation (SS_T) can be divided between variation between people (or groups), SS_B , and the variation within people (or groups), SS_W .

But note that in this case we are not really interested in differences between people, but differences within people, these may reflect different costs of hospital stays, depending on which admission is being considered. But note that SS_W can be divided into two sources of variations:

Variation due to the differences in costs of first, second, or third admissions (SS_A), and variation that cannot be explained by whether the admission is first, second, or third SS_R (or residual variation).

$$SS_W = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_i)^2$$

So for each person, take the difference for each observation and the person's mean, square it and then add them all up.

The variation due to differences between costs is:

$$SS_A = n \sum_{j=1}^m (\bar{x}_j - \bar{x})^2$$

This is the global mean subtracted from the mean for each admission – so it measures how much bouncing around there is within each admission.

$$(\bar{x}_j - \bar{x})^2$$

To get SS_R the easiest way is to take the difference between SS_W and SS_A , but the formula turns out to be:

$$SS_R = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

Where \bar{x}_i is the mean for each person, and \bar{x}_j is the mean for each admission.

So SS_W measures how much bouncing around there is within individuals. This is due to two components. One would be how much bouncing around there is over time (that is differences across admissions), and the second component would just be random bouncing.

The F test in this case basically looks at the ratio of SS_A to SS_R after accounting for degrees of freedom:

The degrees of freedom for SS_A are $(m-1)$ or the number of time units (admissions) minus one.

Degrees of freedom for SS_R are $(n-1)(m-1)$ or the number of groups – 1 times the number of time periods – 1

So the Test is:

$$F = \frac{SS_A / df_{SSA}}{SS_R / df_{SSR}} = \frac{MS_A}{MS_R}$$

Example: The data are for 12 different patients who had 3 admissions to the hospital See the Readmit worksheet in the ANOVAExample.xls file.

Person	Admt	Cost	Person	Admit	To get SSW	To get SSa	To Get SSR
			\bar{x}_i	\bar{x}_j	$(x_{ij} - \bar{x}_i)^2$	$(\bar{x}_j - \bar{x})^2$	
1	1	\$2,319.69	3452.2	3898.9	1282669.5		268305.6
1	2	\$2,898.01	3452.2	3926.9	307170.9		1043.6
1	3	\$5,139.02	3452.2	5714.6	2845226.8		235882.9
2	1	\$4,193.34	4331.3	3898.9	19032.0	377693.6	227158.2
2	2	\$3,243.61	4331.3	3926.9	1183062.3	344022.7	251153.5
2	3	\$5,556.94	4331.3	5714.6	1502201.6	1442646.5	602.3
3	1	\$2,508.31	5036.1	3898.9	6389806.5		3660483.0
3	2	\$4,244.97	5036.1	3926.9	625913.0	2164362.8	41866.2
3	3	\$8,355.07	5036.1	5714.6	11015451.2		4485293.3
4	1	\$7,088.50	5491.0	3898.9	2551910.4	SSa= 25972353.1	4893111.1
4	2	\$4,932.22	5491.0	3926.9	312268.6		768.6
4	3	\$4,452.37	5491.0	5714.6	1078814.6		5016534.8
5	1	\$1,786.17	2815.0	3898.9	1058429.4		171588.3
5	2	\$3,140.76	2815.0	3926.9	106139.1		832335.9
5	3	\$3,517.98	2815.0	5714.6	494223.1		248095.9
6	1	\$4,413.98	2673.5	3898.9	3029421.5		5546454.1
6	2	\$1,599.82	2673.5	3926.9	1152695.7		237268.6
6	3	\$2,006.57	2673.5	5714.6	444737.8		3489382.5
7	1	\$6,767.15	6388.5	3898.9	143398.5		986541.1
7	2	\$2,713.24	6388.5	3926.9	13507315.6		9540040.2
7	3	\$9,685.02	6388.5	5714.6	10867241.9		4390901.4
8	1	\$2,519.60	4504.2	3898.9	3938822.4		1877116.2
8	2	\$7,210.54	4504.2	3926.9	7324023.6		10842714.8
8	3	\$3,782.60	4504.2	5714.6	520773.9		3696963.3
9	1	\$6,934.29	7099.3	3898.9	27234.9		202084.2
9	2	\$4,571.02	7099.3	3926.9	6392300.9		3770453.5
9	3	\$9,792.65	7099.3	5714.6	7254026.5		2226743.7
10	1	\$3,281.44	5705.9	3898.9	5877941.6		3275661.6
10	2	\$6,397.57	5705.9	3926.9	478425.8		1633840.7
10	3	\$7,438.65	5705.9	5714.6	3002468.8		282663.5
11	1	\$2,275.54	3348.0	3898.9	1150184.8		209671.4
11	2	\$1,954.76	3348.0	3926.9	1941136.3		650784.6
11	3	\$5,813.72	3348.0	5714.6	6079742.2		1599241.3
12	1	\$2,698.68	3316.5	3898.9	381656.2		10.3
12	2	\$4,216.57	3316.5	3926.9	810192.0		2210101.8
12	3	\$3,034.14	3316.5	5714.6	79706.5		2200551.4

Global Mean \$4,513.46 **SSw= 105175766.5** **SSR= 79203413.5**

SSw-SSa= 79203413.5

df SSa= 2.0
df SSR= 22.0

MSa= 12986176.5

MSr=		3600155.2
	F=	3.6
Pvalue =		0.0442

Recall that the null hypothesis is that there is no difference across admissions
 So here our pvalue is .044 which implies that there is a 4.4% chance that we could get
 an F value of 3.6 or greater if there were no systematic differences across admission
 rates.

So since .044 < .05 we would (marginally) reject our null hypothesis and conclude
 that there is a difference in costs across admissions.

IV. Factorial Analysis of Variance

In the above example were looking at two variables: cost and the hospital. But
 suppose we also want to know if there is a relationship between a single numerical
 variable and two (categorical) variables:

Eg. Suppose we think length of stay is differs by hospitals within the alliance, but
 also by gender.

Suppose each of the four hospitals divides its admissions for the past year into males
 and females and pulls randomly 10 of each.

One source of variation we are interested in here is the within-cell variation:

$$SS_{WC} = \sum_{k=1}^q \sum_{j=1}^m \sum_{i=1}^n (x_{ijk} - \bar{x}_{jk})^2$$

k indexes the first factor, say hospitals (so q=4), then the second factor is gender so j
 indexes gender (m=2), and i indexes individuals.

This is known as the sum of squares – within cells. This is the proportion of the
 variation in the data that occurs within subgroups. Here we take each observation and
 subtract the hospital by gender mean from each. So take each male in hospital one
 and subtract the mean for males in hospital 1, etc.

Note that this is capturing variation within the subgroups.

Variation due to factor 1 differences can be written:

$$SS_{rows} = mn \sum_{k=1}^q (\bar{x}_k - \bar{x})^2$$

This looks at variations across the first group – or in this case hospitals. It subtracts the global mean from the individual hospital means, squares it and adds them up and then multiplies it by m (the number of the second factor) by n (the number of individuals in each subgroup).

Similarly for factor 2:

$$SS_{cols} = qn \sum_{j=1}^m (\bar{x}_j - \bar{x})^2$$

This is the same formula except for the other factor – in this case gender.

The final type of variation is that due to interaction:

$$SS_{rows * cols} = n \sum_{k=1}^q \sum_{j=1}^m (x_{jk} - \bar{x}_j - \bar{x}_k + \bar{x})^2$$

So take each subgroup mean (gender by hospital) subtract each group mean (overall gender, overall hospitals) and then add the global mean. The intuition is tougher here, but basically it is capturing the fact that the groups may work together to create variation. That is, it is not just hospital alone or gender alone, but somehow two working together results in variation. Eg. Suppose females have a common LOS across hospitals, but male LOS is greater for one or two of the hospitals. In this case gender and hospital work together to explain the variation.

The degrees of freedom on all these are:

SS_T $qmn-1$ (the number of hospitals, times the number of genders, times the number of obs in each cell)

SS_{WC} $qm(n-1)$

SS_{rows} $q-1$

SS_{cols} $m-1$

SS_{r*c} $(q-1)(m-1)$

So what do we do with all this?

There are three tests we can run:

1) Are there differences by factor 1?

$$F_{rows} = \frac{SS_{rows} / df_{rows}}{SS_{WC} / df_{WC}} = \frac{MS_{rows}}{MS_{WC}}$$

2) Are there differences by factor 2?

$$F_{cols} = \frac{SS_{cols} / df_{cols}}{SS_{WC} / df_{WC}} = \frac{MS_{cols}}{MS_{WC}}$$

3) Are there interaction differences between factors?

$$F_{r * c} = \frac{SS_{r * c} / df_{r * c}}{SS_{WC} / df_{WC}} = \frac{MS_{rows}}{MS_{WC}}$$

Example: See LOS – Factorial Analysis worksheet in the ANOVAExample.XLS file. The worksheet LOS- by Hand shows the ugly calculations of all this. The worksheet LOS – Factorial Analysis uses the Excel Data Analysis Tool pack.

Anova: Two-Factor With Replication

SUMMARY	Male	Female	Total
<i>Albemarle</i>			
Count	10	10	20
Sum	36	53	89
Average	3.6	5.3	4.45
Variance	5.377778	30.233333	17.62895

<i>Beaufort</i>			
Count	10	10	20
Sum	40	60	100
Average	4	6	5
Variance	13.77778	15.77778	15.05263

<i>Catawba</i>			
Count	10	10	20
Sum	71	108	179
Average	7.1	10.8	8.95
Variance	24.54444	25.51111	27.31316

<i>Dare</i>			
Count	10	10	20
Sum	50	62	112
Average	5	6.2	5.6
Variance	7.111111	21.95556	14.14737

<i>Total</i>			
Count	40	40	
Sum	197	283	
Average	4.925	7.075	

Variance 13.60962 26.43013

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Sample	245.3	3	81.76667	4.533498	0.005738	2.731809
Columns	92.45	1	92.45	5.125828	0.026586	3.973895
Interaction	17.65	3	5.883333	0.326197	0.806404	2.731809
Within	1298.6	72	18.03611			
Total	1654	79				

So note that we would conclude that there differences both by hospital and by gender, but there are no differences due to interaction between the two.