

## GeoStats for Two Populations [\*Draft 2009-05-09]

The basic approach for this tutorial is to assume I am going to treat a collection of *entities*, with either treatment A or treatment B and observe the outcome. The major question is: *are the effects of these two treatments equal or not?*

The intent of this tutorial is to visually/geometrically present a way to tell if the means of these two treatments are equal or not. The geometry presented here relates vector directions and lengths to standard statistical tests with the last section of this tutorial showing a conventional ANOVA table.

If the entities are people, perhaps treatment A is teaching method A as compared with teaching method B and the observations are test scores. If the entities are inanimate objects, such as gasoline brands, perhaps the treatment is the use of gasoline brand A as contrasted with the use of gasoline, brand B. The observed results might be the miles per gallon produced by each. (This is the experimental context I will adopt in this tutorial).

Both resulting populations of observed values associated with treatment A and B are assumed to follow a Normal distribution, with a common, but unknown variance. The intent of this tutorial is to visually present a way to tell if the means of these two treatments are equal or not.

To distinguish between the case of the means being equal versus their being unequal, I will introduce and use a standard statistical test called the 'F-test'. (This test is named in honor of sir Ronald Fisher, an eminent statistician, and founder of multivariate analysis).

To make this pretty simple I will treat only three representative entities using treatment A and three entities using B. That is, the first three sample observed values will come from treatment A, while the second three will result from treatment B. The reader could think of these two treatments as being the two types of gasoline, A, and B and the measurements reported are the overall miles per gallon achieved under test conditions. (Note: the reader will need some familiarity with vector space ideas to appreciate some of the ideas in this tutorial. For a quick refresh, see the document on this web site, *milagrosoft.com*. The title is: *Vector Operations Quick Look*).

Getting back to the purpose of this note, I want to test the hypothesis that the treatment results from A and B, have the same mean versus the hypothesis that they don't. That is, is the true mean,  $\mu_1$ , of the population of values associated with treating entities with treatment A, the same as the true mean,  $\mu_2$ , of the population of values that come from treating entities with treatment B. Concretely, the question is: does gasoline type A yield the same mean miles per gallon as gasoline type B, statistically speaking?

The alternative hypothesis is that they do differ, *significantly*.

More formally, the *null hypothesis* is that there is no treatment difference, and is written as  $H_0 : \mu_1 = \mu_2$ . An equivalent way to write this, which will relate to *a direction within the vector space*, is:  $H_0 : \mu_1 - \mu_2 = 0$

The *alternative hypothesis* is written as:  $H_1 : \mu_1 \neq \mu_2$  or equivalently,  $H_1 : \mu_1 - \mu_2 \neq 0$ .

Here, the true population mean of treatment A outcomes is  $\mu_1$  and the population mean of treatment B outcomes is  $\mu_2$ . Both of which are unknown and must be estimated, along with their variances. Suppose the raw observational numbers are as below, with the first three values from treatment A and the second 3 from treatment B (concretely, assume these are the sample test mpgs from gasoline A and gasoline B).

$$\mathbf{y} = \{29.8, 28.57, 29.97, 30.33, 31.27, 30.37\};$$

(I'll come back to these actual numbers in a moment after the theoretical discussion in the next section).

## Cut to the Chase

If you have limited time to investigate this tutorial in depth, let me summarize what you will find. See all of this visually from the figure below.

1.  $y_1, y_2, y_3$  are the observations from treatment A, and  $y_4, y_5, y_6$  are observations from treatment B.  $\bar{y}_1$  is the average of  $y_1, y_2, y_3$  while  $\bar{y}_2$  is the average of  $y_4, y_5, y_6$ .

$\bar{y}$  is the (grand) average of  $y_1, \dots, y_6$ .

2. Since the observation vector has 6 components, the natural setting for analysis will be a 6-dimensional vector space. In the very long run, the 6 components of the set of all possible  $\mathbf{y}$  vectors will average out to the true means of gasoline A and B. That particular 6 component vector will look like:  $\{\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2\}$ .

This means that components 1, 2, and 3 of observation vectors will (in the long run) average out to  $\mu_1$ , while components 4, 5, and 6 will average out to  $\mu_2$ . These  $\mu_1$  and  $\mu_2$  are the constant, but unknown mean population parameters. Since there are only 2 distinct components, this vector is embedded within a 2-dimensional space, called the *model* space. This model space holds all such possible true (constant) means.

3. To get an estimate of  $\mu_1$  and  $\mu_2$ , we project  $\mathbf{y}$  down into that model space and get the vector denoted as  $\hat{\mathbf{y}}$ , which consists of the averages of the observation values,  $\bar{y}_i$  shown below. So, it turns out that  $\bar{y}_1$  is the best (least squares) estimate of  $\mu_1$ , while  $\bar{y}_2$  is the best estimate of  $\mu_2$ . Note that I have drawn in the constant  $\{\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2\}$  vector in the diagram just to remind you of what we are trying to estimate.

If we can estimate the  $\mu_i$  though, then we can *estimate their difference* which is what the null hypothesis is all about.

4. To estimate the *difference of the true means*, I calculate the **treatment** vector. Notice how these components are simply the difference of the averages of A and B from the overall average.

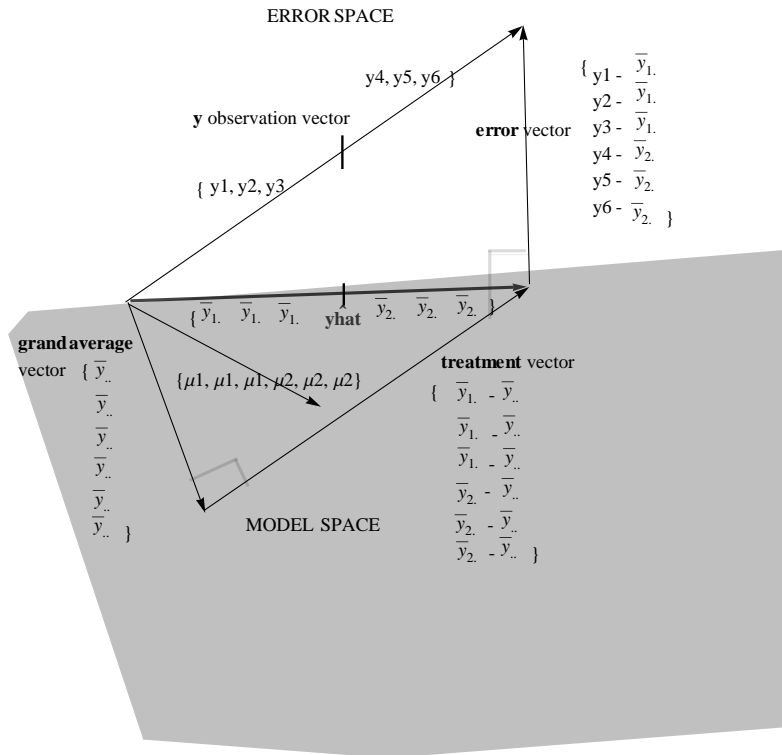
5. To estimate the inherent data variability (called the variance  $\sigma^2$ ), I calculate the **error** vector. Since the total vector space dimension is 6, and the model space occupies 2 dimensions, the error space must be 4-dimensional. In statistical language, the *dimension* of this error space is called the *error degrees of freedom (df)*, and equals 4 in this case. We will use the average squared length of this vector to estimate  $\sigma^2$ . This estimate is called the mean squared error (MSE).

6. Finally, to test whether the null hypotheses of equal means ought to be rejected or not, I calculate the F-statistic. This statistic is simply the ratio of the squared length of the treatment vector over the average squared error vector length. The averaging factor is the degrees of freedom. The F-test is a standard distribution whose values are in tables, keyed by the dimension of the numerator and denominator, which are "1" and "4" in our case.

$$F[1,4] = (\text{treatment vector})^2 / ((\text{error vector})^2/4)$$

That's it! Statistical tests come down to ratios of squared vector lengths. The algebraic formulas you see in the text books are equivalent to finding the squared lengths of the vectors shown here. (\*in this particular example gas milage case, this F-statistics will turn out *not* significant and so I would *not* reject the null hypothesis\*)!

■ \*\*\* End Cut to the Chase \*\*\* \*\* \*\* \*\*



■ **A detailed explanation of the procedure**

The following sections explain the underlying bases for the statistical tests by relating these tests to the directions of vectors and their lengths.

■ **The Model space (long run outcomes)**

In the very long run, if I keep sampling and getting observation vectors like  $\mathbf{y}$  above, *each* of the six components of those observation vectors will *average out* to the true population means,  $\mu_1$  and  $\mu_2$ . The first three components will average out to  $\mu_1$ , and the second three will average out to  $\mu_2$ . These 2 long run averages can be symbolically expressed as the 6-component vector as shown below. The vector space occupied by all these *possible* 'true mean' vectors is a 2-dimensional subspace of the full 6-dimensional vector space, and is called the *model* space. (Note: given that there are 6 components to this vector, this shows that we will need a 6-dimensional vector space to house all 6-component vectors, long run vectors, as well as sample observational vectors). The model space occupies just 2-dimensions of the overall 6-dimensional space since there are only two distinct components. That is, since the long run vector of means has only two different values,  $\mu_1$  and  $\mu_2$ , that translates to a 2-dimensional space as shown next. I will break up the vector on the left into two disjoint components that will show the structure of this 2-dimensional vector (sub)space.

$$\{\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2\} = \mu_1 \{1, 1, 1, 0, 0, 0\} + \mu_2 \{0, 0, 0, 1, 1, 1\}$$

This 2-dimensional space is spanned by the two unit basis vectors :

$$\mathbf{v}_1 = \{1, 1, 1, 0, 0, 0\} / \sqrt{3};$$

$$\mathbf{v}_2 = \{0, 0, 0, 1, 1, 1\} / \sqrt{3};$$

That is, I can write the true ( but unknown) model vector as :

$$\{\mu_1, \mu_1, \mu_1, \mu_2, \mu_2, \mu_2\} = (\mu_1 \cdot \mathbf{v}_1) * \mathbf{v}_1 + (\mu_2 \cdot \mathbf{v}_2) * \mathbf{v}_2$$

#### ■ The assumed structure of the 6 - dimensional space

So, the assumed structure of the 6-dimensional vector space housing all of the possible mpg values from both gasolines, is that the observation vector  $\mathbf{y}$ , consisting of the observed values, the  $y_i$ , can be written as the sum of the true means (the **model** vector) , plus an **error** vector.

The model vector is down in the 2-dimensional *model* space while the error vector is in the 4-dimensional *complement* space. Symbolically, letting the first three components represent the observations from treatment A and the second 3 from treatment B, I can write:

$$\{y_1, y_2, y_3, y_4, y_5, y_6\} = \mu_1 * \mathbf{v}_1 + \mu_2 * \mathbf{v}_2 + \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

where the  $e_i$  are normal deviates whose expected values are zero. If I knew the  $\mu_i$ , I could stop right here, but I don't. Since I don't know the  $\mu_i$  I will have to estimate them from the observational data, enter  $\hat{\mathbf{y}}$ , shown next.

#### ■ $\hat{\mathbf{y}}$ Fitting (estimating) the Model Vector - Projecting $\mathbf{y}$ down onto the model space

Using the  $\mathbf{v}_1$  and  $\mathbf{v}_2$  basis vectors above, I can get least squares estimates of  $\mu_1$  and  $\mu_2$  by simply projecting the observation vector down onto the model space. Symbolically, if  $y_1, y_2, y_3$  are observations from treatment A and  $y_4, y_5, y_6$  are from treatment B then the 'best' estimate of the true means is the  $\hat{\mathbf{y}}$  projection vector shown next.

$$\mathbf{y}_{\text{symbolic}} = \{y_1, y_2, y_3, y_4, y_5, y_6\};$$

$$\hat{\mathbf{y}} = \mathbf{y}_{\text{symbolic}} \cdot \mathbf{v}_1 \mathbf{v}_1 + \mathbf{y}_{\text{symbolic}} \cdot \mathbf{v}_2 \mathbf{v}_2 // \text{FullSimplify}$$

$$\left\{ \frac{1}{3} (y_1 + y_2 + y_3), \frac{1}{3} (y_1 + y_2 + y_3), \frac{1}{3} (y_1 + y_2 + y_3), \frac{1}{3} (y_4 + y_5 + y_6), \frac{1}{3} (y_4 + y_5 + y_6), \frac{1}{3} (y_4 + y_5 + y_6) \right\}$$

Notice the  $\bar{y}_1$  and  $\bar{y}_2$  components of the  $\hat{\mathbf{y}}$  vector. So, the  $\hat{\mathbf{y}}$  vector has components of  $\bar{y}_i$  as the best (least squares) estimates of  $\mu_1$  and  $\mu_2$ . Numerically, I get the numbers shown in the next section.

#### ■ Back to the actual observed vector

Now that you see what the role of the  $\mathbf{v}_1$  and  $\mathbf{v}_2$  basis vectors are, let me do an actual calculation to determine the best numerical estimates of  $\mu_1$  and  $\mu_2$ . At this point I gain some valuable information, namely that my *best estimate* of  $\mu_1$  and  $\mu_2$  are the values, 29.4467 and 30.6567.

$$\hat{\mathbf{y}} = \mathbf{y} \cdot \mathbf{v}_1 \mathbf{v}_1 + \mathbf{y} \cdot \mathbf{v}_2 \mathbf{v}_2 (*\text{the fitted model vector}*)$$

$$\{29.4467, 29.4467, 29.4467, 30.6567, 30.6567, 30.6567\}$$

The error vector would then be the difference between  $\mathbf{y}$  and  $\hat{\mathbf{y}}$

`error = y - yhat`

`{0.353333, -0.876667, 0.523333, -0.326667, 0.613333, -0.286667}`

So,  $\mathbf{y} = \mathbf{yhat} + \mathbf{error}$

### ■ Changing the basis of the Vector Space from $\mathbf{v1}$ and $\mathbf{v2}$ , to $\mathbf{u1}$ and $\mathbf{u2}$

Using the vectors  $\mathbf{v1}$  and  $\mathbf{v2}$  shown above, I could test the sizes of  $\mu_1$  or  $\mu_2$ , to see if they were (significantly) zero or not. This would work since the expected values of the components of the  $\hat{\mathbf{y}}$  vector are  $\mu_1$  and  $\mu_2$ . However, that is not of interest for this exercise although, knowing the relative sizes of the estimates of  $\mu_1$  and  $\mu_2$  already gives me some feeling for the treatment effects.

What I really want though is a coordinate system so that I can test the *difference* of the means

$\mu_1$  and  $\mu_2$  rather than their individual sizes. This is in accord with my interest in the *hypothesis* of testing for no differences. That requires that I choose a different set of two basis vectors that still span the model space but have directions that are statistically helpful. (Note: the numeric projection length components of  $\mathbf{y}$  down in the 2-dimensional model space remain the same no matter what basis set I use. That is, using  $\mathbf{v1}$  and  $\mathbf{v2}$  as a basis or using  $\mathbf{u1}$  and  $\mathbf{u2}$  as a basis won't change the values of  $\hat{\mathbf{y}}$ ). (I show  $\mathbf{u1}$  after this discussion).

One of those new basis vectors that I will choose,  $\mathbf{u2}$ , is down in the model space, and will have a *direction associated with the hypothesis* so that the *expected value* of the projection of  $\mathbf{y}$  down onto that vector will result in  $k * (\mu_1 - \mu_2)$ . This *length* then provides a test for the null hypothesis. If the projection length is near zero, then the null hypothesis is supported, otherwise not.

That is, if I can find such a vector  $\mathbf{u2}$ , then, when this projection length is short and near zero, then  $\mu_1 \approx \mu_2$ , otherwise, the length is long,  $\mu_1$  differs from  $\mu_2$ , and the null hypothesis is rejected.

### ■ Showing symbolically that 'u2' is the best vector to express mean differences

Let me show that by choosing  $\mathbf{u2}$ , and projecting  $\mathbf{y}$  down onto it, this results in a length that relates to the difference in *means* that I want to test. Let me look at this symbolically first to show how the components behave: Again assume the first three components,  $y_1, y_2,$  and  $y_3$  are the result of treatment A while the second three are from B.

`ysymbolic = {y1, y2, y3, y4, y5, y6}; (* the symbolic y vector *)`

`u2 = {1, 1, 1, -1, -1, -1} / Sqrt[6]; (* the proposed direction of the hypothesis *)`

`ex = ysymbolic . u2`

$$\frac{y_1}{\sqrt{6}} + \frac{y_2}{\sqrt{6}} + \frac{y_3}{\sqrt{6}} - \frac{y_4}{\sqrt{6}} - \frac{y_5}{\sqrt{6}} - \frac{y_6}{\sqrt{6}}$$

### ■ The projection length of $\mathbf{y}$ along the direction of 'u2' (where $\mathbf{u2}$ expresses differences in means)

This can be written as :

$$\sqrt{3} / \sqrt{2} * (\bar{y}_1 - \bar{y}_2)$$

and has *expected value* of :

$$\sqrt{3} / \sqrt{2} * (\mu_1 - \mu_2)$$

So, when this projection length is small, the null hypothesis is supported, otherwise not.

### A New Basis for the Model Space

Using  $\mathbf{u}_2$  as one of my new basis vectors, I need another orthogonal vector that will complete the 2-dimensional model space.  $\mathbf{u}_1$  will do this, as shown below, and has the merit of being interpreted as the overall sample average. Here then are the two new basis vectors that span the 2-dimensional model space. The scale factor of  $\sqrt{6}$  makes both vectors of unit length. This is useful since when  $\mathbf{y}$  is projected down on either one, that projection length is due solely to the contribution of  $\mathbf{y}$  in that direction. Remember that the scalar  $\mathbf{y} \cdot \mathbf{u}_1$ , for example, is defined as  $|\mathbf{y}| * |\mathbf{u}_1| \cos[\theta]$ , where  $\theta$  is the angle between  $\mathbf{y}$  and  $\mathbf{u}_1$ . But, since  $|\mathbf{u}_1| = 1$ , the result is  $|\mathbf{y}| * \cos[\theta]$ .

```

u1 = { 1, 1, 1, 1, 1, 1 } / Sqrt[6];
u2 = { 1, 1, 1, -1, -1, -1 } / Sqrt[6];
error = y - (y.u1 u1 + y.u2 u2) // N
{0.353333, -0.876667, 0.523333, -0.326667, 0.613333, -0.286667}

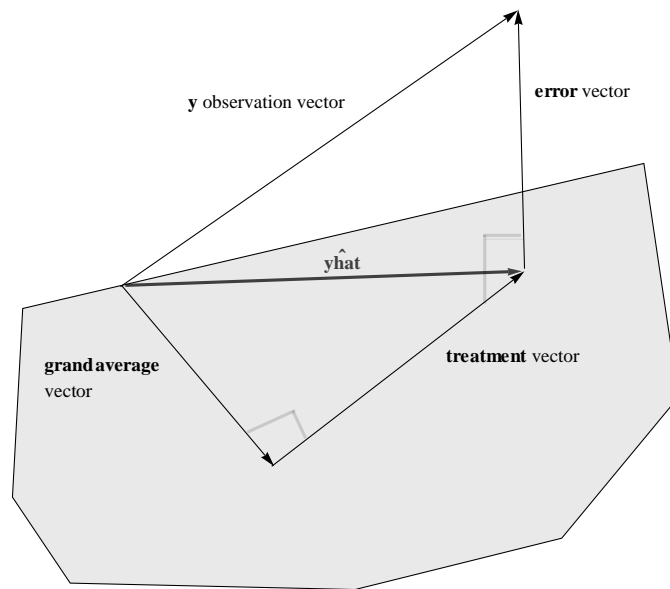
grandAverage = y.u1 u1
{30.0517, 30.0517, 30.0517, 30.0517, 30.0517, 30.0517}

treatment = y.u2 u2
{-0.605, -0.605, -0.605, 0.605, 0.605, 0.605}

```

To recap a little, notice that any orthogonal basis for the model space has the same projection of  $\mathbf{y}$ . That is, whether we use  $\mathbf{v}_1, \mathbf{v}_2$ , or some other basis like  $\mathbf{u}_1, \mathbf{u}_2$ , the projection of  $\mathbf{y}$  remains the same.

$$\hat{\mathbf{y}} = \text{grandAverage} + \text{treatment} = \mathbf{y} \cdot \mathbf{v}_1 \mathbf{v}_1 + \mathbf{y} \cdot \mathbf{v}_2 \mathbf{v}_2 = \mathbf{y} \cdot \mathbf{u}_1 \mathbf{u}_1 + \mathbf{y} \cdot \mathbf{u}_2 \mathbf{u}_2$$



### ■ A little more insight

I have shown that the model space is generated by either the pair  $\mathbf{v1}$  and  $\mathbf{v2}$  or by using  $\mathbf{u1}$  and  $\mathbf{u2}$ . That still leaves the *error* space unaccounted for! It must be 4-dimensional however since the total vector space has dimension six and two are accounted for by the model space. The 4-dimensional error space can be spanned by four more basis vectors. Next I show a possible set that will do this. Once I have these 4, I can combine them with  $\mathbf{u1}$  and  $\mathbf{u2}$  to get a basis for the whole space so that I can write any vector as a unique combination of something in the model space plus something in the error space.

### ■ A possible basis for the error space ( all are mutually perpendicular to each other)

These basis vector must be chosen so that they are mutually perpendicular to each other as well as to the model space vectors  $\mathbf{u1}$  and  $\mathbf{u2}$ . Given these bases vectors I can uniquely write the error vector as shown below. Try out the *dot* product of any of these vectors with any other basis vector and see what you get! Notice that  $\mathbf{u3}$  and  $\mathbf{u4}$  express the variation within treatment A, while  $\mathbf{u5}$  and  $\mathbf{u6}$  express the variation with treatment B.

$$\mathbf{u3} = \{1, -1, 0, 0, 0, 0\} / \text{sqrt}[2];$$

$$\mathbf{u4} = \{1, 1, -2, 0, 0, 0\} / \text{sqrt}[6];$$

$$\mathbf{u5} = \{0, 0, 0, 1, -1, 0\} / \text{sqrt}[2];$$

$$\mathbf{u6} = \{0, 0, 0, 1, 1, -2\} / \text{sqrt}[6];$$

$$\mathbf{error} = y \cdot \mathbf{u3} \mathbf{u3} + y \cdot \mathbf{u4} \mathbf{u4} + y \cdot \mathbf{u5} \mathbf{u5} + y \cdot \mathbf{u6} \mathbf{u6}$$

For completeness I can write any observation vector as a unique combination of the basis vectors as :

$$\mathbf{y} = y \cdot \mathbf{u1} \mathbf{u1} + y \cdot \mathbf{u2} \mathbf{u2} + y \cdot \mathbf{u3} \mathbf{u3} + y \cdot \mathbf{u4} \mathbf{u4} + y \cdot \mathbf{u5} \mathbf{u5} + y \cdot \mathbf{u6} \mathbf{u6}$$

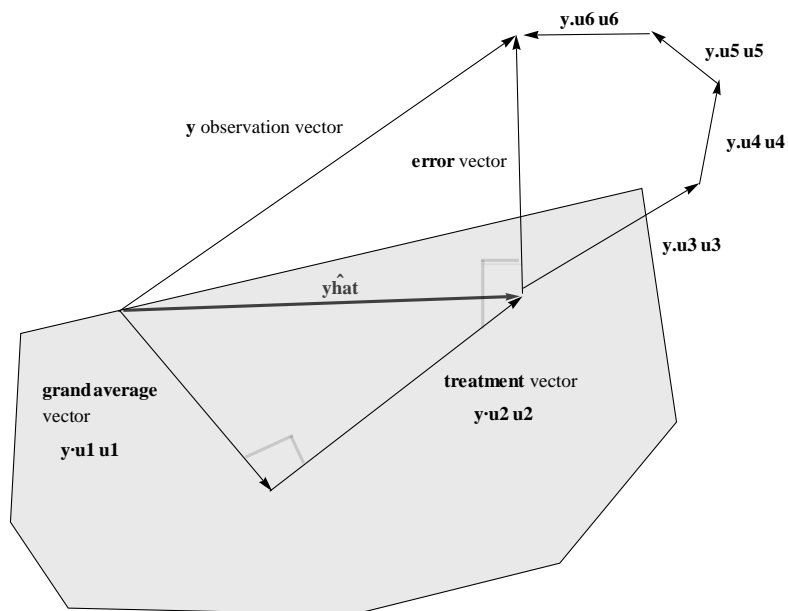
■ **Pythagoras has not left the building**

Check out the diagram below as well as the equations above, which are intended to indicate that the error vector itself can be broken down into mutually perpendicular components. This means that the Pythagorean theorem holds for the unique representation of the observation vector. ( this feature is the basis for ANOVA analyses as shown in the last section of this tutorial)

$$\mathbf{error}^2 = (y \cdot \mathbf{u3})^2 + (y \cdot \mathbf{u4})^2 + (y \cdot \mathbf{u5})^2 + (y \cdot \mathbf{u6})^2$$

Finally, you will note that :

$$y = (y \cdot \mathbf{u1})^2 + (y \cdot \mathbf{u2})^2 + \mathbf{error}^2$$





### ■ Hypothesis tests depend on the ratio of squared lengths

The intent of this experiment was to test whether the mean of the values resulting from treatment A was the same as those resulting from treatment B.

To repeat some of the salient vectors, I have:

```

y = {29.8, 28.57, 29.97, 30.33, 31.27, 30.37};
u2 = {1, 1, 1, -1, -1, -1} / Sqrt[6];
u2 = {1, 1, 1, -1, -1, -1} / Sqrt[6];
error = y - (y.u1 u1 + y.u2 u2)
{0.353333, -0.876667, 0.523333, -0.326667, 0.613333, -0.286667}

error.error
1.73233

```

### ■ The F - statistic for "1" and "4" degrees of freedom

```

F = (y.u2)^2 / ((error.error) / 4)
5.07096

```

### ■ The standard F - test values

Using the *Mathematica* package, I calculate the 95% critical value for the F - statistic, beyond which I would reject the null hypothesis. As it turns out, my F - statistic, 5.071, is less than this critical value, which is 7.708, and so I have no reason to reject the null hypothesis

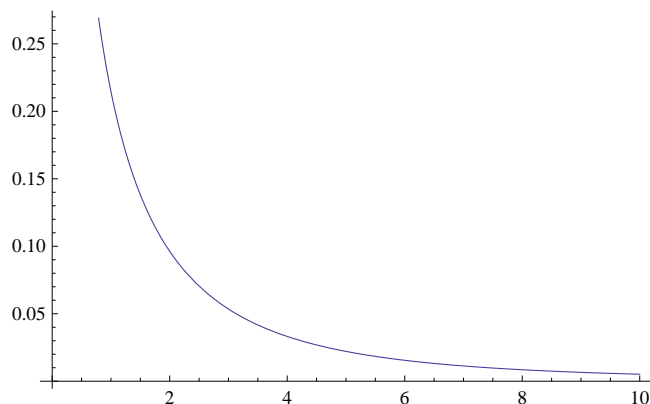
```

Quantile[FRatioDistribution[1, 4], 0.95] (*the value below which lies 95% of the curve*)
7.70865

```

### ■ A Plot of the F[1, 4] distribution

```
Plot[PDF[frat, x], {x, 0, 10}]
```



## The ANOVA table calculations

### ■ Calculating squared lengths for entries in the ANOVA table

```
totalSS = y.y(* total sum of squares= squared length of observation vector*)
5422.54

meanSS =
  (y.u1)^2 (*sum of squares of the grand average vector= squared length of this vector*)
5418.62

treatmentSS =
  (y.u2)^2 (*sum of squares of the treatment vector= squared length of this vector*)
2.19615

errorSS = error.error (* squared length of the error vector*)
1.73233

errorMS = errorSS / 4(* the average squared length of the error vector*)
0.433083

fStatistic = treatmentSS / errorMS
5.07096
```

### ■ Setting up the rows for the ANOVA table

```
r1 = {"source of \nvariation\n=subspace", "df\n=dimension of subspace",
      "SS\n=squared vector length", "MS\n=average squared length", "Fratio"};

r2 = {"mean", 1, meanSS, meanSS, "  "};

r3 = {"treatment", 1, treatmentSS, treatmentSS, fStatistic};
r4 = {"error", 4, errorSS, errorMS, "  "};

r5 = {"Total", 6, totalSS, totalSS, "  "};

v = {r1, r2, r3, r4, r5};
```

### ■ The ANOVA table

Note that the sum of squares of the mean + treatment + error vectors equals the sum of squares of the observation vector. This is a consequence of the Pythagorean theorem again.

Grid[v, Frame → All]

source of variation =subspace	df =dimension of subspace	SS =squared vector length	MS =average squared length	Fratio
mean	1	5418.62	5418.62	
treatment	1	2.19615	2.19615	5.07096
error	4	1.73233	0.433083	
Total	6	5422.54	5422.54	