Analyzing survey data: Contrasting three categories of data *alpha draft* [2009-04-16]

This tutorial introduces you to a way of analyzing your survey or secondary data if you want to inter-compare the *means* of *three* categories of data. For example, if your demographic data had three age categories and you asked your respondents for dollars spent on entertainment last year, those three populations of dollars spent could be analyzed to see if there were significant differences among the age categories. Each inter-comparison you set up can be expressed first as a static (null) hypothesis and, as you will learn below, can then be re-expressed as an dynamic vector equation to be solved and tested.
*The procedure presented here comes under the heading of Analysis of Variance - ANOVA.
There are two basic assumptions made for this and many other statistical experiments: 1) the populations come from underlying Normal Distributions, 2) every population has the same but unknown variance , σ^2 while the population means are unknown and may differ.

Cut to the chase

Check out the diagram shown next. It describes the underlying geometry for inter-comparing three categories of data. The objective is to test various combinations of population means against each other to see whether they differ or not.
In general, what is happening here is that the original observed vector of sample data values is being projected down onto various vectors whose directions represent statistical hypotheses of interest. The lengths of these projections form the basis of testing those hypotheses.
To make the actual tests of hypotheses, we will use what is called the F-test that tests ratios of the *squares* of these projection lengths.

Continuing the age-entertainment example

Consider three populations of dollars from the three age categories. It will turn out that you can ask two independent questions of these populations. That is, you can formulate and test two independent hypotheses. You can pick any one hypothesis question to ask, but, if you want to ask a second independent question, it will depend on the first.

For example, I first might want to ask whether the (dollar) mean of population 1 is equal to the mean of population 2 (the null hypothesis). This is asking if there is a difference between the two younger age groups. Having done this, I can then ask one more independent question. That one is determined to be: is the average of the two means of the younger age groups equal to the mean of the older group.

For the first question, asking if the mean of population 1 is *equal* to the mean of population 2, expresses what is called the *null hypothesis* . But, most importantly for us, this is the same as asking if the *difference* of the means is zero. The reason is that if it's expressed this way, it allows an *equation* to be written called a *contrast*. A contrast can express this zero difference and is the algebraic/geometric equivalent way of setting up to test a null hypothesis. Putting these words into formal statements, where $\mu 1$ is the true (but unknown) mean of population1 and $\mu 2$ is the true(but unknown) mean of population 2, I would write

$H_0 : \mu 1 = \mu 2$ (* this is the null hypothesis*)

$H_1 : \mu 1 \neq \mu 2$ (* this is the alternative hypothesis *)

c1 = μ1 – μ2 (* this is the associated contrast and testing it for zero is the same as testing for equality of the null hypothesis*)

So, to test the null hypothesis I would use the contrast **c1** to guide writing a vector equation using my sample data. I then test **c1** for zero, using geometric/algebraic analyses, and if it is zero, then so too is the null hypothesis confirmed. (In proper statistics language, I would be more careful and say that: *the null hypothesis is not rejected*). If however, **c1** is *significantly different* from zero, then I conclude that the null hypothesis can be rejected and the alternative hypothesis 'accepted'.
So, in summary, the procedure is to turn a null hypothesis into a contrast, a contrast into a solvable geometric vector equation (using sample data), and then test the magnitude of various ratios of vector lengths associated with that contrast. Those ratios constitute the statistical test for rejection or non-rejection of the original null hypothesis.

Background assumed

This tutorial assumes you know how to work with *vectors*, how to calculate a *dot product*, and have an idea about a *vector space* and its associated *basis vectors*. If you need to brush up on these topics you can find all this in tutorials on the milagrosoft.com site -
See - VectorOperationsQuickLook, TTestReference, BasicPhysicalStats, geoStatisticsPartI. I would also very highly recommend the book by Saville and Woods, *Statistical Methods: The Geometric Approach.* They in turn, were inspired by the foundational work of Sir Ronald Fisher and his *Statistical Methods for Research Workers.*

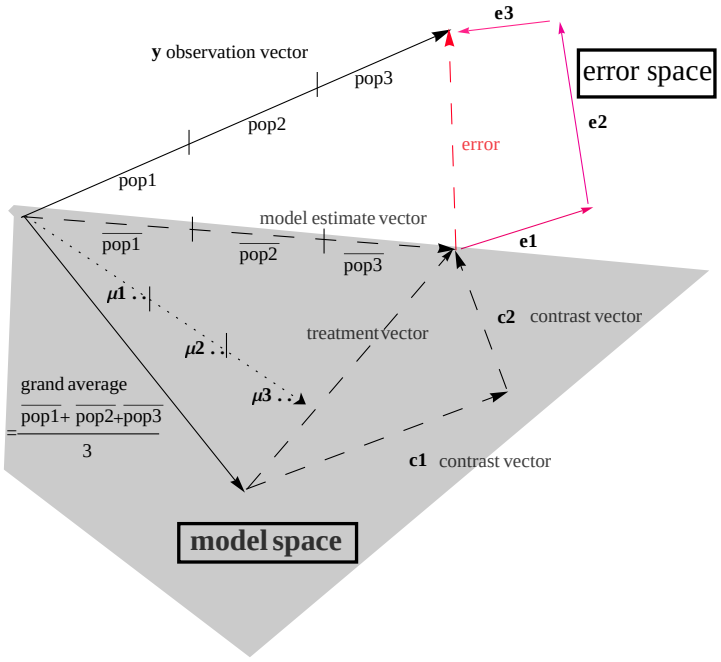
The geometry of the analysis, in general

Let me run through the components of the diagram you see below. (For those who need actual numbers to look at, check out the second diagram). This is a general diagram that would apply to all tests involving three populations. I will substitute in specific numbers in the next section, but for now, try to get a general sense of the *triangles*, *lengths* and *angles* that go into answering statistical questions.
First off, you have a list of numbers from each of the respective populations, this is represented by the observation vector, **y**. The first section of the vector is values from population 1, the second section from population 2, and the last section from population 3. As a first step you find the overall average of all of these values and express that as a vector with that average as each component. That is the *grand average vector*. Its' constant values are the best estimate of the overall average of all the populations combined. (Although an essential "baseline" kind of step, this overall average in itself is not usually of interest).
Next, the best estimate of the underlying population means, μ_1, μ_2, μ_3, is just to average each individual population value set and turn that into a vector as indicated by the *model estimate vector*. Those are indicated by $\overline{\text{pop1}}$, $\overline{\text{pop2}}$, and $\overline{\text{pop3}}$.
I have also drawn in a vector that would represent the true means, if I knew them, which I don't. I don't use this vector in any calculations, but
just drew it in here to emphasize that this is what I am trying to estimate, along with combinations of these true means.
The *treatment vector* carries the effects of the 'treatment' which is the difference between the model estimate vector and the grand average. Generally, this difference vector is not a fine enough discrimination for our purposes, so it needs to be broken down further. What I mean here is that this vector combines all of the treatment effects into one 'length', and even if it were significant, it wouldn't tell you *which* treatment or treatments contributed to the significant difference, that's the job of contrasts.
So, to answer more detailed statistical questions such as those asked in this tutorial, we decompose the *treatment vector* into component vectors. These component vectors are perpendicular to each other and thus support independent hypotheses. These vectors are the *contrasts* and represent hypotheses tests such as **c1** as described in the initial discussion above and **c2** that will be described later.

Notice the blue colored (shaded) area of the diagram, this is meant to represent the 3 - dimensional model space where all of the long run outcomes of the experiments are located. In particular, the true population means reside here as well as the grand average, treatment vector, and the various contrasts.

The ' red' vectors, e1, e2, e2, and error, represent the ' error' between the observation vector and the model estimate vector. That error vector, error, can be decomposed into three independent, perpendicular (orthogonal) vectors as well, e1, e2, and e3. These construct the 3 - dimensional error space.

So, the error vectors build out the remainder of the 6 - dimensional space that is the complement to the model space. The statistics faithfully follow the picture.



- **A few insights to be gained from the diagram**

Since **all** the triangles in this diagram are right triangles, the Pythagorean theorem *applies to every triangle you see*. So, the sums of squares rules apply to the triangles shown below.

**The equality of sums of squares for right triangles is the basis for Analysis of Variance (ANOVA) as you will see in that section.

Insight 0 : The whole vector space of 6-dimensions, that embeds the **y** observation vector, is partitioned into two mutually perpendicular 3-dimensional subspaces, the *model space* and the *error space*.

Insight 1: y observation vector= grand average vector + treatment vector + error vector

Note: given that the right hand side vectors are mutually perpendicular, Insight 2 follows from Pythagoras' theorem :

Insight 2: (*y* observation vector)^2 = (grand average vector)^2 + (treatment vector)^2 + (error vector)^2

Note: When we get to the section on *Analysis of Variance* the (*y* observation vector)^2 will be called Total Sum of Squares (SS_{total}), the (grand average vector)^2 will be called (SS_{mean}), (treatment vector)^2 will be ($SS_{treatement}$), while (error vector)^2 is known as (SS_{error}).

Insight 3: treatment vector = c1 + c2, and (treatment vector)^2 = c1^2 + c2^2

Insight 4: y- grand average = treatment vector + error
(* This difference vector is usually called the 'corrected observation vector, of the observation vector corrected for the mean'. Note that I haven't drawn this vector in the picture, perhaps you would like to do that - from the tip of the grand average vector to the tip of the observation vector*)

(y – grand average)^2 = (treatment vector)^2 + error^2

Insight 5: error^2 = e1^2 + e2^2 + e3^2

Insight 6: error^2/3 = best estimate of σ^2. This is called the pooled variance estimate and is called the sample variance s^2. When we start using the vocabulary shown in the text books, this s^2 will be called the mean square error (MSE).

Insight 7: (the length of vector **c1**)^2 /s^2 is the F-test and the size of that ratio determines acceptance or rejection of the null hypothesis represented by **c1**.
similarly (the length of vector **c2**)^2 /s^2 is the F-test for the null hypothesis represented by **c2**.

Insight 8: ** If I had drawn this diagram to scale, I could have read off lengths from the diagram itself to compute ratios for testing purposes. This is simply good engineering practice, where a diagram is as good as an equation. In our case we have both.

- **What combinations of means can I test?**

Given three populations of numbers you could, in theory, make six comparisons of their various means, that is, you could set up six different null hypotheses, together with their alternative hypotheses and so six *contrasts*. However, only two of these hypotheses will be independent within a particular experiment. In a given situation though, you can choose any contrast hypotheses/contrasts to set up, based on your particular knowledge and intent. Once you pick that first contrast though, the requirement that the second contrast be perpendicular pins down what *population means* that second contrast must compare. Here is a list of the potential hypothesis pairs where 1, 2, 3, represent the respective population means. That is, if you choose to test 1 versus 2 and another hypothesis as well, it must be 3 versus the average of (1and 2)

- **Pairs of independent hypotheses to potentially test**

{1 versus 2, 3 versus average (1 and 2)}
{1 versus 3, 2 versus average (1 and 3)}
{2 versus 3, 1 versus average (2 and 3)}
(you *could* also test whether or not the overall mean is zero, but this is rarely of interest)

Now, you get to pick which two hypotheses/contrasts to actually test. In the example considered below, I will first test the mean of population 1 versus the mean of population 2, that was contrast **c1**. For my second hypothesis/contrast, if I want it to be independent, I must test the mean of population 3 versus the average of the means of populations 1 and 2, and that is contrast **c2**.

■ **Illustrating the example with actual numbers**

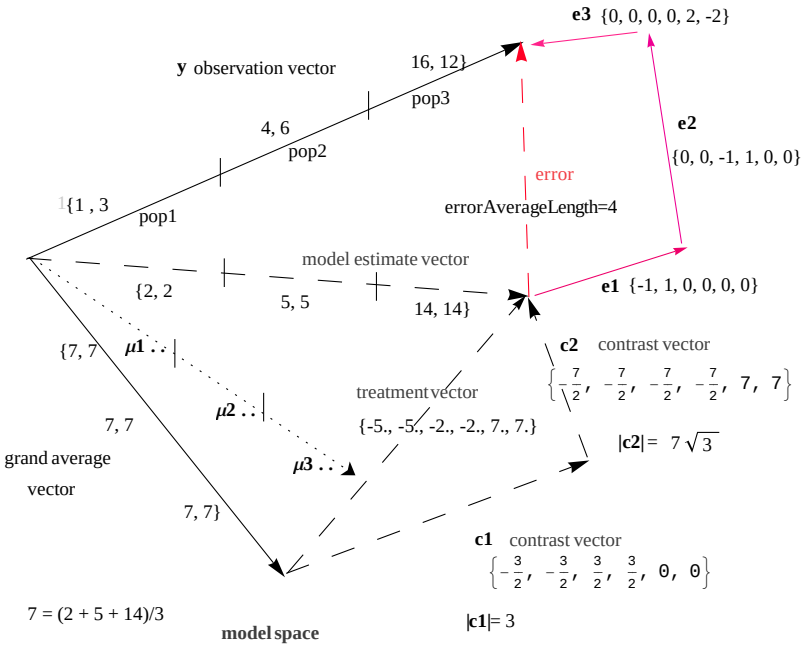
To get more specific, suppose I have 6 respondents with 2 each from three age *categories.* The age categories are: A=[20-30], B=[31-50], C=[51-70]. I also asked the respondents to indicate the number of dollars spent on entertainment during the past year. This then gives me a connection between age and dollars spent on entertainment.
The *populations* referred to here are the *numbers* representing the number of dollars spent by individuals in a given age category.
The observation vector shown below, and denoted by **y,** will have 6 elements matching age categories and dollar values. The first two numbers come from population 1, the second two from population 2, and the last two from population 3. Here are the dollar values (scaled). Obviously, older people have considerably more fun:

```
y = {1, 3, 4, 6, 16, 12};
```

So, let me denote the true mean of population 1 as μ1, the true mean of population 2 as μ2, and the true mean of population 3 as μ3. Since I don' t know these true means, I will have to estimate them from the sample data, that is, from the observed vector of sample data, **y.**

So, as things stand now, my *best estimate* of μ1 is the average of the data from population 1, which gives (1 + 3)/2 == 2, Similarly μ2 is best estimated by (4 + 6)/2 == 5 while (16 + 12)/2 ==14 estimates μ3. I could stop here and present these values as my experimental results. However, reporting that μ1 differs from μ3, while true for this particular sample, doesn't take into account the inherent variability of the data that can mask any actual differences if the experiment were to be repeated. That is, while this particular sample difference of the average of population 1 versus the average of population 2 is (2-5) == -3, this number may be misleading due to inherent data variability. What I actually want to test is whether this differences is *significant* in the long run, taking into account inherent data variability.

■ **The specific geometry of this example**



■ **The first hypothesis and its associated contrast, c1**

For my first hypothesis, I have chosen to test the dollar mean of the age category A, (population 1) versus the dollar mean of the age category B (population 2). The *null* hypothesis would be written as:

$H_0 : \mu 1 == \mu 2$

$H_1 : \mu 1 \neq \mu 2$ (* this is the alternative hypothesis *)

The equivalent contrast would look like

c1 = $\mu 1$ - $\mu 2$

And the test involving that contrast would ask if, when sample data is substituted for the true means, it approximates zero. So, substituting in actual data I could look at the difference between the means of population 1 and population 3 which is -12. Is that really significant given the small sample size and inherent variability to be expected? Read on!
Below is the unit length direction vector that expresses this mean difference

```
u2 = {1, 1, -1, -1, 0, 0} / Sqrt[4]
```

$$\left\{\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, 0, 0\right\}$$

The length of the contrast vector **c1** is:

```
c1length = (y.u2)
 (* this is the 'signed' length but usually I will take the absolute value of this number *)
```

```
-3
```

The actual contrast vector, **c1**

```
c1 = (y.u2) u2
```

$$\left\{-\frac{3}{2}, -\frac{3}{2}, \frac{3}{2}, \frac{3}{2}, 0, 0\right\}$$

y dotted with **u2** is actually the *projection length* of the observation vector **y** along **u2**. Notice that **u2** is in the direction of the hypothesis! That projection length reflects how close to zero the difference of the means is. In this case I get a value of -3. Is that significant enough to say that the underlying true means actually differ? (to save some suspense, it turns out that the square of this length divided by an average square of error vectors is the appropriate test to use and results in a ratio of :

```
f13 = (-3)^2 / 4.
```

```
2.25
```

Using the standard values associated with the F-test having 1 and 3 degrees of freedom, that value of 2.25 turns out to be insignificant at the 95% confidence level. Using *Mathematica*, I can actually calculate the critical value beyond which my result would be significant. So, the number 10.128 is what you would see in a table and means that numbers larger than this occur less than 5% of the time when the null hypothesis is true. Since 2.25 is way under this value, I have reason to reject the hypothesis that the two means are equal. (Keep this number in mind though since I will use it for the second hypothesis test as well).

```
Quantile[FRatioDistribution[1, 3], 0.95](* Mathematica calculation *)
```

```
10.128
```

```
10.128 (* the 95% value. Less than 5% lies to its right*)
```

```
10.128
```

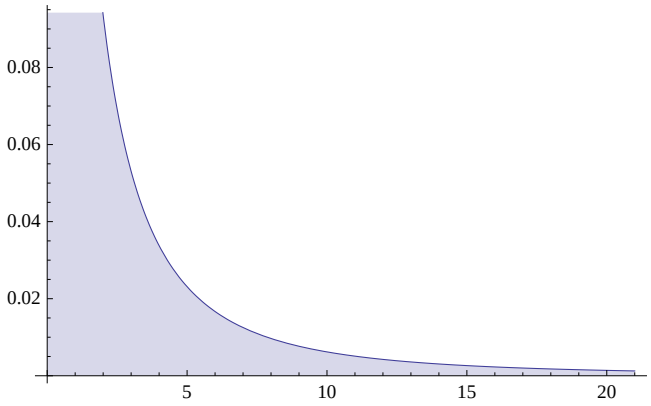
```
Quantile[FRatioDistribution[1, 3], 0.99]
(* Mathematica calculation for the 99th percentile. Less that 1% lies to its right*)
```

```
34.1162
```

- **A picture of the F[1, 3] distribution**

The ratio $\frac{(y.u2)\text{^}2}{\text{avgerageError}^2}$ follows an FRatio distribution with 1 degree of freedom in the numerator and 3 degrees of freedom in the denominator, formally called an F[1,3] distribution. The diagram below shows the theoretical distribution of this ratio under the assumption that **y.u2** is zero. So, if indeed y.u2 is zero, then we shouldn't be getting values beyond 10 or so. That is the test criterion.

```
Plot[ PDF[FRatioDistribution[1, 3], x], {x, 0, 21}, Filling → Axis]
```



- **The second hypothesis and its associated contrast , c2**

Now I want to ask a second question: I want to see if older age category Cs values significantly differ from the average of those of the youngsters A and B. So this is asking if the older people spend significantly more money on entertainment than do the young people.
The hypothesis to test, the null hypothesis, would be written as:

$H_0 : \mu 3 == (\mu 1 + \mu 2) / 2$

An equivalent contrast to use as a test would be :

```
c2 = μ3 - (μ1 + μ2) / 2
```

$$\frac{1}{2} (-\mu 1 - \mu 2) + \mu 3$$

Slightly rearranging this contrast makes it easier contrast to work with, since it involves only whole numbers:

```
c2 = 2 * μ3 - (μ1 + μ2)
```

$-\mu 1 - \mu 2 + 2\, \mu 3$

The test involving this last contrast is to see, if we substitute in sample data estimating the various means, whether of not it approximates zero. Remember, I don't actually have the true means to work with and so must estimate them from the sample data I do have.

- **A specific outcome**

So, below are the actual numbers I have to deal with. The question then is whether this sample difference represents a significant difference when
sample variability is taken into account?

The direction of the hypothesis is given by **u3**, which is one of the special basis vectors of the vector space That is, when **u3** is dotted into **y** , this result measures the difference between the mean of population C versus the average of the means of populations A and B (up to a scale factor).

```
u3 = {-1, -1, -1, -1, 2, 2} / Sqrt[12];
```

The projection *length* of the contrast **c2**

measures the extent to which the observation vector supports the null hypothesis. If the length, then the mean differences are small and thus supports the null hypothesis. If this length is relatively long, then the alternative hypothesis is given more weight.

```
c2length = y . {-1, -1, -1, -1, 2, 2} / Sqrt[12]
```

$7\sqrt{3}$

The contrast vector **c2** is :

```
c2 = (y.u3) u3
```

$\left\{-\frac{7}{2}, -\frac{7}{2}, -\frac{7}{2}, -\frac{7}{2}, 7, 7\right\}$

- **Testing the result with the F distribution**

The ratio $\frac{(y.u3)\text{^}2}{\text{avgerageError}^2}$ follows an FRatio distribution with 1 degree of freedom in the numerator and 3 degrees of freedom in the denominator, formally called an F[1,3] distribution.

```
f13 = (7 Sqrt[3.])^2 / 4
```

```
36.75
```

From the previous discussion, this ratio is highly significant and I can conclude that the younger people spend less money on entertainment than do the older people. (This number is actually past the 99% confidence level)

Further discussions of the vector space

At this point I assume you know what a vector space is and how a particular set of vectors can generate it. If you need to brush up on these ideas, go to the
VectorOperationsQuickLook tutorial on the milagrosoft.com site. That one will lead you more deeply into the wonders of *linear algebra*, as desired.

- **A particular basis for the 6-dimensional vector space that embeds 'y'**

Here are 6 vectors that form a basis for the 6 - dimensional vector space that holds the observation vector **y**. That is, any vector in the space can be expressed *uniquely* by these 6 basis vectors. The above discussion explained why I chose the first three vectors, **u1**, **u2**, and **u3**. These vectors generated the 3-dimensional *model* space and let me express estimates of the various combinations of the population means. The remaining three dimensions are spanned by three additional vectors, denoted **u4**, **u5**, **u6**. These three vectors build out the rest of the space, the last 3 dimensions, called the error *space*. These three basis vectors point in the direction of the 'errors'.

*Recall that perpendicular vectors mean that the cosine of the angle between them is zero or equivalently, that their dot product is zero.
The basis vectors break up into natural groups, one group constructs the so-called model space while the second group constructs the error space, which is the complementary space to the model space.

The way I have written them insures that each vector is of unit length and that all are mutually perpendicular to each other.

- **Model space vectors: @ 3-dimensions**

```
u1 = {1, 1, 1, 1, 1, 1} / Sqrt[6];(* grand mean direction*)

u2 = { 1, 1, -1, -1, 0, 0} / Sqrt[4];(* μ1 - μ2 direction *)

u3 = { -1, -1, -1, -1, 2, 2} / Sqrt[12];(* 2 μ3 -( μ1 + μ2)2 direction *)
```

- **Error space vectors: @ 3-dimensions**

```
u4 = {-1, 1, 0, 0, 0, 0} / Sqrt[2];(*error variation within population 1*)

u5 = { 0, 0, -1, 1, 0, 0} / Sqrt[2];(*error variation within population 2*)

u6 = { 0, 0, 0, 0, -1, 1} / Sqrt[2];(*error variation within population 3*)
```

Discussion of the model space basis vectors

- **'u1' points in the direction of the overall grand average**

The first vector, **u1**, points in the direction of the overall grand average. The scale factor of $\sqrt{6}$ makes this vector of unit length. The observation vector projected down onto **u1**, describes the overall vector representation of the average, '7'.

```
u1 = {1, 1, 1, 1, 1, 1.} / Sqrt[6];

y.u1 (* the projection coefficient of y down onto the  overall average direction *)

17.1464

y.u1 u1 (* the projection vector of y onto u1*)

{7., 7., 7., 7., 7., 7.}
```

Notice that the grand average, 7, is the sum of all the values divided by their number, which is 6. If I were interested in the mean of all the populations combined, this number would be my best bet. That is, the overall population mean, $\mu = (\mu 1 + \mu 2 + \mu 3)/3$ is estimated by '7'.

- **' u2' points in the direction of the hypothesis $\mu 1 - \mu 2$**

y dotted into **u2** is a length which is a measure of how different $\mu 1$ is from $\mu 2$.

- **' u3' points in the direction of the hypothesis of $2\,\mu 3 - (\mu 1 + \mu 2)$**

y dotted into **u3** is a length which is a measure of how different the mean of population 3 is from the average of the means of population 1 and 2.

The error space vectors

Each of the three error vectors, **u4**, **u5,** and **u6**, when dotted into **y**, represents a *length* , and the square of that length estimates σ^2.
That is, the squared lengths of each of these vectors individually estimates the underlying population's variance, $\boldsymbol{\sigma^2}$. Averaging these squares results in a pooled estimate that is the best estimate possible for $\boldsymbol{\sigma^2}$. This is also called the mean square error (MSE).
That is MSE = ($(\boldsymbol{y}.\mathbf{u4})^2 + (\boldsymbol{y}.\mathbf{u5})^2 + (\boldsymbol{y}.\mathbf{u6})^2$)/3 = 4

```
e1 = y.u4 u4
```

```
{-1, 1, 0, 0, 0, 0}
```

```
e2 = y.u5 u5
```

```
{0, 0, -1, 1, 0, 0}
```

```
e3 = y.u6 u6
```

```
{0, 0, 0, 0, 2, -2}
```

```
error = e1 + e2 + e3
```

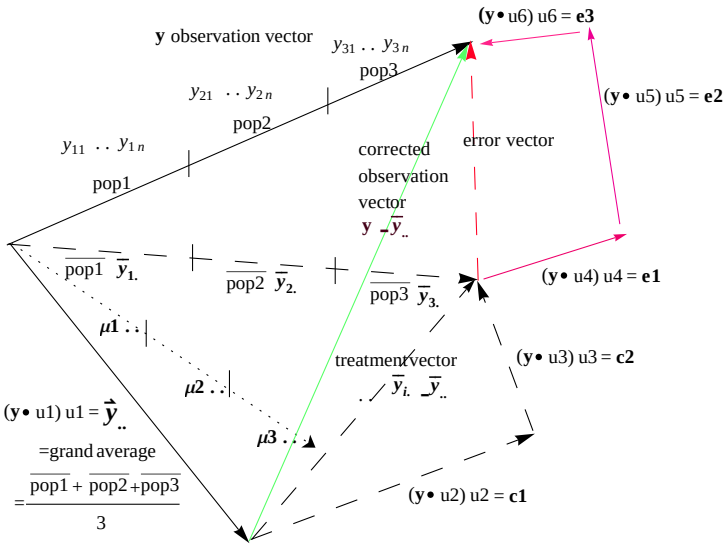
```
{-1, 1, -1, 1, 2, -2}
```

```
meanSquareError = (e1.e1 + e2.e2 + e3.e3) / 3
```

```
4
```

Generalizing the example and giving algebraic solutions

From the initial example you dealt with just a few easy numbers. Real questions generally have a lot more numbers. To handle these cases I have introduced some general notation. Now each population is indexed with 'i' and individual values within that population are indexed with 'j'. So I would write y_{23} for the third value of population 2 and y_{2n} for the nth value of population 2, assuming that 'n' is the total number of elements sampled from that population. Similarly for the other populations. There will still be only the 6 basic vectors but they will each be 'longer'.



Calculating the grand average (y.u1) u1, which consists of adding up all the values of the observation vector and dividing by the total number of elements, 3n.

$$\bar{y}_{..} = \frac{1}{3\,n} * \sum_{i=1}^{3}\sum_{j=1}^{n} y_{i,j}$$

Total sum of squares = $SS_{total} = 3* n * \bar{y}_{..}^{\,2}$

Calculating each population average

$$\bar{y}_{i.} = \frac{1}{n} * \sum_{j=1}^{n} y_{i,j}$$

Treatment sum of squares

$SS_{treatment} = n * \sum_{i=1}^{3}(\bar{y}_i - \bar{y}_{..})^2$

Calculating the error squared value

error ^{2} = e1^2 + e2^2 + e3^2= (y.u4)^2 + (y.u5)^2 + (y.u6)^2

$SS_{error} = \sum_{i=1}^{3}\sum_{j=1}^{n}(y_{i,j} - \bar{y}_{i.})^2$

MSE = SS_{error} / (3 * n - 3)

- **ANOVA tables go here (*** in progress ***)**

  ** ** working on this, be patient!! ** ** *

```
y
```

```
{1, 3, 4, 6, 16, 12}
```

```
v1 = {1, 1, 0, 0, 0, 0} / Sqrt[2]
```

$$\left\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, 0, 0\right\}$$

```
v2 = {0, 0, 1, 1, 0, 0} / Sqrt[2]
```

$$\left\{0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0\right\}$$

```
v3 = {0, 0, 0, 0, 1, 1} / Sqrt[2]
```

$$\left\{0, 0, 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}$$

```
u1
```

$$\left\{\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right\}$$

```
yhat = y.v1 v1 + y.v2 v2 + y.v3 v3
```

```
{2, 2, 5, 5, 14, 14}
```

```
yhat - y.u1 u1 // N
```

```
{-5., -5., -2., -2., 7., 7.}
```