

The Geometry of Statistical Decisions - Part I

Draft 2009-04-03

Introduction

The purpose of this tutorial is to introduce you to a consistent, *geometric* way of looking at statistical decisions, basing those decisions on the measures of *Normal* populations. That is, those populations of numbers that follow the *Normal Distribution*. (See the Normal Distribution in this tutorial).

If you need a refresher on some of the ideas in this tutorial, then check out the background documents listed below - all are on the milagrosoft.com website: Basic Trigonometry, BasicPhysicalStats, VectorOperationsQuickLook, and TTestReference. The follow-on tutorial is GeoStatisticsPartII where I will describe how to do multiple population comparisons, also known as Analysis of Variance (ANOVA). The same ideas of lengths and angles apply to multiple comparisons as to the single comparison discussed here.

I emphasize the *geometry of statistics* since the sample data you gather, the data you calculate, and the data you analyze, can all be represented by vectors (directed line segments), characterized by lengths and the angles between them. Statistical hypotheses, for example, *have an associated vector direction* and, the magnitude of that direction vector, relative to the magnitude of associated averaged *error* vectors, constitutes a statistical test. This test determines the rejection or non-rejection of the hypothesis. (Don't worry, I'll illustrate all this, but maybe you would first like to read the short section *Cut to the Chase* on page 1).

The take home idea is that *Statistical Questions depend on Directions and Lengths* you can visualize! The direction expresses the hypothesis, and, its' associated length, relative to an 'error' length, expresses that hypothesis' rejection criteria. I would suggest that you use the geometry you learn here to understand the algebraic formulas you see in your text books - it's just that knowing the underlying geometry will allow you to understand and follow those formula derivations.

So, I'm going to show how the geometric approach exposes the *lines, lengths, and angles* of a statistical experiment, making interpretation easier and more memorable. (As an aside: Can you remember any of the stats formulas you saw at one time in some textbook? Congratulations if you can, but for those of us who can't, I hope to show you how geometry will make stats visual* as well as understandable).

Cut to the Chase - Is the Population Mean Zero, or Significantly Not?

Below is an example of what I am talking about. I will revisit this same experiment in much more detail later, but for now, consider this the prototypical visual for making statistical decisions. This picture represents an experiment of taking two readings from the same *Normal* population, over and over again. See the diagram below.

Consider taking two readings of a thermometer, with the thermometer inserted in an ice and water solution, read and then withdrawn. Take these pairs of readings over and over, assuming that the population of readings follow a Normal Distribution.

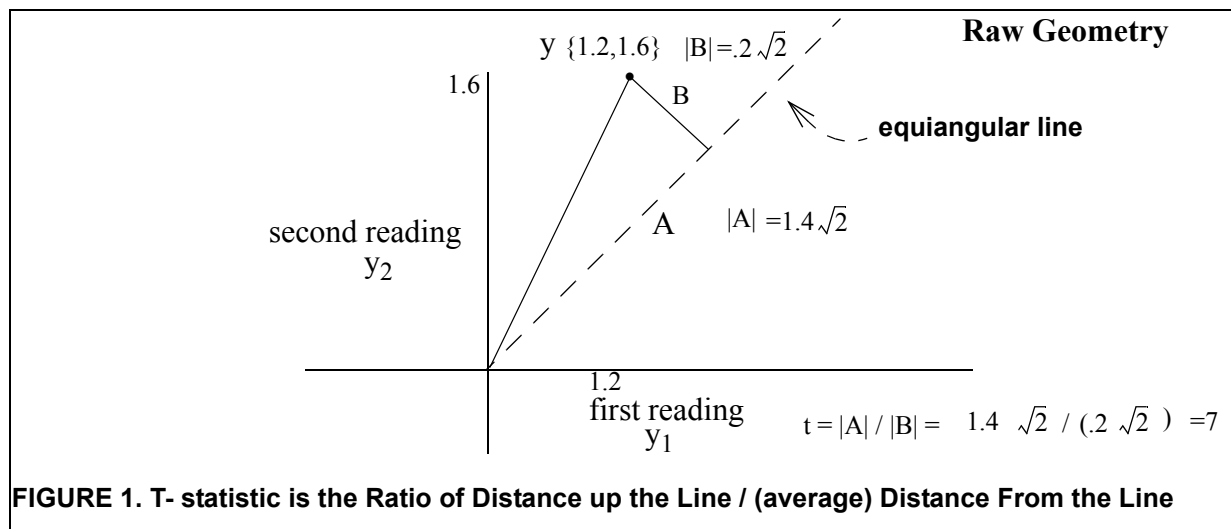
*Note: the thermometer *should* read zero under these conditions since an ice and water solution should be at zero (centigrade). If the thermometer doesn't read zero, there is some kind of error, either random or there is an actual bias away from zero, one way or the other.

The result of such pairs of readings will be a *cluster* of outcomes, as in the diagram below, clustered around some point in 2-Dimensional space (stay with me here). The hypothesis question is, where do the reading cluster? If the mean, that is, the overall population average μ , was approximately zero, then the readings would cluster about the point $\{0,0\}$, right? If the true population mean is not zero, then the cluster should move up the equiangular line away from zero and cluster around a point $\{\mu, \mu\}$ where μ is significantly different from zero. The *direction of the hypothesis* is along that direction of $\{1,1\}$, the equiangular line. The length of an observation vector, like the point $(1.2, 1.6)$, projected down on that equiangular line shows how far away that observation is from $\{0,0\}$. If the observation vector has a long length along this equiangular line, it's unlikely that $\mu = 0$ since the 'cluster', based on this one pair, seems to be moved up the line, away from zero.

Here I show only one possible pair of such readings, namely the observation vector $\{1.2, 1.6\}$. I am going to use this one observation to decide if the overall population mean is zero or is something *significantly* different from zero. Check out the diagram, and locate A, and B. If the ratio of the length of the line A, denoted by $|A|$ divided by the length of the line B, denoted by $|B|$, exceeds a tabulated value, then the (Null) hypothesis that $\mu=0$ is rejected, otherwise no rejection. That ratio of line lengths is called the t-statistic while its' square is called F-statistic. (more on this later).

In this case, the ratio of '7' doesn't give sufficient reason to reject the (null) hypothesis of no bias since a t-test reference value with one degree of freedom gives a rejection value of +12.7.

Statistical tests come down to comparing ratios of lengths (or squares of lengths) of vectors with tabulated values.



General Commentary

So, data vectors, the *vector spaces* they generate, and the pictures they create, give a coherent backdrop for the kinds of decisions that can be made within the statistics paradigm

A 'paradigm' is a mind set that tells me *which* questions are important to answer, *how* to go about answering them, and *what* a 'legitimate' answer will look like. As we look at the 'statistical paradigm' questions in this tutorial, you will see which kinds of questions are asked (is the population mean zero?), how questions are approached (draw a sample and

construct a test statistic like the t-statistic or the F- statistic to answer the question), and what constitutes an answer (a rejection or a non-rejection of the hypothesis that the population mean is zero, with perhaps an accompanying confidence interval).

The other foundation of this tutorial is the use of the Normal distribution. It provides the most important (and simplest) of the *reference* distributions in statistics. It is not only a good approximation to a lot of the observation data sets you will encounter in your research work, but, by taking averages of your raw observations, the Normal is the distribution that these averages will tend to. This tendency of *any* initial distribution, under sample averaging, to morph into the Normal distribution is the *unique* property of the Normal. For these and other reasons that I will introduce as we go along, the Normal makes a valuable and consistent basis for *all* our statistical investigations. I use the properties of the Normal to set up experiments as well as interpret their results. Very importantly, for these tutorials I will assume that all samples are drawn from Normal Distributions with unknown means and unknown variances, but I assume that while all variances are unknown, they are equal. More advanced work will relax this restriction.

Is the Mean of a Distribution Zero

In this section, an elaboration of the *Cut to the Chase*, page 1 discussion, I will test whether the mean of a distribution of numbers is zero or is not zero. I'll illustrate the standard tests: the t-test and the F-test, in the simplest possible manner. That means taking samples of size *two* from a Normal population. After this introductory example, I will then take samples of size three and test again for a mean of zero. Finally, to show this approach in more generality, I'll take a sample size of 5 and perform the same analyses and tests. To do these analyses I will use the example introduced below, expanding it to show additional features of the geometric approach as more dimensions are introduced.

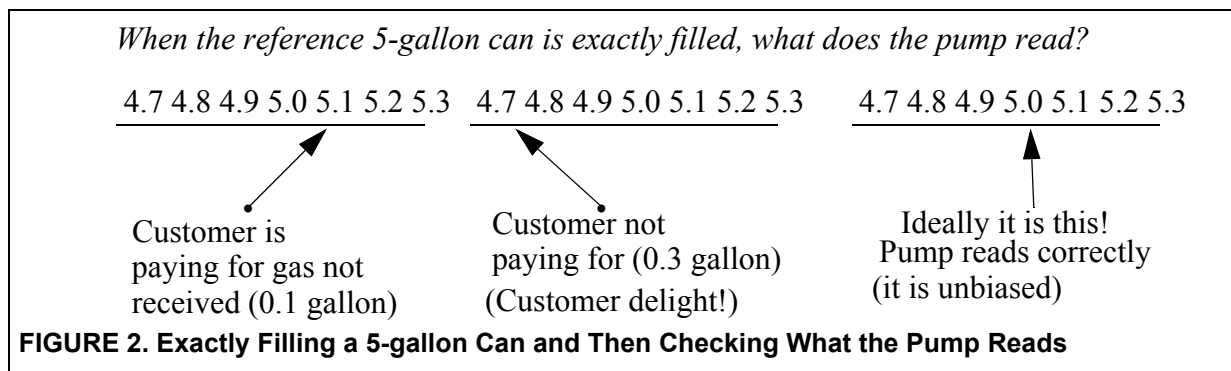
Example - Department of Weights and Measures

The Arizona Department of Weights and Measures is charged with certifying the accuracy of various scales and pumps that serve the Phoenix valley public. Suppose we work for this department and are interested in checking that a particular gas station pump gives true readings. That is, it is *unbiased* so that when the pump meter reads '5.0' gallons, it really has pumped exactly 5 gallons. We will test this pump (in an innovative fashion) by repeatedly *exactly filling up a 5 gallon can* and seeing if the pump meter reads 5 gallons on the nose, or perhaps above or below that value. A single reading of course, can be influenced by a lot of factors such as hose imperfections, temperature, humidity, or even grade of gas, so we expect some variability and decide to repeat this procedure on subsequent days.

What we come to realize as we do this work, is that we have to make judgement calls by analyzing *clusters of readings*, that is, we are back to analyzing batches of numbers (EDA principles)! We never get the exact same readings, we get clusters of readings. So, we end up making judgements about where that cluster is centered. In the pump example, the question is: are the reading deviations from the standard 5-gallons centered on zero, or is the cluster center significantly away from zero?

Over time and repetitions, these pump deviation readings could average out close to zero, above, or below. If the deviations average out *above* 5 gallons then the customer is being overcharged, while a long run average of *under* 5 gallon readings would lead the station to lose legitimate revenue. That is, since I exactly fill a five gallon can and then look at the pump reading, if the *reading* says more than 5 gallons the customer is overcharged, while a reading of less than 5 gallons means the customer is getting 5 gallons but is being undercharged. So, to summarize the experiment, if

the deviations are centered at zero, then the pump is a ‘good’ one that is unbiased and represents fair value to both customer and merchant, otherwise someone is getting short-changed. The hypothesis that the true mean is zero is called the *null hypothesis* and the *alternative hypothesis* is that the true mean is not zero.



The Testing Protocol - Taking Pairs of Readings

To make the math and the diagrams easy to follow, assume I take only two measurements of deviations from a nominal 5-gallon reading, at two different times each day, over multiple days. Further assume that these deviations will follow a Normal distribution with an unknown mean and common variance, σ^2 . (This is often a justifiable assumption when errors are the results of multiple causes, each contributing a little bit). When I say unknown *mean*, I allow the possibility that the scale is actually biased either high or low, and, in the long run, the mean of the deviations might be non-zero. That’s what we want to check.

The diagram below (Figure 3 on page 5) Part (a), illustrates a typical experimental outcome where the two deviation readings, y_1 and y_2 , come from a distribution with a true mean of zero, even though, individually, they themselves are not zero. That is, I am proposing that they come from an unbiased pump meter whose *long run deviations around ‘5’, average to zero*. Part (b) shows what happens when I take a large number of such pairs of readings. I would expect these readings to cluster around $\{0,0\}$. That is, the long-run mean (μ) of *each component* of the pair of readings is zero.

This is a subtle statement that is not immediately obvious (at least it wasn’t to me). What I am saying is this: think of doing the readings over and over. What will the *average* of the readings head toward? If the readings really come from a normal distribution of deviations, then in the long run these readings will average out to the mean of that distribution. Think about that! Now, this is true for *each component* of the readings. So, for the pair of readings, each component, in the long run, will average out to the mean of that same distribution. This is why I am saying that the cluster is centered on the point $\{\mu, \mu\}$ in 2-dimensional space. Where μ is the true mean that the readings are averaging toward. This also explains why the equiangular line is so important since, whatever the long run average deviation, the pair of these ‘final’ values lie somewhere along the equiangular line. To emphasize this idea, suppose you took 50 readings a day, over days and days, then the long run averages of each component of this 50-component vector would lead to a point in space with 50 equal components of $\{\mu, \mu, \dots, \mu\}$, again lying on a 1-dimensional equiangular line that is now embedded within 50-dimensional space.

As an alternative to a zero distribution, Part (c), shows readings coming from a distribution with an assumed true mean of 1.5. (That’s what I am assuming here just to show the method and it’s variations). In this case, since the long run average of each component is $\mu=1.5$, the outcome points

will cluster around the equiangular line (the 45° deg. line whose coordinates are equal). That is, the pairs of readings will cluster around the point with coordinates of $\{\mu, \mu\}$, which in this case is $\{1.5, 1.5\}$. That is the message of Part (d) of the diagram.

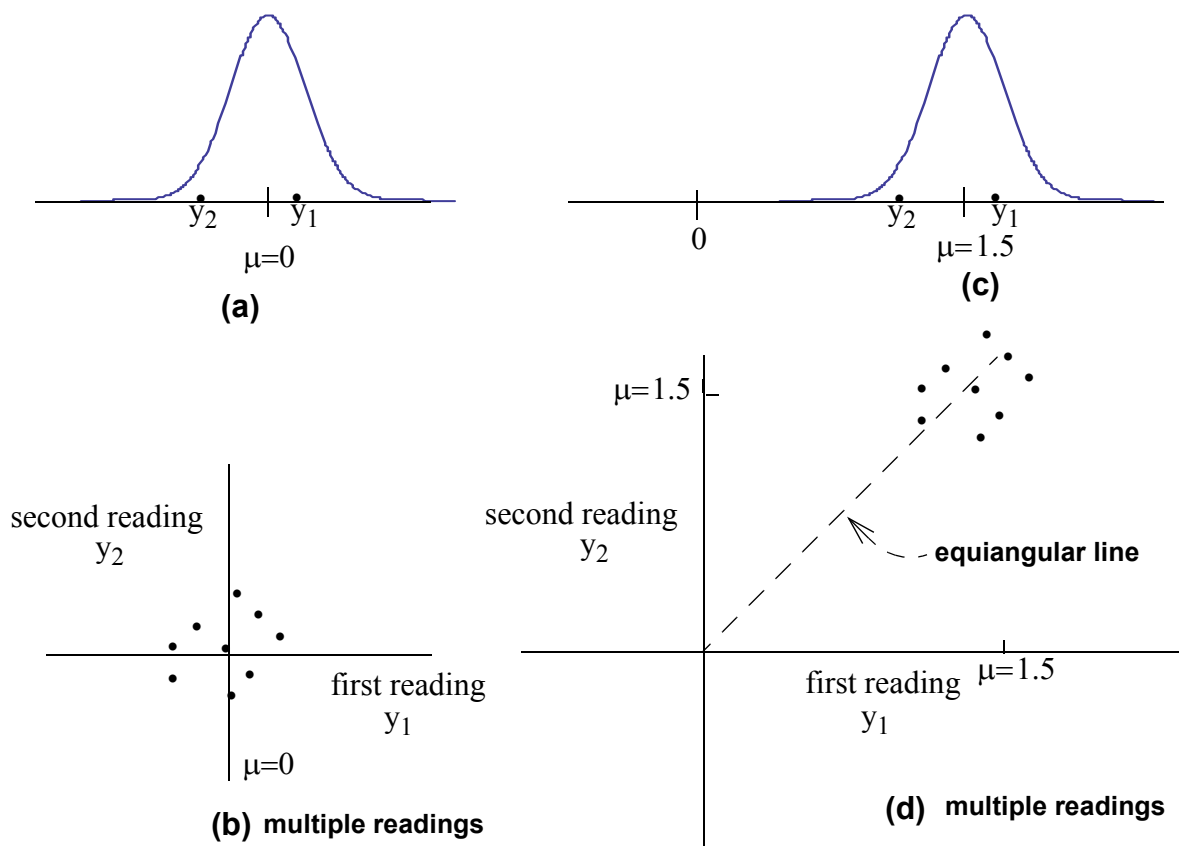


FIGURE 3. Clusters of pairs of readings around a true μ of zero and for an alternative $\mu = 1.5$

Naturally, we don't know the actual true mean and must estimate it from our sample readings. Suppose we take a pair of reading and get outcome deviations from '5.0' as the 'y' vector shows below. Note that I have scaled these deviations by '10' just to make the math easier to check. That is, "1.2" means 0.12 gallons.

$y = (y_1, y_2) = \{1.2, 1.6\} = \{6/5, 8/5\}$. This means that while the 5-gallon can was exactly filled, the pump readings were 5.12 and 5.16. That is, the customer was being overcharged.

Does this suggest that the scale is biased? Although these two values do suggest the chance of bias, we also have to take into account the variability of the 'y' data itself which could very well mask any long-run deviations from zero. The vector 'y' has two components, each assumed to be drawn from the *same* distribution of deviations. Further, I will assume that the variability (variance), of the deviations, denoted by σ^2 , stays the same from observation to observation, which is a pretty reasonable assumption here.

We need what is called a *test statistic* to help us distinguish the case where μ is zero (think of the cluster as being centered at $\mu = \text{zero}$), from the case where μ is really not zero (the cluster is centered significantly away from $\mu = \text{zero}$).

What Kind of Test Statistic Do We Need to Discriminate Zero From Non-Zero?

A test statistic won't be an absolute discriminator, but will give more or less strong indications as to whether our *null hypothesis* is true, that is, whether the cluster of readings is zero centered.

Here is where the geometry kicks in: If μ really were zero, then the cluster of points shouldn't move *very far* up or down the equiangular line, and ought to cluster around $\{0,0\}$ as was shown in Part (b) of Figure 3 on page 5, above. If, on the other hand, the cluster *moves up (or down)* a ways along the equiangular line, then this might suggest that the true mean is not zero (depending on how far it moves and also depending on the inherent variability of the deviations, which could swamp any actual difference).

To iterate a point (again!) that often confuses new researchers, is the phrase in hypothesis testing which says: "the true mean is zero versus, the true mean is not zero". What this is getting at is that the *distribution/cluster/cloud* of outcomes is either clustered around zero or is clustered around some other significantly different value. To say that the true mean is zero doesn't mean the outcome has to be *exactly zero* but only that the experimental outcomes *cluster* around zero. The intent of an experiment that attempts to determine the true mean, comes down to checking *where* the center of the cluster of outcomes lies (as well as accounting for inherent variability of the observation vector as a possible masking factor).

It turns out that the statistic that we need to distinguish $\mu = 0$ from $\mu \neq \text{zero}$ is a *ratio of distances*. The *distance up/down the equiangular line* of the sample point (called ' |A| ' in the figure), divided by the *average distance from the equiangular line* of the sample point (called ' |B| ' in the figure), provides just what we need. It's merits are that it doesn't depend on the units we choose to measure distances, in this case 'gallons', plus, the ratio is small when μ is zero and large when μ is not. This is the essence of a test statistic: it is 'large enough' or 'small enough' to swing our opinion one way or the other as to where the center of a distribution (a *parameter* of the distribution) lies. As a look forward, this same general idea will be used to swing our opinion as to where any *parameter* of one or more distributions lie.

The T-Test Finally Revealed

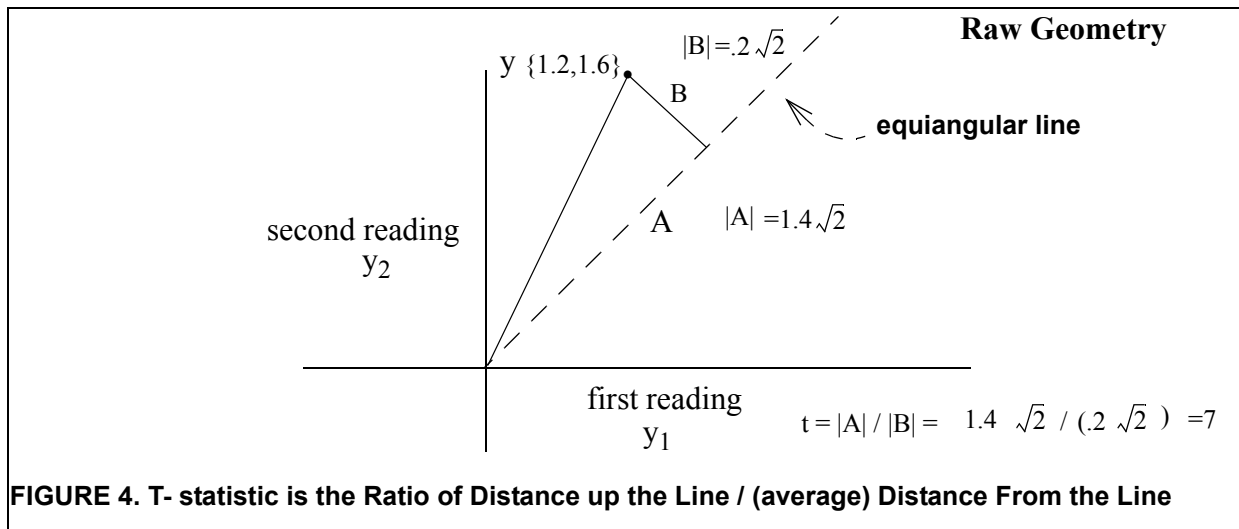
This ratio of distances (signed length of A /absolute length of B) *IS* the T-Statistic and the distribution of these ratios *IS* the T-Distribution (when the true mean μ is actually zero)! Once I have this ratio of lengths, I check that number against a standard 't-table' to see if that ratio is big enough for me to reject the hypothesis that the pump is unbiased. I will go into delightful detail in the next section about how to calculate and think about these numbers, but for now though, revel in your new found insight that the big bad t-test is just a simple ratio of distances. The numerator is the distance *up* the equiangular line of the cluster center while the denominator is a measure of the distance *away* from the equiangular line. That distance away from the line represents the potential to mask the actual true mean of the 'y' observations. As a bonus, the big bad 'F' test is simply the *square* of the t-test statistic. One is the square of the other, that's their relationship.

Warning: Later you will learn that it is not the raw distance of the sample point from the line, but rather the *average* distance of the sample from the line, as you will see in the next example. Taking this into account, the denominator distance is still a *distance from the equiangular line*. That average distance is usually denoted by 's', the sample standard deviation. We will see all this next.

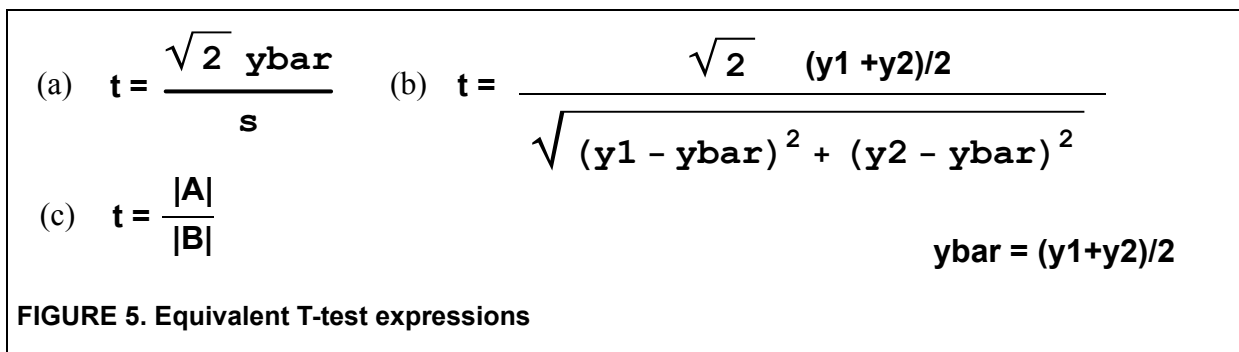
O.K - Raw Geometry Lets Us Really Look Under the Hood

Take a look at the figure below: the distance up the line is given by the length of the vector A, denoted by |A|. The distance away from the line is the length of the vector B, given by |B|. The t-statistic is written as a ratio:

$|A| / |B|$, and is equal to 7.



In traditional texts, this ‘t’ ratio is written as below in (a) or (b). Our geometric interpretation is shown as (c). Notice that ‘s’ is identified with the length of B, while the Sqrt[2] properly belongs with the ybar since (Sqrt[2] * ybar) is the distance up the line.



Testing the Result of the Two-Dimensional Experiment

To see if this value of the t-test statistic is unusual, we look in the reference table for the t-distribution. For a two sided test, with samples drawn from a Normal distribution, and for t with 1 degree of freedom, at 95% confidence.

Note: The *degrees of freedom* value for the t-test equates to the dimension of the error space in which the ‘B’ vector lies. In our case this dimension is ‘1’ and so the degrees of freedom is ‘1’. I will say more about this parameter as we look at higher dimensional experiments. But in general, the degrees of freedom always equals the *dimension of the sub- space under discussion*.

The t-table shows 12.7, as a critical value beyond which we would reject the null hypothesis. Since our calculated value is ‘7’, that’s *not big enough* to reject the null hypothesis that the true mean is centered on zero.

This means we don’t have enough evidence to reject the null hypothesis claim that μ is zero. So we report to our boss that the pump seems to give ‘honest’ unbiased readings, as far as this very limited experiment goes.

We also remind the boss that two readings are not much to base this conclusion on and so request more funds to do more extensive experiments (note: always request more funds and resources)! Since we know a little statistics, we finally get authorization to take more extensive readings, a process that is detailed in the next section “Taking Three Readings at the Pump” on page 10.

Optional insight! The degrees of freedom for the t-test equates to the dimension of the error space. That number tells how many independent basis vectors generate the error space. When I project the observation vector ‘y’ down onto *each* of these error space basis vectors, the square of each projection length is an *unbiased* estimate of the variance. *Pooling* all of these estimates together gives me the best estimate for the variance, σ^2 . Pooling consists of dividing the sum of the squared projection lengths by the number of basis vectors, which is the dimension of the error space, which is conveniently called the degrees of freedom -Now *you* know what degrees of freedom for the t-test means!

Recasting the Experiment - Vector Spaces

Embedding our analyses and syntheses inside of vector spaces means we have ready-made structures and operations from which we can construct, for example, various subspaces such as error spaces and model spaces. So, working in the rich context of a vector space will give us the basis for the statistical decisions with associated T-test and F-test geometric insights to come. Specifically, we will see that certain directions within the vector space represent the *directions in which hypotheses can be tested* while other directions within the space let us estimate potential *masking variability* that could cause us to miss legitimate conclusions.

So, let me re-present our experiment using the idea of vectors embedded inside of a vector space. You are already familiar with the geometric lines and triangles of ‘y’, ‘A’, and ‘B’, from the above discussion, but now I’ll explicitly consider them as vectors as in the diagram below. The *vector* ‘y’, for example, has two components and so can be thought of as embedded within a 2-dimensional vector space. Using vector ideas I can calculate the statistical t- tests and F-tests.

Constructing the Vector Space: Model Space and Error Space Vectors

O.k, I have embedded my ‘y’ observation vector inside of a 2-dimensional vector space. To allow me to do detailed analyses of the size and direction of the observation vector ‘y’, I will first break up the overall 2-dimensional space into two *1-dimensional subspaces*. and then allocate (project) y into those two complementary spaces. I will need two ‘basis’ vectors to generate those two subspaces. These are chosen with the *experimental hypotheses* in mind. In the test for the mean, one required basis vector would be the equiangular vector, representing the ‘mean’ hypothesis direction. That is the direction along which the vector $\{\mu, \mu\}$ lies and that we approximate by the length of the part (projection) of the y vector that lies along its direction.

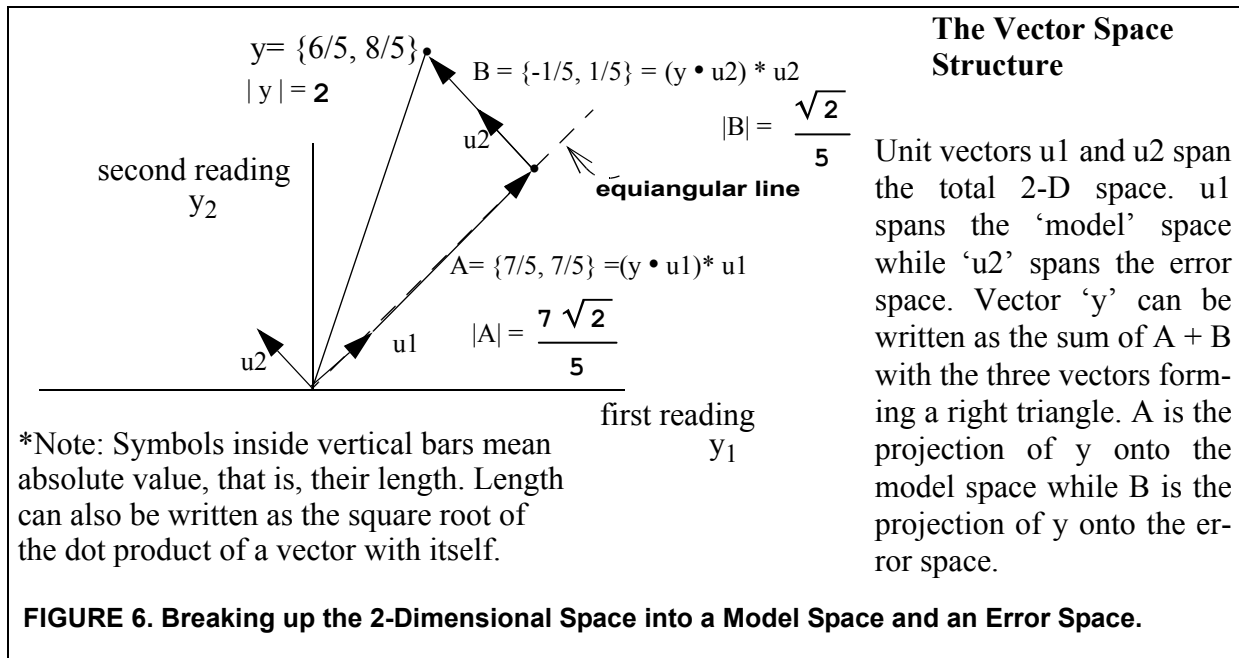
So, I construct a special vector (I will call it ‘u1’) that generates this 1-dimensional space along which the true mu lies. This 1-dimensional space is called the *model* space.

The other basis vector (called ‘u2’) generates or spans a complementary (perpendicular) 1- dimensional space called the *error* space.

The plan will be to write the observation vector as being uniquely made up of a vector in the model space plus a vector in the error space. I will use the length of the projection of y along the model space vector ‘u1’ to estimate the mean, while I use the length of the projection along ‘u2’ to estimate variability.

From Figure 6 on page 9, below, I can see that ‘u1’ points along the equiangular line, the one at a 45 degree angle, while the basis vector ‘u2’ points at right angles to that. The A vector in the dia-

gram, is the *projection vector* of y onto the space spanned by ‘ u_1 ’, while B is the projection vector of y onto the space spanned by ‘ u_2 ’. This way, I can look at y as made up of an estimate of the true mean (the vector in the model space) together with an error component (the vector in the error space).



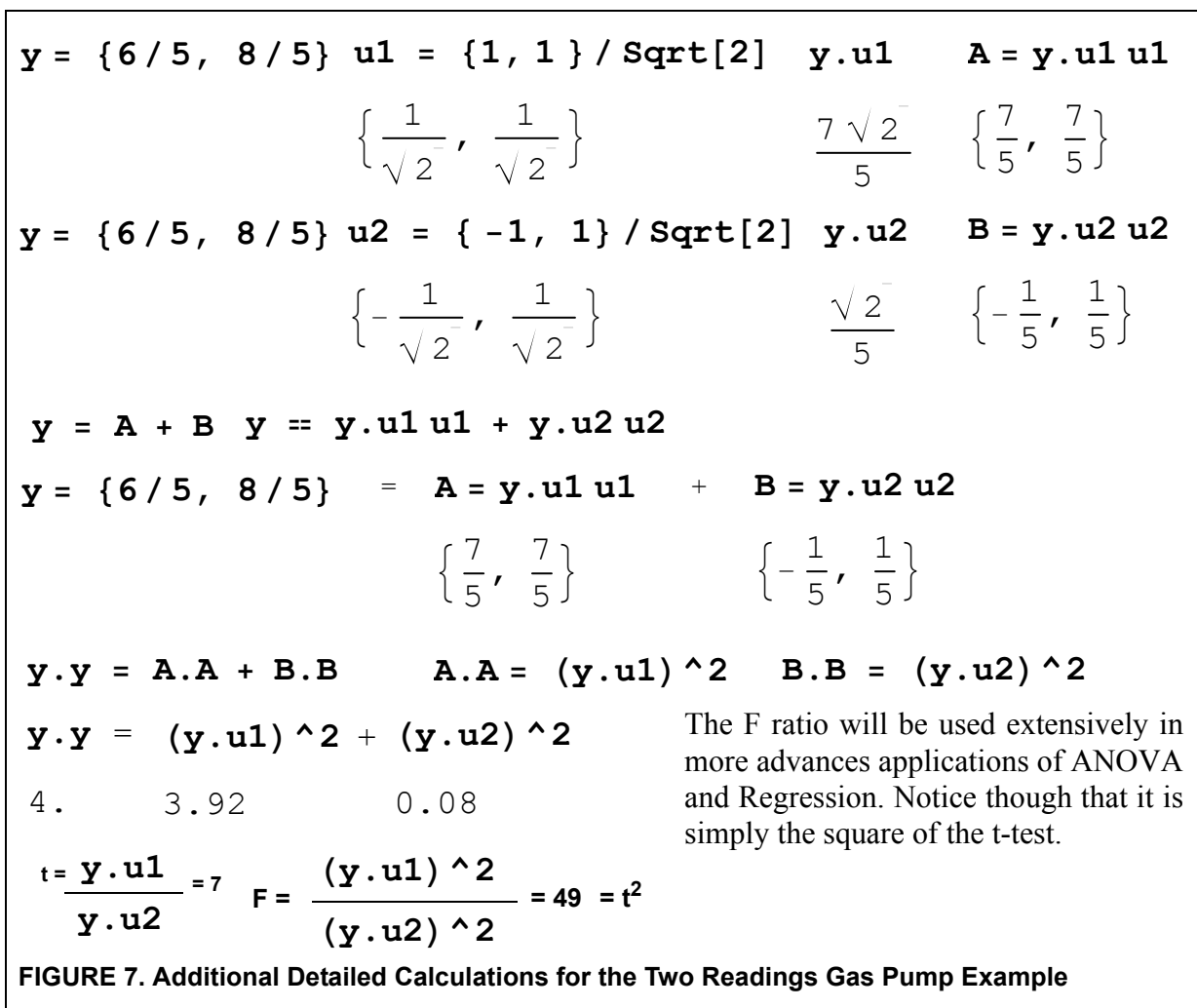
Vector Details for the Pump Example

For even more detail, I have written down all of the various vector components used in the calculations for the t-test. This is the content of “Additional Detailed Calculations for the Two Readings Gas Pump Example” on page 10 shown next.

Let me list what you will see in this figure and give a description of its usage.

- y - of course is the observation vector, which is all we have to work with. Since it has two components, it requires that I set up a 2-dimensional space to hold it.
- u_1 - is a vector of unit length (called a ‘unit vector’) with two equal components that point in the direction of the equiangular line, which is called the *model* space. Multiples of u_1 actually can be thought of as generating the model space. This equiangular *direction is used* to test the hypothesis of a zero mean. Note that u_1 is scaled by square root of 2 so that its length works out to be “1”.
- u_2 - is a vector of unit length that is perpendicular to u_1 . That means that their *dot product* is zero. u_2 and its multiples generate the error space. Note that u_2 is scaled by square root of 2 so that its length works out to be “1”.
- $y \cdot u_1$ - is the dot product of y with u_1 and is equal to the length of the projection of y onto u_1 . If this length is short that suggests that the true mean is zero (or doesn’t differ significantly from zero)
- $y \cdot u_1 * u_1$ - is the projection *vector* of the y component in the model space, whereas, $y \cdot u_1$ is it’s *signed length*.

- $y \cdot u_2$ - is the dot product of y with u_2 and is equal to the length of the projection of y onto u_2 . Its length provides a measure of the variability of the y vector
- $y \cdot u_2 \cdot u_2$ - is the projection *vector* of the y component in the model space, whereas, $y \cdot u_2$ is its *signed length*.
- The ratio of $y \cdot u_1 / |y \cdot u_2|$ is the t-statistic and its value is compared with standard tables to see if it is large enough to suggest rejecting the *null hypothesis* that the true mean of the deviations is zero.
- $y = (y \cdot u_1) \cdot u_1 + (y \cdot u_2) \cdot u_2$ and is a vector equation that expresses a right triangle (as a 2-dimensional figure) with y as the hypotenuse.
- $y \cdot y = (y \cdot u_1)^2 + (y \cdot u_2)^2$ - says that the square of the length of the y -vector equals the sum of squares of the other two vectors.



The F ratio will be used extensively in more advanced applications of ANOVA and Regression. Notice though that it is simply the square of the t-test.

Taking Three Readings at the Pump

Now I would like to extend our gas pump example by taking *three* readings at a time. This will introduce additional complexity but well within our capability to handle by vector means. See Figure 8 on page 12. Again, the idea is to see how far up the equiangular line the *cluster of out-*

comes is situated. Remember, if the true mean is μ , then the outcomes should cluster around the point in 3-space that looks like $\{\mu, \mu, \mu\}$. Is this μ (statistically) equal to zero? That is the question. In effect, we are going to project the observation vector 'y', down onto the equiangular line, a 1-dimensional subspace generated by 'u1', and use the tip of that projection as the estimate for $\{\mu, \mu, \mu\}$. To account for inherent variability of the observation vector, we use the average of the squares of the error space projections to estimate σ^2 . I'll discuss all this next.

Our model space is still 1-dimensional and is still spanned by the u1 equiangular basis vector but now, since we are in an overall 3-dimensional space, the remaining two dimensions making up the error space require two basis vectors to generate it. I will call those basis vectors u2 and u3. The details of u1, u2, and u3 are shown in Figure 8 on page 12, below. The plan is now to break up the observation vector into *three* components, with the projection component of y along u1 being an estimate of the population mean while the squared projection components of y along the other two basis vectors are each unbiased estimates of the population variance.

Characteristics of the Squared Projection Lengths

Note: It can be shown (see the site tutorial *Basic Statistics*) that the square of the projection length of y onto u1 is also an estimate of σ^2 *plus a factor involving μ^2* . That is, the square of the projection length of y onto u1 could be *inflated* beyond σ^2 . I will call this square 'c'. If it is not inflated, then it estimates σ^2 , just like the error squares do. If it is inflated then the ratio of c/d (see 'd' below) will exceed "1" by various amounts.

$$c = (y \cdot u1)^2$$

Second, the squares of the lengths of the projections along u2 and u3 each estimate σ^2 in an unbiased and independent way. The best estimate of σ^2 therefore suggests that we *pool* the two estimates by averaging them. That is, we add $(y \cdot u2)^2$ and $(y \cdot u3)^2$ and then divide by 2. I will simply call that 'd'.

$$d = ((y \cdot u2)^2 + (y \cdot u3)^2) / 2$$

That divisor of '2' expresses how many independent estimates of σ^2 that I have. And, you will also notice that it is the dimension of the *error* space. This is the basis of the term *degrees of freedom* as I described earlier in an extract. The degrees of freedom is simply the dimension of the error space which, in turn, equates to the number of independent estimates we can make of σ^2 .

The ratio of c/d is called the F-ratio and the square root of this ratio is the t-test.

If the mean truly is zero, the top and bottom of this ratio should both approximate σ^2 and so eventually average out to around "1". But, if the mean differs from zero then the numerator will be *inflated* and we get larger and larger numbers for the t-test and the F-ratio, leading to a possible rejection of the null hypothesis that the mean really is zero.

The observation vector 'y' can be written as a combination of the three basis vectors, u1, u2, and u3. u1 spans the 1-dimensional model space while u2 and u3 span the 2-dimensional error space. Notice that the basis vectors are mutually perpendicular. The projection of y down onto u1 is an estimate of the mean, μ , while the square of the projections onto u2 and u3 estimate the variance, σ^2

```

u1 = {1, 1, 1} / Sqrt[3]
u2 = {1, -1, 0} / Sqrt[2]
u3 = {1, 1, -2} / Sqrt[6]
lengthProjectionyOnu2 = y.u2
-0.141421
lengthProjectionyOnu3 = y.u3
-0.0816497

```

```

errorVector = y.u2 u2 + y.u3 u3
{-0.166667, 0.0333333, 0.133333}

```

```

|B|=averageLengthErrorVectors =
Sqrt[ ((y.u2) ^2 + (y.u3) ^2) / 2]
0.152753

```

```

|A|=lengthProjectionOfyOnu1 = y.u1
2.36714

```

$$y \cdot y = (y \cdot u1)^2 + (y \cdot u2)^2 + (y \cdot u3)^2$$

```

t = lengthProjectionOfyOnu1 /
avgerageLengthErrorVectors
20.5 t test value

```

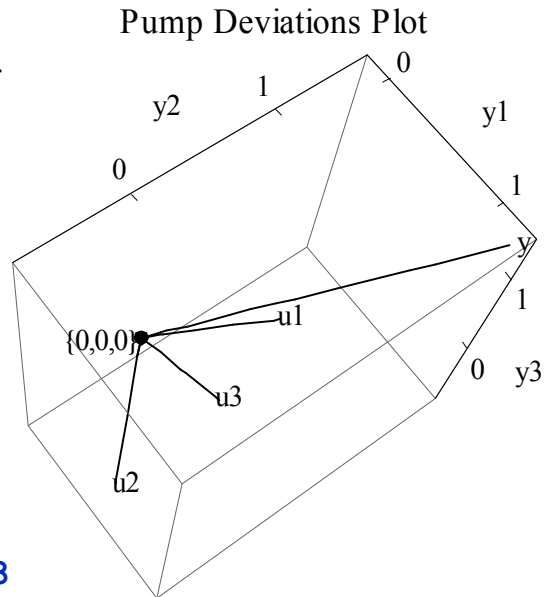
```

F= (lengthProjectionOfyOnu1) ^2 /
avgerageSquareLengthErrorVectors
240.143 F test value

```

FIGURE 8. Three Readings at the Pump

The Vector Space Structure



```

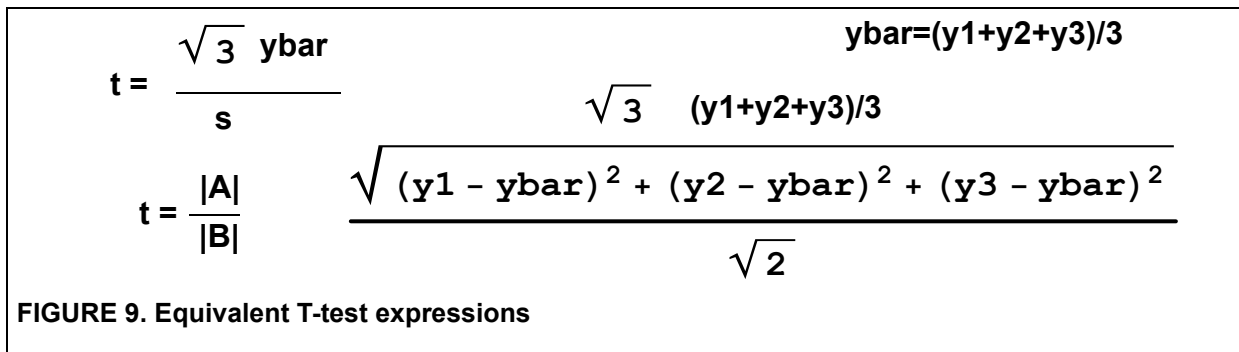
y = { 1.2, 1.4, 1.5}
ybar = Mean[y]
1.36667

```

$$\begin{aligned}
 y &= \left\{ \frac{6}{5}, \frac{7}{5}, \frac{3}{2} \right\} \\
 y \cdot u1 \ u1 & \left\{ \frac{41}{30}, \frac{41}{30}, \frac{41}{30} \right\} \\
 + y \cdot u2 \ u2 & \left\{ -\frac{1}{10}, \frac{1}{10}, 0 \right\} \\
 + y \cdot u3 \ u3 & \left\{ -\frac{1}{15}, -\frac{1}{15}, \frac{2}{15} \right\}
 \end{aligned}$$

Equivalent t-test Expressions

Compare these with the previous expressions of the t-test for two readings (See “Equivalent T-test expressions” on page 7).

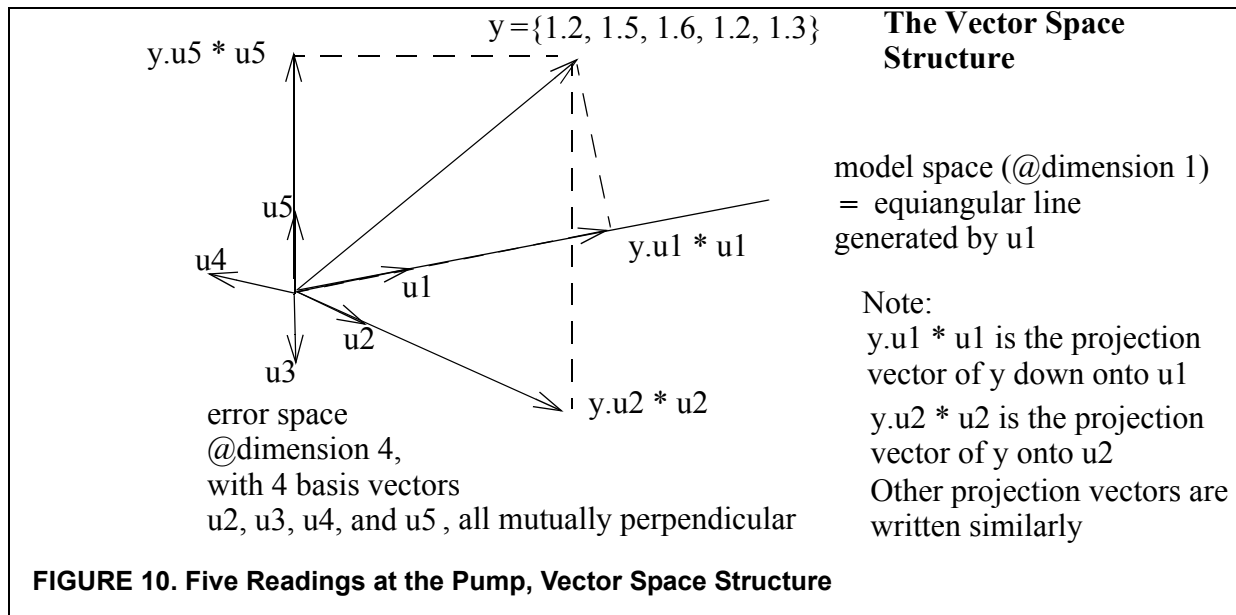


Summary

Since the critical t value for a 95% confidence interval is about 4.3, my value of 20.5 strongly suggests that the mean is NOT zero. Similarly, the F-Test says the same thing. For an F distribution[1,2], the critical point is 38.5. Since my F ratio turned out to be 240, this is pretty convincing evidence that the mean deviation is not zero.

Taking Five Readings at the Pump

Let me now introduce the last experiment, a five-reading protocol that will demonstrate how to think about the structure of 5-dimensional vector space in terms of our desired statistical tests. It’s a little hard to visualize 5-space but I can show in general how that works, using our vector approach.



The Test Protocol

This is a continuation of the previous two experiments. Now I'm going to take samples of size five. Each sample will consist of five readings of the deviations from the nominal reading of five gallons. The deviations are in terms of 10th of a gallon and I have multiplied those readings by 10 in order to make the math transparent.

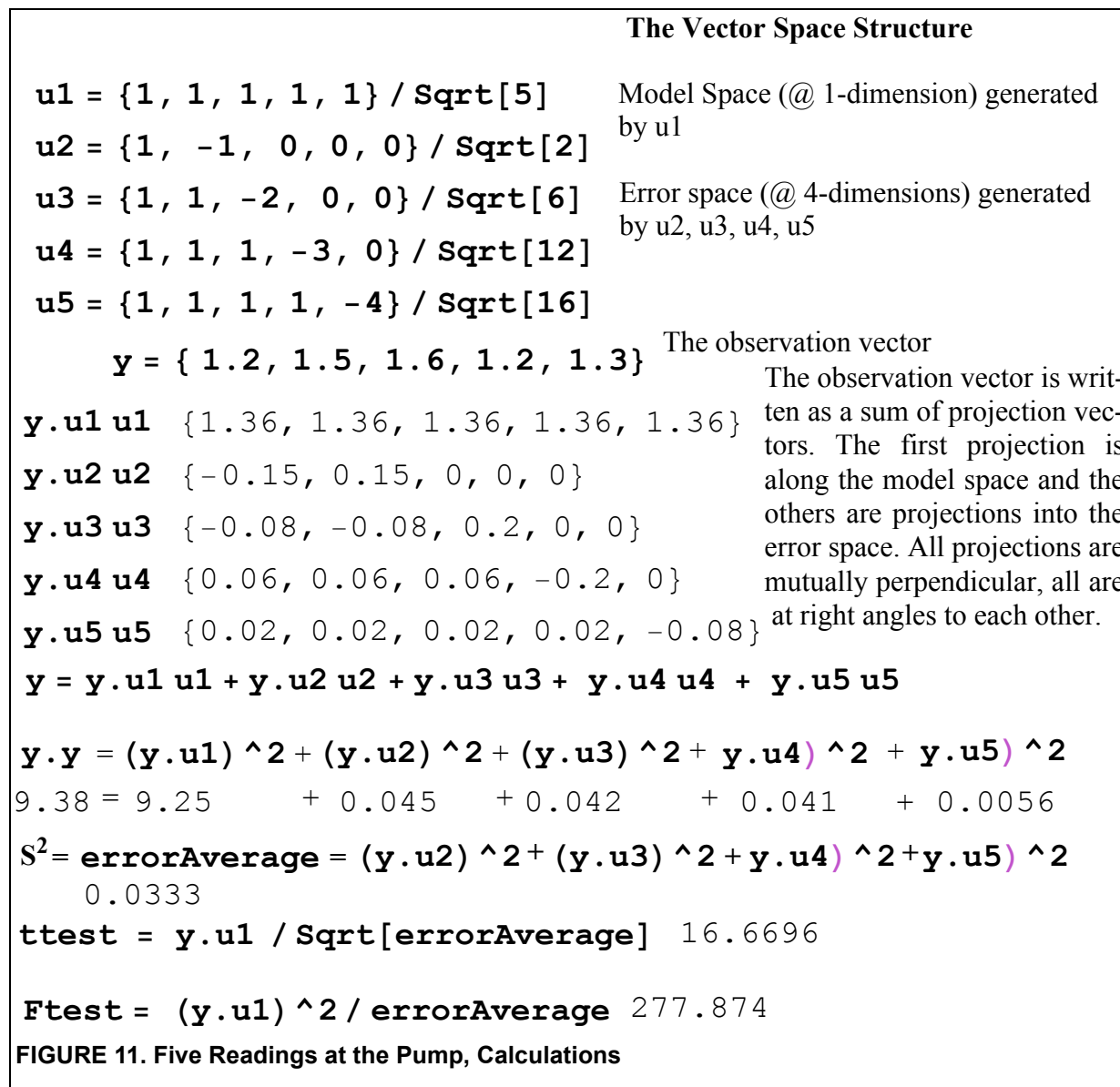
The Vector Space Structure

Since the observation vector has five components, the embedding vector space has five dimensions as well. Given five dimensions I could choose any five independent vectors to generate that space, however, my purpose is to test for a non-zero mean. That purpose determines my choices for these five vectors. As in the two previous experiments, I want one basis vector to point along the equi-angular line that represents the hypothesis direction of the *mean*, in the *model space*. That unit vector I will call 'u1' and it will look like:

$$u1 = \{1,1,1,1,1\}/\text{Sqrt}[5]$$

The $\text{Sqrt}[5]$ in the denominator scales the vector to be of unit length. That is $u1 \cdot u1 = 1$

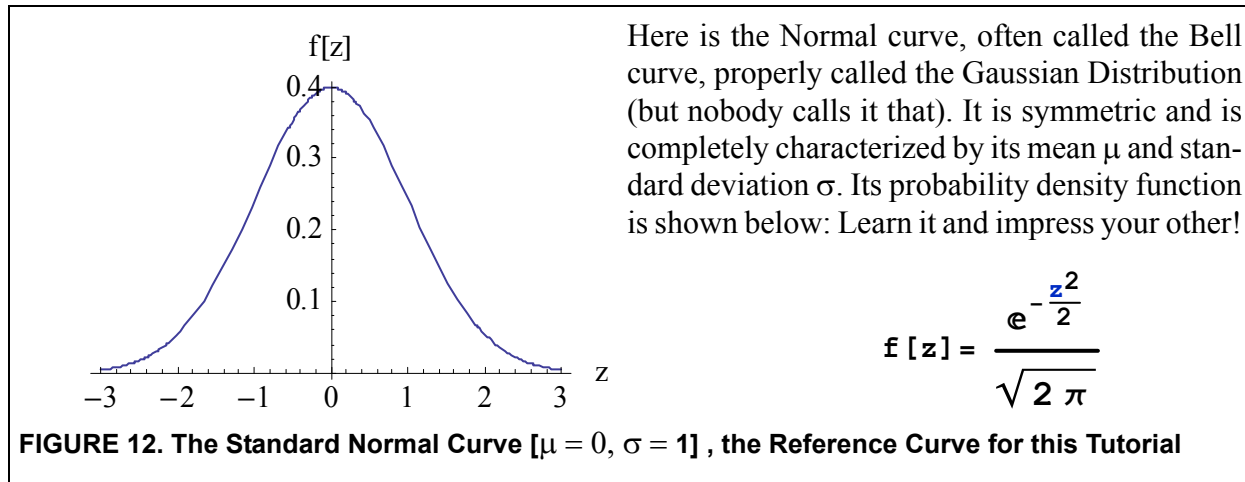
I then choose four other mutually perpendicular vectors to generate the remainder of the vector space. I will label these as u2, u3, u4, u5 and they generate the four dimensional *error space*. The set of five basis vectors that generate the five-dimensional vector space are noted below: "Five Readings at the Pump, Calculations" on page 15. This set of five mutually perpendicular basis vectors is called an orthonormal basis. Perpendicular simply means that every dot product between these vectors is zero. This also means that each is at right angles to the other.



The Normal Reference Distribution, the Foundation of Statistical Tests

Below is a graph of the Normal reference distribution having mean μ , $\mu = 0$ and standard deviation σ , $\sigma = 1$. The distribution actually extends to infinity in both directions but the bulk of the area, 99.73%, lies between -3 and +3. You will also see the function that generates that curve, called the PDF (probability density function), $f[z]$. Take a moment and check out the fact that the function is symmetric, when 'z' is zero it has a max height of $(1 / \text{Sqrt}[2 \text{ Pi}]) = 0.4$, falls off exponentially as you move away from zero, and the total area under the curve = 1.0. The area under the curve between any two 'z' values equates to the probability of finding a value from the data distribution between those two points. For example, the area between $z = -1$ and $z = +1$ is 0.683 which says that the probability of finding a number from this distribution between $z = -1$ and $z = +1$ is 0.683 or 68.3%. You can also think of these 'z' values along the horizontal axis as 'sigma' scores. That is, the area between ± 1 sigma is 0.683. That's the same as $\pm 1 z$. The area between ± 2 sigma is 0.9545 or 95+%.

To make this Normal curve relevant, think of a process, say making and testing Centigrade thermometers. At zero degrees, a tested thermometer ought to read “0”. If the process is ‘in control’, most of the thermometers will have readings clustered around zero. As we move away from zero, fewer and fewer will be off by that much. Finally, those off more than 3 degrees are definitely suspect! This bell shaped curve will often be a pretty good graph of such process errors.



Why Vectors and Vector Spaces?

The geometry I will use in the examples in this tutorial has a natural home in what is called a *vector space*. The geometry of a decision can be related to a vector space since every set of experimental observations can be represented as a *vector* embedded in the space. The embedding vector space has a dimension of the same size as the observation vector, plus a rich set of operations we can use to analyze, synthesize and manipulate various other vectors we can construct and interpret within this space. The space and its ‘fine’ structure will turn out to have statistically important interpretations.

Primarily, representing statistical decisions within vector spaces allows us to use our innate spatial competence to understand, implement, and test statistical concepts.

Note: If your understanding of vectors and vector operations is a little hazy, you might want to check out another tutorial on this site called *Vector Arithmetic and Vector Operations*. If you like, you can go back even further with the *Trigonometry Basics* tutorial, especially if you would like to refresh yourself on topics like pythagoras’ theorem, dot products, and trig operations. There is an irreducible amount of math involved in all my tutorials, but I urge you to learn with me as you come to experience the pleasure of understanding and implementing statistics and quantitative decision making at a deeper level.

A Little History

This geometric approach, pioneered by the agricultural statistician Sir Ronald Fisher in the 1920’s and 1930’s, allowed workers to *visualize* what was going on when they analyzed *multivariate* statistical questions. After Fisher’s time though, his approach fell into obscurity since people wanted quick, efficient, algebraic solutions that were easy to calculate, not requiring (perceived) complex geometry. Consequently, visual insight was replaced by the dreaded ‘cookbook formula’ approach. Even today, almost all texts still present the ‘cookbook’ approach that involves plugging and chugging with various formulas to give hypothesis tests and confidence intervals. These equations hide

mystery upon mystery that the book's authors attempt to explain verbally and algebraically, but only geometry can dispel.

Good news! In the 1990's, several authors have revived Fisher's approach and turned it into a workable, modern process, while still retaining his basic geometric insights. Of particular importance is the work of David Saville and Graham Wood, two statisticians from New Zealand and Australia. Their papers and textbooks have provided the inspiration for these notes (See references). I have taken their overall perspective and applied it to the needs of the students I work with, giving lots of graphical guidance. I ought also to mention the work of Box and Hunter [Statistics for Experimenters] where they also use geometry to explain some parts of statistics.

Statistical Decision Making Tests

From these notes you will soon come to appreciate that deciding on the interpretation of an experiment uses one or both of two very closely related tests, called the *t-test* and the *F-test*. These two tests are used throughout statistical decision making so, having familiarity with them is essential for your further progress. Along with setting up decision examples, the approach I will introduce here is not just focused on statistical tests, but on the accompanying ideas and methods as well. These ideas and insights will serve you well when you work on Analysis of Variance (ANOVA), and Regression (Trend Line) questions. I will introduce ANOVA ideas in a follow-on Tutorial.

For those of us who need a good picture or diagram of what is going on during a statistical analysis and synthesis (or any other analysis/synthesis for that matter), the geometry presented in these notes will hopefully meet this desire. So, let me start with an example.

Mini - Glossary

For the following discussion I will explain and use several math ideas from vector spaces. If you get lost, come back to here and review these definitions: (you might want to check out another tutorial on this site called *Vectors Spaces and Vector Operations*).

- Vector: a directed line segment, also an ordered list of components. For example $v = \{1,2,3\}$
- Vector Space: A collection of vectors, together with operations you can perform on them, generated by a basis set of vectors. For example 3-dimensional space can be generated by the basis set: $x=\{1,0,0\}$, $y=\{0,1,0\}$, $z=\{0,0,1\}$. Any 3 dimensional vector can be written in terms of x , y , and z . So the vector $v = \{1,2,3\} = 1*x + 2*y + 3*z$. In a vector space you can multiply vectors by a scalar as well as add and subtract them.
- Model Space: a subspace of a vector space consisting of all possible combinations of the population *parameters*. For example if an observation vector $v = \{1,2,3\}$ has each component coming from the same Normal distribution with (unknown) mean μ , then the model space is the sub-space generated by $\mu * \{1,1,1\}$. The equiangular line in 3-space.
- Error Space: the vector space complement of the Model Space. That is, for the Model Space generated by $\mu * \{1,1,1\}$, the Error Space would be 2-dimensional and consist of a plane perpendicular to the Model Space vector $\{1,1,1\}$. The two basis vector $u_2 = \{1,-1,0\}$ and $u_3 = \{1, 1, -2\}$ could generate this error space plane. Notice that these three vectors are mutually perpendicular. (Take the Dot product of each with the others to see this). Taken together, the Model Space and the Error Space generate the whole vector space.
- Observation vector: This is the starting point for the statistical analyses. From given observed values which make up your sample, you can construct a vector. If the vector has, say, 3 components then you are automatically talking about a 3-dimensional vector space. If all three com-

ponents come from the same Normal distribution and you are testing for a zero mean, then the model space will be 1-dimensional while the error space will be 2-dimensional as in the bullet points discussion of Model Space and Error Space above.

- ****Projection Vectors (informal explanation):** A common physical example of a projection vector is when a sun-dial (the observation vector) casts a shadow (the *projection vector*) down onto the table space. To push this analogy a little further, the table space is the Model Space (where all the shadows are realized) and the Space perpendicular to the table space (that is, the vertical dimension) is the Error Space. There is a complementary projection of the sundial vector onto that Error Space as well. It would be the vector from the tip of the shadow up to the tip of the sundial vector.
- **Projection Vectors (formal discussion):** Formally, a projection vector (for our purposes) is the vector that results from taking the Dot product of the observation vector down onto one of the basis vectors (which is a scalar) times that basis vector. (This assumes that each basis vector is of unit length).

Summary

With more sample readings as in this 5 sample experiment, the precision usually goes up, In this case, consulting the standard t tables for degrees of freedom = 4 and at 95% confidence, I calculate 2.78. So my results strongly reject the null hypothesis and I conclude that I have good reason to believe that the pump readings are biased above zero. So when a motorist see '5' gallons on the pump, they are not getting 5 gallons in their tank. The standardized F-test value for 1 degree of freedom in the numerator and 4 in the denominator shows: 12.22. Wow, am I ever over that value! So, for sure I can't keep saying the pump readings are unbiased with a mean deviation of zero.

The follow-on tutorial Geometric Statistics - Part II, will test whether the means of two distributions differ. After testing the difference of means for two populations, I'll conclude with a test between the means of three or more different populations, that will introduce ANOVA (Analysis of Variance) procedures.

References

- Box, G. and Hunter, X, (1995) *Statistics for Experimenters*, Prentice Hall.
- Saville, D. and Graham Wood (1991) *Geometric Statistics*, Springer-Verlag
- Saville, D. and Graham Wood(1996) *Statistical Methods A Geometric Primer*, Springer-Verlag
- Rucker, R.(2008) *Vector Spaces and Vector Operations*, milagrosoft.com
- Rucker, R.(2008) *Trigonometry Basics*, milagrosoft.com
- Rucker,R. (2008) *Basic Statistics*, milagrosoft.com