Simple t - Tests as Reference Examples [*draft 2009-02-19]

*Note : The following examples are so simple that they could be used as a check when you are testing out your own calcula-tions or those of a software package. Just use the numbers in the examples below and see how they correlate to your own or your packages' output.

Reference example #1 : Check out a t-test for a hypothesis of a zero population mean $\mu = 0$ versus $\mu \neq 0$.

Reference example #2 : Check our a t-test for a hypothesis that the means of two populations are equal

(In fact, I have used the *Mathematica* package to confirm my geometric analyses. So, when you see methods or functions starting with a capital letter, you will know these are from the *Mathematica* package. For example, I use the built in **Mean** method to calculate averages.

Reference Example #1: Is the Population Mean Zero, $\mu = 0$?

Consider testing to see if a sample data set of 4 numbers *could likely* have come from a population with a true mean of zero. (*Note: this is more general than it seems since you can *subtract off* a fixed number from the observations and test *that* distribution by exactlly the same methods shown below). We assume the population of numbers is Normal with a common variance, and I am drawing independent samples from the same population. Both the true population mean and the true variance are unknown however, and will have to be estimated from the sample data. This is an example of a hypothesis test that asks if the population mean, μ, is likely to be zero : H0: $\mu = 0$ (the *null* hypothesis) while the *alternative* hypothesis says the true mean is unlikely to be zero H1: $\mu \neq 0$. For small samples (or large ones for that matter), a calculated value, called the 't-statistic', can be used to distinguish between the two hypotheses. Consider the data set:

y = { 1., 2., 4., 3.}

 Could those four numbers have come from a population whose true mean is zero? On the face of it, since the grand average is 2.5, a zero true mean seeems unlikely. Unfortunately, variation in the data set will often mask the actual values of population parameters like the mean. Calculating a *t-statistic* value is one way of determining this issue. You will see that the t-statistic is simply the *ratio of lengths of vectors.* That ratio is the length of a *grandAverage* vector divided by the square root of the average squared lengths of the *error* vectors. Using this ratio we wil accept or reject the null hypothesis that the true popula-tion mean is zero. All this is developed below.

 A word on terms: the *model* space is the set of vectors that represent *all* the possible true means of the population underlying the given data set. These potential vectors will all lie along an equiangular line. This viewpoint will become clear as we look at some diagrams below. In the case of testing for a mean of zero, the model space will always be 1-dimensional The remain-ing dimensions belong to the *error* space. That is, if I have 4 observations the model space dimension is 1 while the error space dimension is 3. For a sample of 50 observations, the model space is 1-dimensional while the error space dimension would be 49. The dimension of the error space is called *degrees of freedom* (df) in the statistics world.

```
y = {1., 2., 4., 3.}; (* the sample data set, the observed values *)
grandAvg = Mean[y] (*average of all the sample  values*)

2.5
```

■ **Summary Calculations for Use in Checking a t-test for a Zero Population Mean**

 I'll just list the outcomes here as a way to check your own work or the output of a statistics package. All of these calculations are worked out below but are summarized here.

grandAverage : 2.5 (this is the best estimate of the population mean μ, using the sample data)

sample variance , $s^2 = 1.66667$ (this is the best estimate of the population variance σ^2)

sample standard deviation = 1.291 (this is the square root of the sample variance

t-statistic (calculated t value) = 3.873 (this is used to determine whether to accept or rejection of the null hypothesis)

Two sided p-value = 0.03 (this is the chance of getting a t-statistic larger than 3.873 or smaller than -3.873)

lower and upper critical values at 95% confidence (@ 3 df) = -3.182 , 3.182 (the area outside of these limits is 0.05, these two critical values come from standardized tables)

confidence interval for the true mean at 95% confidence = $0.44574 \le \mu \le 4.55426$

Notice that the t - statistic , 3.873, is outside the critical values which means it is outside of where 95% of values would fall if indeed the zero mean hypothesis were true. Another signal of rejection is that the confidence interval doesn't include zero! That is, a value of *zero* for the true mean is highly unlikely.

■ Exploring the Geometry of the T - Test for a Population Mean of Zero

To go a little farther in explaining the bases of the calculations of the t - test, I will discuss the process in its natural setting, a vector space. (Readers wishing a little more background on vectors, vector spaces, bases, and the dot product, may want to check out another tutorial by typing this address in your browser:

http://milagrosoft.com/VectorOperationsQuickLook.pdf

If you are familiar with those ideas, read on!

■ Introducing Vectors and Vector Spaces

Let me use vectors to represent the observations as well as describe the vector space these observations are embedded within. (Vector spaces are the natural home of statistics and provide considerable insight during exploration and analysis. In particular, geometric operations can be used to derive the usual algebraic formulas found in text books). Recall the observation vector **y** = {1., 2., 4., 3.}

Since there are 4 components to this vector, that requires a *4 -dimensional vector space* for its representation. To describe that whole space itself, I use 4 special basis vectors that let me write any vector with 4 components uniquely in terms of them. In particular, I can write **y** in terms of 4 mutually perpendicular *basis* vectors having special directions. The 4 basis vectors I have chosen that generate the 4-dimensional vector space are: The **u1** unit vector pointing along the equiangular line that contains all possible true mean values while the three remaining vectors span (cover) the error space. Three more vectors, **u2**,**u3**, and **u4**, are required to generate the remaining 3-dimensional error space.

```
u1 = {1, 1, 1, 1} / Sqrt[4];(*points along the grand avg*)
u2 = {1, -1, 0, 0} / Sqrt[2];(*error vector *)
u3 = {1, 1, -2, 0} / Sqrt[6];(*error vector*)
u4 = {1, 1, 1, -3} / Sqrt[12];(*error vector *)
```

■ Projecting y Along Each Basis Vector Gives a Unique Decomposition of 'y'

I can now write the observation vector in a unique way using these 4 basis vectors. Projecting **y** along the **u1** vector yields the vector, **grandAvgVec** whose each component is the grand average. Projecting **y** along **u2**, **u3**, and **u4** represent the error components of the observation vector.

Projecting **y** along the **u1** basis vector is the best estimate of the population mean. (It' s all we' ve got!). Notice that the **u1** vector points along an equiangular line that encompasses all possible true mean values. So the best estimate of the population mean is 2.5.

```
grandAvgVec = y.u1 u1
```

$\{2.5, 2.5, 2.5, 2.5\}$

The other three basis vectors cover the error space and will be used to estimate the variance of the data at a later stage of analysis. Right now, I will project **y** onto each of these vectors. Look at the diagram below to see where I am going with this. From the calculations below you can see that if you add up the first four projection vectors you will get the observation vecto**r** **y**. So, I have written **y** as a unique sum of four vectors that I can interpret in terms of the hypotheses. That is, I can reduce the hypothesis test question to a matter of calculating a ration of lengths. So, if the length of the **grandAvgVec** is *long* relative to the average of the error vector lengths, this suggests that the true mean differs from zero otherwise, there is not sufficient reason to reject the hypothesis that the true mean is zero.

grandAvgVec = y.u1 u1;	$\{2.5,$	$2.5,$	$2.5,$	$2.5\}$
error2Vec = y.u2 u2;	$\{-.5,$	$.5,$	$0,$	$0\}$
error3Vec = y.u3 u3;	$\{-0.833,$	$-0.833,$	$1.67,$	$0\}$
error4Vec = y.u4 u4 ;	$\{-0.167,$	$-0.167,$	$-0.167,$	$0.5\}$
y =	$\{1,$	$2,$	$4,$	$3\}$

At this point **y** is decomposed into 4 mutually perpendicular vectors.

y = grandMeanVec + error2Vec + error3Vec + error4Vec

A 4-dimensional vector space spanned by u1, u2, u3, u4
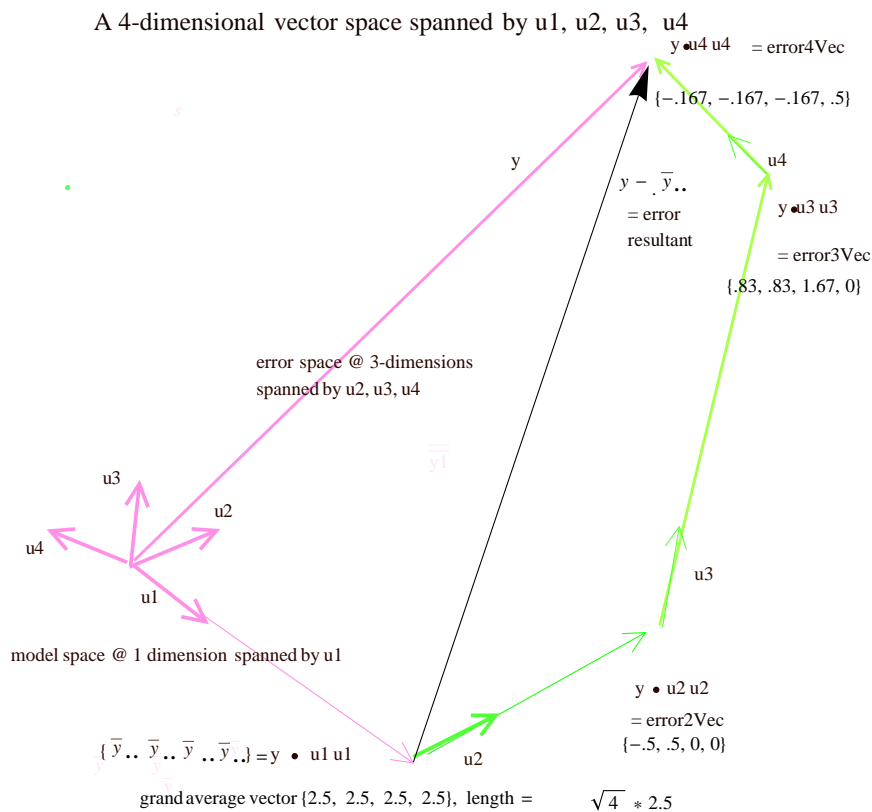
■ **Using Geometry to Calculate the t-Statistic**

s2 = sample variance is the average of the squared lengths of the three error vectors. This is the Mean Square Error 'MSE' column in computer package outputs. Notice that the averaging is done over the *three* error vectors squared lengths. This is where the term degrees of freedom (df) comes from and here we see that *df = 3*. This average is called the *pooled variance estimate* since it is the average of three independent estimates of the population variance.

```
s2 = ( (y.u2)^2 + (y.u3)^2 + (y.u4)^2) / 3

1.66667
```

sample standard deviation =

```
s = Sqrt[s2]

1.29099

lengthOfGrandAvgVector = y.u1

5.

tstatistic = lengthOfGrandAvgVector / s

3.87298
```

This last number, the t - statistic, is a ratio of lengths we can use to test the hypothesis.

■ **Calculating the Chance of Getting that t - statistic Value**

Using the Mathematica package built-in table generators for the T - Distribution, I can find the chance that the value 3.873 or larger, would occur under the assumption that the true population mean was actually zero. I first find the probability density function (PDF) and then use that to find the area of the t-distribution to the right of the t-statistic 3.873. This involves integrating from -Infinity up to 3.873.

The result, calculated below, is that the chance of getting such an outcome under the null hypothesis is about 1.5 %. So, there is good reason to reject the null hypothesis and we conclude that the sample values suggest a true mean different from zero.

```
probabilityDensityFunction = PDF[StudentTDistribution[3], x];

1 - Integrate[probabilityDensityFunction, {x, -Infinity, tstatistic}]

0.0152331
```

■ **Critical Values of T - Distribution for 3 Degrees of Freedom**

Find the t value such that 0.975 of the area of the T-distributin lies to its left. Note that the negative of this number will be the critical value such that 0.025 lies to *its* left. You can find these numbers in standard 't-table' but I can use *Mathematica* to calculate them as well.

```
tcrit = Quantile[StudentTDistribution[3], 0.975]

3.18245
```

■ **Finding a 95 % Confidence Interval**

Subtracting off the true mean from an estimate of the population mean
gives me a random variable whose mean is zero and whose variance is σ^2 That is, (2.0 grandAvg - 2.0 μ) is a N [0, σ^2]
random variable.
Dividing this by the sample standard deviation, s, gives me a variable that follows the t-distribution.
If this varible does follow the t-distribution, then its 95% confidence interval is bounded by those two
critical values found earlier, that is, -3.18245 and + 3.18245. Given these conditions, I can isolate the μ variable in a couple
of steps. (I have reproduced the code I used in *Mathematica* but of course the final interval is the interesting outcome rather
than the intermediate steps. But just in case you are interested)!

```
confi = -tcrit ≤ (2.0 grandAvg - 2.0 μ) / s && (2.0 grandAvg - 2.0 μ) / s ≤ tcrit;
```

$-3.18245 \le 0.774597 (5. - 2. \mu) \&\& 0.774597 (5. - 2. \mu) \le 3.18245$

```
Reduce[confi, b]
```

$0.44574 \le \mu \le 4.55426$

■ **The Standard Algebra of the t - Statistic**

Let me translate the geometry into the standard algebraic formulas that appear in text books. The general formula is
$t = \overline{y}/(s/\sqrt{n}$)where $\overline{y}$ is the grand mean, **n** is the sample size and **s** is the sample standard deviation. If I call each observation
y[i], and let i go from 1 to n, I can write the general formulas as:

$$\sqrt{n}$$

$$\overline{y} = \left(\sum_{i=1}^{i=n} y[i] \right) \Big/ n$$

$$s2 = \left(\sum_{i=1}^{i=n} (y[i] - \overline{y})^2 \right) \Big/ (n-1)$$

$$s = \sqrt{s2}$$

Plugging in numbers for the particular example we started with, we get

$$\overline{y} = 2.5, \quad \sqrt{n} = 2, \quad s = 1.291$$

$$t = 5.0 / (1.291)$$

3.87297

Reference Example #2: Are Two Population Means Equal, $\mu 1 = \mu 2$?

*Note : This example can be used to check the output of a two population t - test. That is, you can use this example just to
check and then interpret what a statistics package prints out when you ask for the two-sample t-Test (Students T Test)

This example takes the same data set above but now pretends that the 4 numbers come from two different populations, at 2 each. Are the means of those two populations equal (statistically)?

First off, assume population I consists of just the two sample readings {1, 2}, while population II consists of the readings {4, 3}. The sample average of the readings of population I is 1.5 and the sample average of the readings of population II is 3.5. The difference of the averages is 3.5 - 1.5 = 2. Does that mean that the actual underlying *population* means are equal? If there were no variability in the values of the data sets that would be true but, variablity can mask true differences. So lets see what the t-test statistic can tell us.

```
p1 = { 1., 2.};(*population I sample data*)
p1avg = Mean[p1](* find the sample average for population 1 values*)

1.5

p2 = { 4., 3.};(*population II sample data *)
p2avg = Mean [p2](*find the sample average of population 2 values*)

3.5

y = Join[p1, p2](* the 4-dimensional observed data vector*)

{1., 2., 4., 3.}

grandAvg = Mean[yobs](*overall (grand) average of observed sample data*)

Mean[yobs]
```

▪ Cut to the Chase Section

If you don' t have time to go through the derivation of this t-statistic, you can just read the summary below and check out your stat package outputs.

- **Null Hypothesis: are the true means (statistically) equal, that is H0: $\mu1 = \mu2$?**
  The alternative hypothesis is that they are (statistically) unequal. H1: $\mu1 \neq \mu2$

best estimate for population I's true mean $\mu1$ = sample average = 1.5

best estimate for population II's true mean $\mu2$ = sample average = 3.5

Best estimate for difference of the true means is 3.5 - 1.5 = 2.0 (* note

that this will not turn out to be statistically significant due to inherent variability)

DF error degrees of freedom = 2

pooledSampleVariance = 0.5

sampleStandardDeviation = Sqrt[0.5] = 0.707107

Squared length of the treatment vector (difference of means) = 4

Squared length of the overall error vector= 1

Average squared length of error vectors = 1/2 = 0.5

F-Ratio value = Squared length of treatment vector/Average squared length of error vector = 4/0.5=8

A calculated value that tests whether there is a detectable difference between true means based on 2 degrees of freedom

t statistic = 2.828

For a 2-sided test, the critical t-values at 95% , for 2 degrees of freedom (this is the dimension of the error space)

t.025 = -4.302

t.975 = 4.302

So, if I get a t value outside of this range I would reject the null hypothesis of equal means. Since my value, 2.828 is within this range, I don't have sufficient reason to reject the null hyp. We would say that the t statistic value is not significant. Apparently, the variability of the data has masked any true differences we might detect.

The confidence interval is the interval that is consistent with this data set and represents where the differences should fall, for 95% of identical experiments.

Confidence interval for $\mu1 - \mu2$

Noitice that the confidence is another signal that the t - statistic is not significant since zero is included in the confidence interval as a possible value. That is, the null hypothsis is supported by this confidence interval

- **Using the Mathematica Package to Confirm the Geometric Analysis**

The data entered below is the syntax required for Mathematica. The first component of each vector is the population number while the second value is an observation. For example, the first vector {1, 1} means population I with observed value of ' 1'. The third entry of the list, {2, 4} refers to population II with observation 4. Since the overall mean of the combined popula - tions is not usually of interest, packages omit that output and simply show the difference of means data, called the *model*. You can see that the SumOfSq is the squared length of the y **. u2 u2** treatment vector.

```
Needs["ANOVA`"]
Needs["HypothesisTesting`"]

ANOVA[{{1, 1}, {1, 2}, {2, 4}, {2, 3}}]
```

$\left\{ \text{ANOVA} \rightarrow \begin{array}{llllll} & \text{DF} & \text{SumOfSq} & \text{MeanSq} & \text{FRatio} & \text{PValue} \\ \text{Model} & 1 & 4. & 4. & 8. & 0.105573 \\ \text{Error} & 2 & 1. & 0.5 & & \\ \text{Total} & 3 & 5. & & & \end{array}, \text{CellMeans} \rightarrow \begin{array}{ll} \text{All} & 2.5 \\ \text{Model}[1] & 1.5 \\ \text{Model}[2] & 3.5 \end{array} \right\}$

■ **Embedding the Statistics Question Within A Vector Space**

Let me describe the Vector Space that is the natural home of the Observed vector **y**. It is 4 - dimensional since y has 4 components. This space that **y** inhabits can be generated by 4 mutually orthogonal unit vectors, called **basis** vectors: u1, u2, u3, and u4. The u1 vector points in the direction of the overall grand average, the u2 vector points in the direction of the difference of the population averages, whose hypothesis we want to check, while u3 and u4 generate the 'error' space.

```
u1 = {1, 1, 1, 1} / Sqrt[4]; (*special vector in the model space,
thath points in the direction of the  overall average*)

u2 = {-1, -1, 1, 1} / Sqrt[4]; (*special vector in the  model space
 that points in the direction of the difference of sample averages*)

u3 = { 1, -1, 0, 0} / Sqrt[2]; (* error space vector,
related to variability within population I*)

u4 = { 0, 0, 1, -1} / Sqrt[2]; (* error space vector,
related to variability within population II*)
```

■ **Now I can write any vector in this 4 dimensional space by using the 4 basis vectors, u1, u2, u3, u4. In particular, I can write the observation vector 'y' as a sum of 4 vectors, the projections of y along each of the u1, u2, u3, and u4 vectors.**

```
y = {1., 2., 4., 3. }; (* our observation vector *)

grandAvgVec = y.u1 u1 (* the projection vector of y along u1*)

{2.5, 2.5, 2.5, 2.5}

a = y.u1 (* length of the grand average projection vector*)

5.

differences = y.u2 u2 (* the projection vector of y along u2*)

{-1., -1., 1., 1.}

b = y . u2 (*(signed) length of the differences projection vector *)

2.

error1 = y.u3 u3 (* the projection vector of y along u3*)

{-0.5, 0.5, 0, 0}

c = y.u3 (* length of this error vector *)

-0.707107

error2 = y.u4 u4 (*the projection vector of y along u4*)

{0, 0, 0.5, -0.5}

d = y.u4 (* length of this error vector *)

0.707107
```

- **Check that y · y = overall · overall + differences · differences
  + error1 · error1 + error2 · error2**
    We will use the length b as a measure of the differences in means of the two populations and the square root of the average of the
    squares of c and d as error estimates (this square root is the sample standard deviation 's') . The ratio of b/s is the t-statistic.

A 4-dimensional vector space spanned by u1, u2, u3, u4
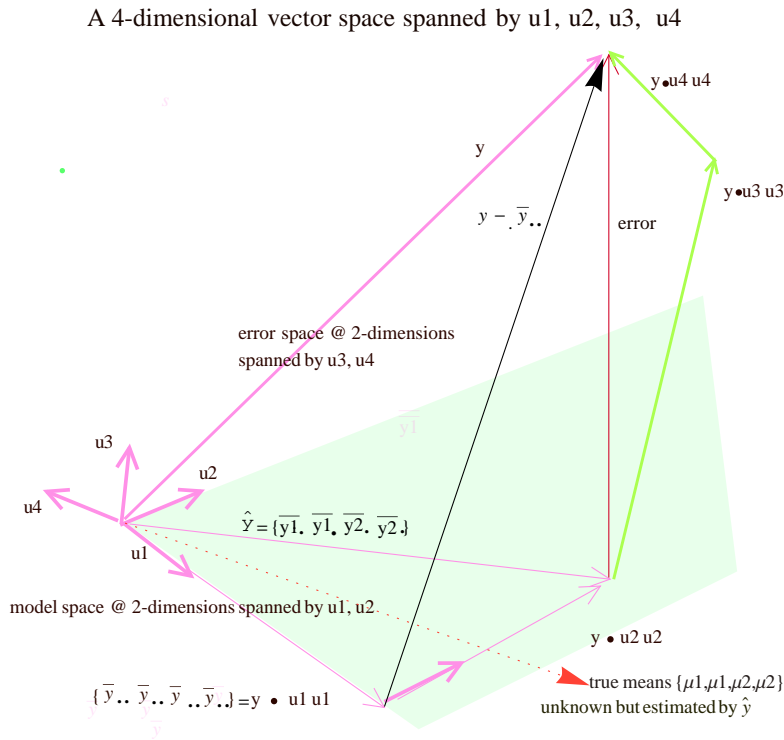
- **T - Test Analysis**

```
b (* length  of
```

```
2.
```

```
c^2
```

```
0.5
```

```
d^2
```

```
0.5
```

```
pooledSampleVariance = (c^2 + d^2) / 2
```

```
0.5
```

```
sampleStandardDeviation = Sqrt[pooledSampleVariance]
```

```
0.707107
```

```
t = b / sampleStandardDeviation
```

```
2.82843
```

Checking the values of a t - distribution with 2 degrees of freedom

Since I have a math package I can do an independent check for any parameters I like

```
tdist = StudentTDistribution[2];

tpdf = PDF[tdist, x](*probability density function*)
```

$$\left(\frac{1}{2 + x^2}\right)^{3/2}$$

```
t975 = Quantile[tdist, 0.975](* what t value such that 0.975 of the area lies to its left?*)

4.30265
```

- **Checking that value by actual integration**

```
Integrate[tpdf, {x, -Infinity, t975}]

0.975

s = sampleStandardDeviation

0.707107

delta = (μ1 - μ2)

confi = (-t975 * s ≤ (p1avg - p2avg) - delta) &&
  ((p1avg - p2avg) - delta ≤ t975 * s)
```

$-3.04243 \le -2. - \text{delta} \&\& -2. - \text{delta} \le 3.04243$

```
Reduce[ confi, delta]
```

Reduce::ratnz :

  Reduce was unable to solve the system with inexact coefficients. The answer was obtained by solving a
      corresponding exact system and numericizing the result. ≫

$-5.04243 \le \text{delta} \le 1.04243$

So, this means that $(\mu1 - \mu2)$ lies in that interval.

- **Verifying this result using the Mathematica package (it defaults to a 95% confidence interval)**

```
MeanDifferenceCI[{1, 2}, {4, 3}]
```

{-5.04243, 1.04243}

- **Extra Verbiage**

Compare the populations of returns from two energy conservation plans, A versus B. Plan A yields savings (in $10,000) of { y11, y12 } = { 1.0, 2.0} while plan B yields savings of {y21, y22} = { 4, 3} . These observed outcomes can be combined and written as a 4 component vector: y = { y11, y12, y21, y22} = { 1,2,4,3}, called the observation vector. The true, but unknown, underlying means of the two populations are $\mu 1$ and $\mu 2$. The null hypothesis is that they are (statistically) equal, H0: $\mu 1$ = $\mu 2$. This can be written in a more useful way as: H0 = ($\mu 1$ - $\mu 2$)=0 . There are 4 vectors that generate this 4-dimensional space.

The overall mean vectors has direction u1 = { 1,1,1,1}/Sqrt[4]

The unit direction associated with the null hypothesis is: u2= {1,1, -1, -1}/Sqrt[4]

The two error vectors are u3 = { 1,-1, 0,0}/Sqrt[2] and u4={0,0,1, -1}/Sqrt[2]

A 4-dimensional vector space spanned by u1, u2, u3, u4