

DISTRIBUTIONS USED IN STATISTICAL WORK

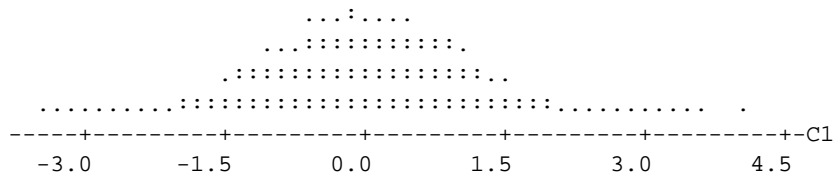
In one of the classic introductory statistics books used in Education and Psychology (Glass and Stanley, 1970, Prentice-Hall) there was an excellent chapter on different distributions commonly used in statistical work (normal, chi square, F, and t). This chapter developed each and showed interrelationships amongst these distributions. I have not really found in newer texts, any discussions on a par with this one. What typically happens is that WHEN a test is used, say the first time a t test is introduced, THEN that theoretical distribution is discussed. Generally, more recent books don't show or discuss connections amongst distributions; they are all isolated bits of information. Giving full credit to Glass and Stanley, I want to use Minitab in this handout to do some simulations that supplement what their chapter did. NOTE: There clearly are other distributions of value to present but, I am limiting myself to what was presented in the Glass and Stanley book. See the end for a summary of the distributions and some example uses of each.

NORMAL DISTRIBUTION

Think about a random variable where the sample space makes up a normal distribution where the mean = 0 and the standard deviation = 1. This is your standard unit normal distribution. Now, consider taking repetitive random observations from that distribution and then, forming a picture of that. You would see in the long run, this most familiar shape.

```
MTB > rand 10000 c1
MTB > dotp c1
```

Dotplot: C1



```
MTB > desc c1
```

Descriptive Statistics: C1

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	10000	0.0046	0.0038	0.0033	1.0042	0.0100

Variable	Minimum	Maximum	Q1	Q3
C1	-3.4060	4.0883	-0.6571	0.6803

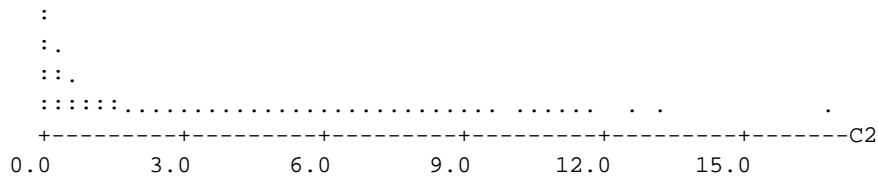
Consider this the BASE distribution.

CHI-SQUARE DISTRIBUTIONS

Now, what if we take a large set of randomly obtained values from the unit normal distribution and, SQUARED each of these single values? What we would have in that case would be a random variable that is made up of SQUARED z scores. We could do this easily in Minitab and, data would look something like:

```
MTB > let c2=c1**2 <<<< NOTE: I just squared the 10000 z values from the above
MTB > dotp c2
```

Dotplot: C2



```
MTB > desc c2
```

Descriptive Statistics: C2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C2	10000	1.0083	0.4501	0.8070	1.4225	0.0142

Variable	Minimum	Maximum	Q1	Q3
C2	0.0000	16.7139	0.1033	1.3371

What we clearly see here when we square the values in a random normal variate, is a radically + skewed distribution with a mean close to 1.

This is called a chi square distribution with one (1) degree of freedom. 1 degree of freedom in this case simply means squaring ONE z value from a normal variate and then making a plot of the resulting squared (single) z scores.

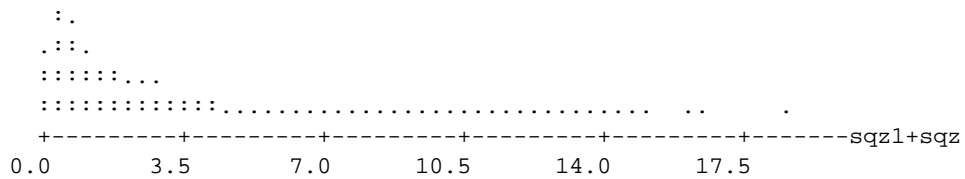
Now, what if instead of taking ONE set of normal deviate values and squaring them, we take TWO separate and independently drawn, normal deviate variables. That means we would have TWO columns of z values. Square each of these z values and then ADD THE 2 SQUARED z values together. I will show you some of these results and then let you see the distribution that is generated from this process.

Row	z1	sqz1	z2	sqz2	sqz1+sqz2
1	-1.07693	1.1598	-0.05103	0.0026	1.1624
2	-0.28851	0.0832	1.31326	1.7247	1.8079
3	-0.36138	0.1306	-0.37990	0.1443	0.2749
4	0.25661	0.0658	1.60017	2.5605	2.6264
5	-1.07087	1.1468	0.69643	0.4850	1.6318
6	2.65793	7.0646	-0.44119	0.1946	7.2592
7	-0.51160	0.2617	0.96389	0.9291	1.1908

8	2.14181	4.5873	0.76252	0.5814	5.1688
9	1.36193	1.8549	0.19187	0.0368	1.8917
10	-0.44002	0.1936	-0.79115	0.6259	0.8195

MTB > dotp c5

Dotplot: sqz1+sqz2



MTB > desc c5

Descriptive Statistics: sqz1+sqz2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
sqz1+sqz	10000	2.0219	1.3830	1.7915	2.0300	0.0203

Variable	Minimum	Maximum	Q1	Q3
sqz1+sqz	0.0003	18.7082	0.5687	2.8166

If you compare the two skewed distributions, you will see that in this second case, there is NOT as much skew; ie, the peak has moved a bit to the right. Note that when we add together TWO squared z values and look at the distribution, the mean now is approximately 2, and note that the standard deviation has increased too.

Call this a chi square distribution with two (2) degrees of freedom.

Now imagine taking more than 2 independent random normal variates, squaring and adding those together, then making a distribution out of the results. Say we took 5. Here is what you would find.

z1	z2	z3	z4	z5
-0.91347	2.05764	1.70838	-1.64319	0.62376
1.52128	2.59559	-1.05094	0.16472	-1.53489
0.87871	0.44068	1.10548	-0.24896	-0.08487
-0.51881	-0.52567	0.26883	-0.12788	-0.95803

-0.95218 0.83189 -0.36857 0.39893 -1.83481

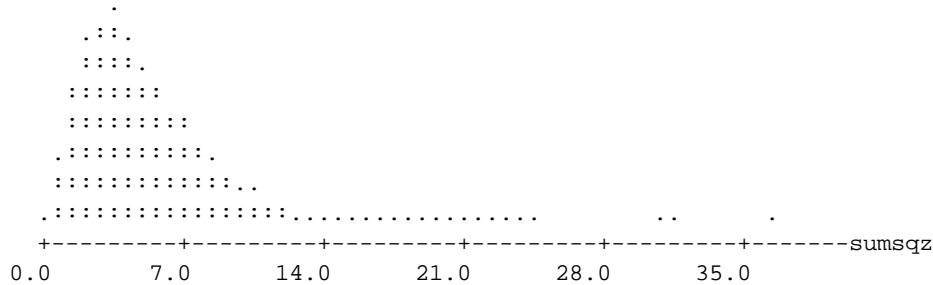
Of course, this is just 5 rows of the 10000 I generated. If we squared each of the z values and then added (across the rows) them up we would find:

sqz1	sqz2	sqz3	sqz4	sqz5	sumsqz
0.8344	4.2339	2.9186	2.7001	0.3891	11.0760
2.3143	6.7371	1.1045	0.0271	2.3559	12.5389
0.7721	0.1942	1.2221	0.0620	0.0072	2.2576
0.2692	0.2763	0.0723	0.0164	0.9178	1.5519
0.9066	0.6920	0.1358	0.1591	3.3665	5.2602

Then we could examine the set of squared z values; ie, adding the squared zs across the rows and then making a graph of this set of 10000 sums.

MTB > dotp c11

Dotplot: sumsqz



MTB > desc c11

Descriptive Statistics: sumsqz

Variable	N	Mean	Median	TrMean	StDev	SE Mean
sumsqz	10000	4.9884	4.3209	4.7445	3.1672	0.0317

Notice that there is now even LESS skewness to the data set and, that the mean is approximately 5 AND the variability as measured by the standard deviation is larger than the previous two cases (degrees of freedom 1 and 2).

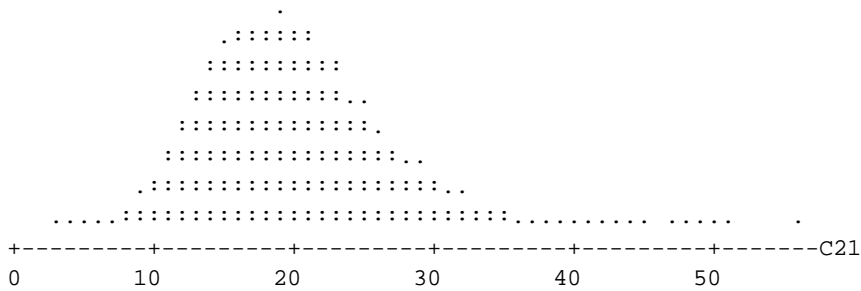
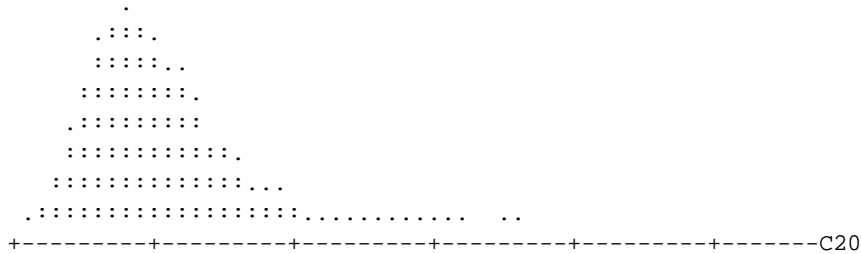
Call the above case: chi square distribution with 5 degrees of freedom (since we added together the squares of FIVE (5) squared normal deviates).

What seems to be happening is that as the number of squared deviates we add together increases, the chi square distribution gets less and less + skewed, the mean increases, and so does the variability (standard deviation).

Without showing all the data, here is what would happen IF we had taken 10 independent random normal deviates, squared the zs, added them together (and also 20), and then had a look at the distributions that result.

```
MTB > dotp c20 c21;
SUBC> same.
```

Dotplot: C20, C21



```
MTB > desc c20 c21
```

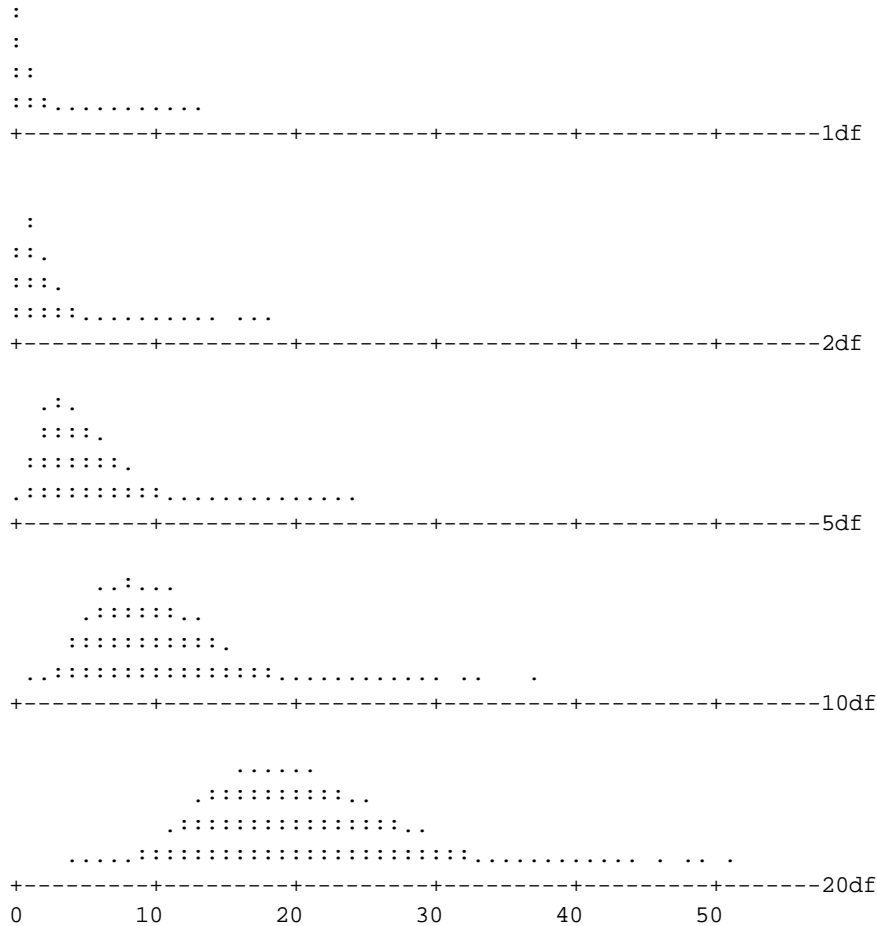
Descriptive Statistics: C20, C21

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C20	10000	9.9979	9.3375	9.7502	4.4667	0.0447
C21	10000	19.907	19.292	19.654	6.286	0.063

We see even less and less skewness and, the means are getting larger and larger, and so too are the standard deviations. Call the two distributions above cases of chi square distributions with 10 and 20 degrees of freedom.

Here, in a quick summary form, is what we have seen as a general pattern (NOTE: I have done new simulations to show the data below; it is just simpler for me to do this).

Dotplot: 1df, 2df, 5df, 10df, 20df



MTB > desc c30-c34

Descriptive Statistics: 1df, 2df, 5df, 10df, 20df

Variable	N	Mean	Median	TrMean	StDev	SE Mean
1df	10000	0.9812	0.4524	0.7840	1.3831	0.0138
2df	10000	2.0207	1.4097	1.7981	2.0058	0.0201
5df	10000	5.0019	4.3742	4.7664	3.1445	0.0314
10df	10000	10.086	9.447	9.840	4.523	0.045
20df	10000	20.116	19.458	19.886	6.327	0.063

Note that the left side end value is 0. What seems to be happening as the df increase, is that someone over on the right side is grabbing the tail of the distribution (with the left side fixed in place) and pulling it more and more to the right. As this gets pulled more and more to the right, the “hump” in the distribution is moving more and more out towards the MIDDLE of the distribution. Thus, clearly if the hump is moving out to the right, the AVERAGE value (mean) is getting larger and larger. Also, if we are dragging it more and more to the right WHILE keeping the left end (more or less) firmly held in place, the overall width of the distribution is increasing; ie, more variability (standard deviation gets bigger).

In summary, here is what we seem to have with chi square distributions.

1. Chi square distributions are created by adding together squared normal deviate values.
2. When only using 1 squared normal deviate, we have 1 degree of freedom. If squaring and adding together 2 or more normal deviate values, we have degrees of freedom 2 or more, depending on how many squared deviates you have drawn independently and added together.
3. As degrees of freedom increase, we see that: A) the distribution becomes less and less + skewed, B) the mean gets larger and larger [in fact, the mean IS = to the degrees of freedom], and C) the standard deviations get larger and larger too [in fact, the formula for the standard deviation is the square root of 2* df].

F DISTRIBUTIONS

We first (very briefly) looked at unit normal distributions and then, with that as the base distribution, expanded our foray into statistical distributions by linking chi square distributions TO unit normal distributions. We now move one step up the statistical distribution ladder to what are called F distributions (F coming from Sir Ronald Fisher).

F distributions are iterations of chi square distributions just like chi square distributions were iterations of unit normal distributions. So, here we go!

What if we first generate a chi square distribution with 3 degrees of freedom. This would mean that we take 3 independently generated unit normal variates, square the zs in each case, and then add them together. Call that Set 1.

Now, independent of Set 1, we generate another set of chi square values, perhaps this time with 5 degrees of freedom. This would mean having 5 unit normal random variates, squaring the zs in each case, and then adding those FIVE values together. Call this Set 2.

F distributions are essentially RATIOS of two chi square distributions. So, if Set 1 is the first chi square distribution with 3 degrees of freedom and Set 2 is the case of a chi square distribution with 5 degrees of freedom, our RATIO in this case would be (in a sense) values where we divide a 3 degree of freedom chi square value by a 5 degree of freedom chi square value.

NOTE: Recall we saw that the MEANS of chi square distributions were = to the degrees of freedom value ... keep that in mind as I continue.

Technically, F distributions are the ratios of two average chi square values that take the following form:

$$F (dfN,dfD) = (\text{chi square } N/df N) / (\text{chi square } D/df D)$$

N = numerator or the TOP chi square distribution

D= denominator or the BOTTOM chi square distribution

An F distribution with 3 and 5 degrees of freedom (Note: F distributions have TWO degrees of

freedom values ... one for the numerator and one for the denominator) would be formed in the following way:

1. First generate a chi square variate with 3 degrees of freedom (this means having 3 independently generated unit normal variates, squaring these, and then summing together).
2. Using the values in 1, take the AVERAGE of these values by dividing each by 3 ... or its degrees of freedom value.
3. Now generate a chi square variate with 5 degrees of freedom (this means having 5 independently generated unit normal variates, squaring these, and then summing these).
4. Using the values in 3, take the AVERAGE of these values by dividing each by 5 ... or its degrees of freedom value.
5. Divide the values you have in 2 by the values you have in 4.

Now you will have a RATIO of two averaged chi square values or distributions, and that is called an F distribution with 3 and 5 degrees of freedom. Let's see if a small Minitab simulation will help. I sure hope so!

Row	sumsqz3	avchis3	sumsqz5	avchis5	F3and5
1	1.3809	0.46031	3.0708	0.61417	0.7495
2	1.9476	0.64920	3.1034	0.62067	1.0460
3	1.7120	0.57066	4.5572	0.91143	0.6261
4	4.1059	1.36862	6.8605	1.37210	0.9975
5	4.1805	1.39351	4.1833	0.83666	1.6655
6	0.8523	0.28408	1.8808	0.37616	0.7552
7	0.7517	0.25058	6.4200	1.28401	0.1952
8	0.8417	0.28055	2.0673	0.41346	0.6786
9	1.3041	0.43471	3.6515	0.73030	0.5953
10	1.4495	0.48316	4.2770	0.85539	0.5648

The sumsqz3 column is the result of taking 3 random unit normal deviates, squaring them, and adding them together. This is your regular chi square value with 3 degrees of freedom. Then, I divided that by $df = 3$ and that produced the avchis3 column. **THIS WILL BE THE NUMERATOR VALUE IN THE F DISTRIBUTION.**

Then I took 5 new independently drawn unit normal variates, squared them, and added them together. This produced the sumsqz5 column. This is a regular chi square value with 5 degrees of freedom. I then divided that by $df = 5$, and got the avchis5 column. **THIS WILL BE THE DENOMINATOR FOR THE F DISTRIBUTION.**

All that is left to do is to **DIVIDE** the avchis3 value (numerator) by the avchis5 value (denominator) and this will then be our F distribution with 3 and 5 degrees of freedom. Here is what distribution looks like.

```
MTB > dotp c22
```

Dotplot: F3and5

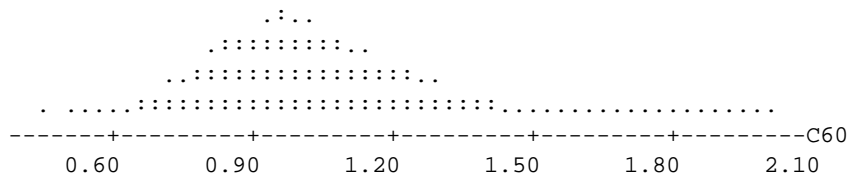
40and10	10000	1.2555	1.0541	1.1662	0.8013	0.0080
Variable	Minimum	Maximum	Q1	Q3		
20and10	0.1548	15.9568	0.7186	1.5321		
30and10	0.1812	12.4927	0.7333	1.5115		
40and10	0.1916	12.5616	0.7590	1.5057		

When the numerator df is larger (and I vary it) while keeping the denominator df smaller and constant, we see the means of F distributions are about 1.25 and, that the distributions are still seriously + skewed.

This might lead you to the conclusion that all F distributions are radically skewed but, that is not the case. Consider the following F distribution with 100 and 100 degrees of freedom.

MTB > dotp c60

Dotplot: C60



MTB > desc c60

Descriptive Statistics: C60

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C60	10000	1.0202	0.9983	1.0126	0.2066	0.0021
Variable	Minimum	Maximum	Q1	Q3		
C60						

Here we see that familiar normal like shape thus, with certain combinations of numerator and denominator df values, the F distribution can look more normal. Note that the mean in this case is around 1, but see that the standard deviation is much SMALLER compared to the other F distributions we saw.

Thus, while the means hover in the vicinity of 1 (maybe somewhat larger), the standard deviations can vary widely depending on the specific N and D df combinations.

It should not be too surprising that some F distributions can look normal like. Remember, F distributions are ratios to two chi square distributions and, we saw that when df values got larger for chi square, the chi square distributions themselves looked more normal like. So, when you have an F distribution where the df value for N and D is in the larger vicinity, you are essentially dividing (to get the F values) two normal like values; hence, the resulting F distribution also can look similar to a normal

distribution. Note I did NOT say that it IS a normal distribution; only that it looks LIKE a normal distribution.

What can we summarize about F distributions? Well, it is a bit more complicated than what we saw the summary was for chi square distributions but, here goes anyway.

1. F distributions are ratios of two AVERAGED chi square distributions.
2. F distributions have 2 df values; one for the N and one for the D.
3. Many F distributions have that radical + skewness shape BUT, some combinations of N and D df values produce F distributions that look similar to a normal distribution.
4. Mean values of F distributions are NEAR 1. [NOTE: actually the mean is about $(df D / (df D - 2))$].
5. Standard deviations of F distributions can vary widely depending on the N and D df values.

t DISTRIBUTIONS

The final distribution we will look at is called the t distribution. Just as the F distribution had a direct connection to the chi square distribution, we will see that the t distribution has a direct connection to BOTH the unit normal distribution AND the chi square distribution. Here we go.

Similar to an F distribution, the t distribution is formed by considering a ratio of two random variables. In this case, we will have in the numerator a regular unit normal distribution value and, in the denominator we will have a chi square variable. Thus, at a generic level, the t distribution looks like:

$$t = \text{unit normal} / \text{chi square}$$

But, let's be a bit more specific.

What if we first dip into a unit normal random variable and select one value. This will be the numerator part in this t formula above. Then, independently of this, we select some chi square distribution (say chi square 4), generate 4 values from a unit normal distribution, square them and add them together, and take an average of these as we did when looking at F distributions. Finally, we form the t ratio as follows:

$$t = \text{Unit normal value} / \text{SQRT}(\text{chi square} / \text{df})$$

Here are the steps:

1. Generate one random unit normal value.
2. Decide on a df value for chi square; say it is 4.
3. Generate 4 random unit normal values and square and add them together.
4. Divide the sum you get in 3 by the df value (in this case 4).
5. Take the square root of the value in 4.

6. Divide the value from 1 (make it the numerator) BY the value in 5 (make this the denominator).
7. Call the value in 6 a t value with a df value of the chi square value in the denominator; ie, in this case, we have a t value with 4 degrees of freedom.
8. Now, think about repeating the steps of 1 to 6, over and over thousands of times so that you could then make a graph of the results and see what happens.

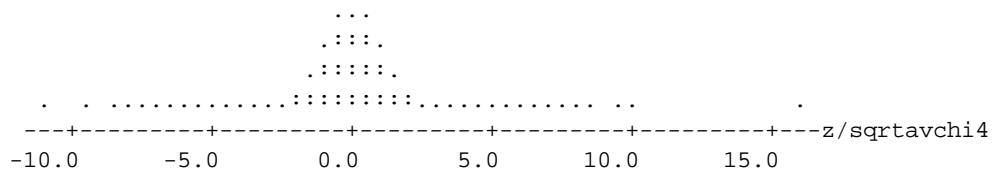
Here is a Minitab simulation that does this.

Row	zN	sumsqz4	avchis4	sqrt	z/sqrtavchi4
1	0.35955	14.6502	3.66256	1.91378	0.1879
2	0.89285	4.2375	1.05938	1.02926	0.8675
3	0.94712	4.5503	1.13758	1.06657	0.8880
4	0.89602	1.5794	0.39485	0.62837	1.4259
5	0.31100	5.0217	1.25542	1.12045	0.2776
6	0.55220	1.5488	0.38721	0.62226	0.8874
7	-0.56848	1.2031	0.30079	0.54844	-1.0365
8	0.05694	5.0047	1.25119	1.11856	0.0509
9	-1.82256	1.7303	0.43259	0.65771	-2.7711
10	1.50361	2.9829	0.74572	0.86355	1.7412

I first generated 10000 values from a unit normal distribution. These are in column zN. Then, independently of that, I generated 10000 sets of 4 unit normal values, squared each of the 4, then added the squared values together. This addition gave me a chi square 4 value. These are in the sumsqz4 column. Then I divided that summed chi square value by the df of 4 and this gives me the averaged chi square value and it is in column avchis4. Then I took the square root of that averaged chi square value and that is in the sqrt column. Finally, to complete the t value calculation, I divided the z value from the zN (numerator) column by the sqrt of the averaged chi square 4 value in the sqrt column, and that gave me the values in the column z/sqrtavchi4. These are our final t values ... and they have 4 degrees of freedom (the df comes from the chi square df in the denominator). Here is what that distribution looks like.

MTB > dotp c112

Dotplot: z/sqrtavchi4



MTB > desc c112

Descriptive Statistics: z/sqrtavchi4

Variable	N	Mean	Median	TrMean	StDev	SE Mean
z/sqrtav	10000	0.0019	-0.0062	-0.0034	1.3941	0.0139

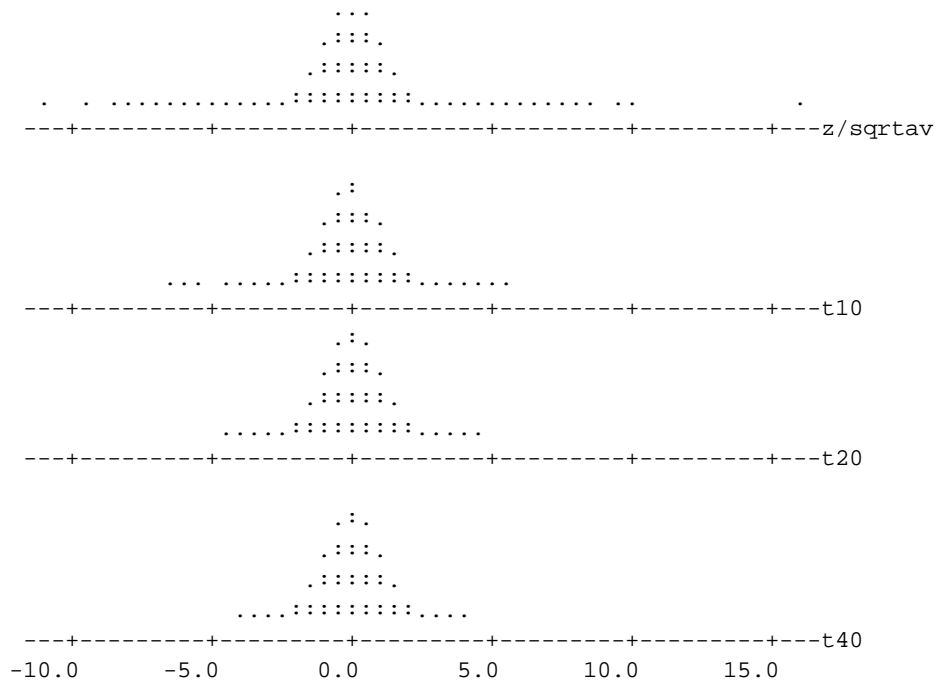
Variable	Minimum	Maximum	Q1	Q3
z/sqrtav	-10.9034	15.9094	-0.7316	0.7369

Note that unlike most of the chi square and F distributions we saw that were radically + skewed, this t distribution is taking on that familiar symmetrical unimodal looking shape; ie, the normal distribution shape. Note that the mean is about 0 and the standard deviation is about 1.4.

While the numerator in this t distribution is always a regular unit normal value, the denominator can vary depending on whether we are using a chi square 4 distribution, or a chi square 9 distribution, or other df values for chi square distributions. Thus, while the N seems to be rather constant, the D will change depending on the df for the selected chi square distribution. What this means is that the RATIO will vary depending on the D. And, if the D varies depending on the df for the chi square distribution, then the t DISTRIBUTION that results from this N and D division or ratio, might be and look different. Without going though all these other possible simulation scenarios, let me use the “power” of Minitab to simply generate random distributions from t distributions that have different df values. This is easily accomplished in Minitab by a simple random generation command. What I am showing below is our case of df=4 plus, df values of 10, 20 and 40. Here is what those different t distributions look like.

```
MTB > dotp c112-c115;
SUBC> same.
```

Dotplot: z/sqrtavchi4, t10, t20, t40



```
MTB > desc c112-c115
```

Descriptive Statistics: z/sqrtavchi4, t10, t20, t40

Variable	N	Mean	Median	TrMean	StDev	SE Mean
z/sqrtav	10000	0.0019	-0.0062	-0.0034	1.3941	0.0139
t10	10000	-0.0198	-0.0308	-0.0178	1.1177	0.0112
t20	10000	0.0052	0.0033	0.0045	1.0603	0.0106
t40	10000	0.0087	0.0034	0.0060	1.0297	0.0103

Note that all of these t distributions, from the first one with $df=4$ to the last one with $df=40$... all of these look like the typical normal distribution. Also note that the centers or means of these t distributions are approximately 0. Finally, note that the WIDTH of these t distributions seems to get narrower as we INCREASE the df value. In fact, with larger df values, the standard deviations look very similar to the standard deviation value we get in a unit normal distribution AND, to boot, the mean of 0 looks the same! In fact, t distributions look very much LIKE unit normal distributions and more and more like them (approach in width) when df gets larger and larger. So, what can we say about t distributions?

1. t distributions are ratios of unit normal values TO chi square values.
2. The shapes of t distributions are symmetrical and unimodal, like normal distributions.
3. The mean or center of t distributions is 0.
4. The width or standard deviations of t distributions get closer and closer to 1 as df gets larger.
5. In summary, t distributions look more and more like unit normal distributions as df increases more and more.

If you think about the formula for a t distribution, the fact that values from it produce a distribution that is similar to a unit normal distribution is not that surprising. First remember that the N part IS a unit normal distribution. But, in reality, though it does not look like it ... the denominator is like a unit normal distribution too. Remember that we SQUARED z values and added them together to find a chi square distribution. And recall that these were usually radically + skewed. But, if you start with the chi square distribution that is radically + skewed, and then take an average and square root, it is in a sense, like starting BACK with individual z values from a unit normal distribution. So, in effect, a t distribution is like a ratio of dividing one unit normal distribution BY another unit normal distribution thus, the result of dividing one normal distribution by another normal distribution is ... well, just ANOTHER normal distribution. And that in fact is what t distributions look like.

SUMMARY OF CHARACTERISTICS OF THE DISTRIBUTIONS

1. We start with the unit normal distribution which has mean=0 and standard deviation=1.
2. Chi square distributions are formed by taking unit normal deviate values and squaring them. How many you take and square and sum together, determines how many degrees of freedom there are.
3. Chi square distributions tend to be rather + skewed with small degrees of freedom; have a mean = to df, and a standard deviation = $\sqrt{2*df}$. At larger df values, chi square distributions can look more normal like.
4. F distributions are essentially a ratio of TWO chi square variates. This means that there is a df for the numerator and a df for the denominator. Means of F distribution are near 1 or a bit

larger and, the standard deviations can vary widely. Many F distributions are also radically + skewed but, certain combinations of N df and D df, produce more normal looking F distributions.

5. t distributions are essentially a ratio of a unit normal variate and a chi square variate. Degrees of freedom for the t distribution are determined by the chi square variate in the denominator. t distributions look very much like unit normal distributions with a mean of 0 and, a standard deviation that gets closer to 1 as df increases.

SAMPLE USES OF THE DISTRIBUTIONS

1. Normal Distribution

If you took many random samples of say $n=70$ each from a population that was normally distributed, and then make a frequency distribution graph of the MEANS you get from all your samples, the SHAPE of that “sampling distribution of means” would look like the normal distribution. This would allow you to say such things as 68% of all sample means would fall between the population mean and 1 standard error unit (ie, the standard deviation of the distribution of sample means) on either side of the population mean.

2. Chi Square Distribution

A simple inferential technique is to take a simple SAMPLE poll like: males/females as one variable and agree (to some issue) or NOT agree (as the other variable) and ask if there is some relationship between these two variables in the larger population from which the sample was drawn. This would give us a 2 by 2 frequency cross tabs table and, we can test the null hypothesis that there is 0 relationship. We would calculate a chi square test statistic based on our sample data and then consult and chi square table to make a decision with respect to retaining or rejecting the null hypothesis.

3. F Distributions

A very popular use of F distributions is in the application of a technique called “analysis of variance” (ANOVA). A typical setting would be in an experiment where we manipulated both TYPE of instructional method and AMOUNT to time allowed for student practice. In this two factor design, we would calculate an ANOVA summary table with test statistic F ratios for each of the factors (TYPE and AMOUNT, and interaction between TYPE and AMOUNT) and compare our values with critical values in the appropriate F tables. Then we would retain or reject the various null hypotheses.

4. t Distributions

Perhaps the most common (but not only) application of the t distribution would be in a simple

two group experiment (experimental and control groups) where we are interested in the potential difference in the two respective POPULATION means. We look at the difference in the sample means, calculate our t test statistic, then compare it to our table of critical t values, and make a decision to retain or reject the null hypothesis.

There are dozens and dozens of uses of the above distributions and, the ones mentioned only touch the surface. However, the normal, chi square, F and t distributions are very frequently used and popular within statistical work.