Chi-Square and F-Distributions, and Dispersion Tests

Recall from Chapter 4 material on:

$E(Y)=\mu$
$E(Y-\mu)^2=\sigma_Y^2$

If Y is normally Distributed, then $Z=\frac{Y-\mu}{\sigma_Y}$ and

$E(Z)=\mu_Z=0$

$$E(Z-\mu_Z)^2=E(Z)^2=\sigma_Z^2=E\left(\frac{Y_i-\mu}{\sigma_Y}\right)^2=\frac{E(Y_i-\mu)^2}{\sigma_Y^2}=\frac{\sigma_Y^2}{\sigma_Y^2}=1.0$$

If the population distribution of Y is normally distributed, then Z will also be normally distributed, and is called the **standard Normal Z**.

Now <u>consider the distribution of squared Z's</u>:

$$Z_i^2=\frac{(Y_i-\mu_Y)^2}{\sigma_Y^2}$$

Imagine what the shape of the distribution so Z^2 would look like. Think of squaring each value under a standard normal distribution. This produces an outcome which eliminates all negative value, and yields a probability distribution which appears to peak at zero, and show declining probabilities as Z^2 increases. Notice where it appears to be located (*i.e.*, the expected value).

This squared Z distribution is called a **Chi-Squared Distribution**.
Its expected value is 1.0, and its variance is 2.0. It is thus the distribution of one squared Z.

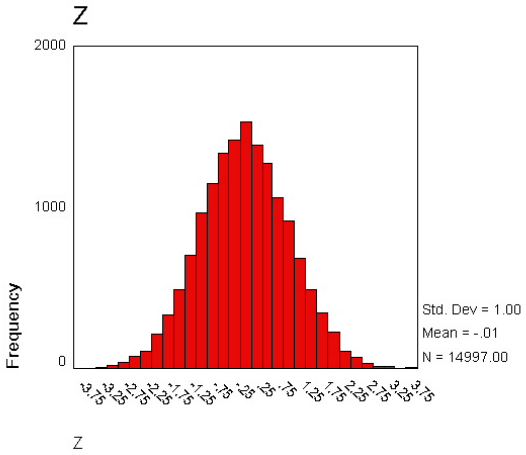
Now consider summing squared Z's, where each is randomly, and independently, sampled. Thus we produce a new Random variable, where the value is the sum of squared Z's.

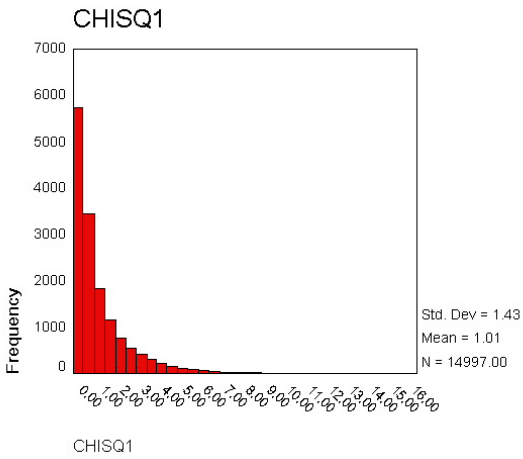
$Z_1^2+Z_2^2=\chi_2^2,$
$Z_1^2+Z_2^2+Z_3^2=\chi_3^2,$ *and in general,* $=$
$\sum_{i=1}^{v}Z_i^2=\chi_v^2$

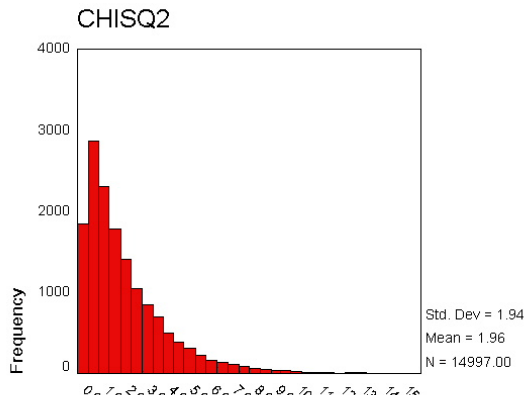
It is also the case that $\chi_{v_1}^2+\chi_{v_2}^2=\chi_{v_1+v_2}^2$ (Also can take diffs where df are also diff)

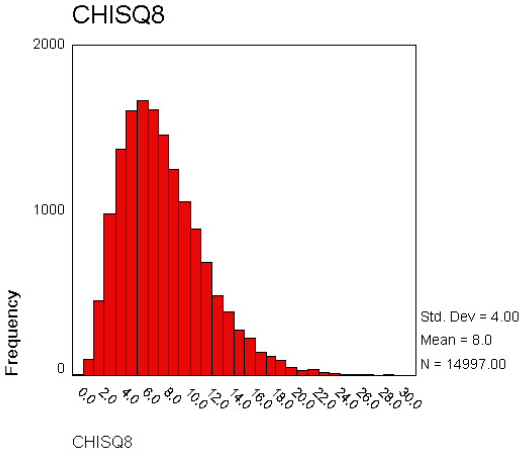
Simulations of Std Normal Z and Chi-Square

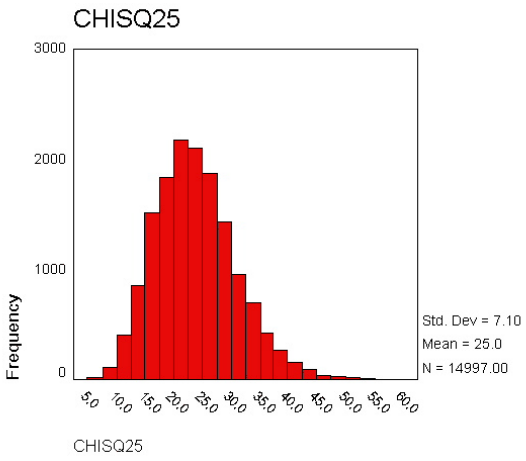
Run with N=14,997 and graphed in SPSS as a "Histogram". Thus the intervals represent grouping of the X values into equal intervals for purposes of production of these histograms. The actual distributions are continuous. Each Chi-square distribution is titled CHISQ(df) where the number represents degrees of freedom. Note the asymmetry of Z even with this large N - simulation is not perfect.











Note on all distributions,

$E\left(\chi_v^2\right) = v$, where v=df, and

$\sigma_{\chi_v^2}^2 = 2v$. Thus each of these five distributions is "centered" on their expected value.

Algebraic Basis relating Chi-Squared Variables, Population and Sample Variances, and the One-Sample Dispersion Test

Starting point is re-examination of s^2, in light of above: $s_Y^2 = \frac{\sum_{i=1}^{N}\left(Y - \mu\right)^2}{N - 1}$

Thus,

$$\frac{s_Y^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{N}\left(Y-\mu\right)^2}{\left(N-1\right)\sigma_Y^2} \; and$$

$$\frac{\left(N-1\right)s_Y^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{N}\left(Y-\mu\right)^2}{\sigma_Y^2} = \sum Z^2 = \chi_{N-1}^2$$ Recall that N-1 = df

Therefore, for a sampling situation where N data points are sampled randomly, and a variance calculated:

$$\frac{\left(N-1\right)s^2}{\sigma^2} = \chi_{N-1}^2$$

also, from a two−group pooled situation:

$$\frac{\left(N_1+N_2-2\right)s_{Pooled}^2}{\sigma^2} = \chi_{N_1+N_2-2}^2$$

Thus rearranging the above formula $\frac{\left(N-1\right)s_Y^2}{\sigma_Y^2} = \chi_{N-1}^2$ gives $s^2 = \frac{\sigma_Y^2\chi_{N-1}^2}{N-1}$.

Thus

$\frac{s^2}{\sigma_Y^2} = \frac{\chi_{N-1}^2}{N-1}$ and more generally, $\frac{s^2}{\sigma_Y^2} = \frac{\chi_v^2}{v}$ and $\frac{s^2}{\sigma_Y^2} = \frac{est\,\sigma_Y^2}{\sigma_Y^2} = \frac{\chi_v^2}{v}$

These algebraic substitutions enable the creation of a test about population dispersion hypotheses.

Goal: (1) We want to test the hypothesis that a population variance (unknown) equals a specified value.
(2) Use the standard approach of NHST
(3) Create a test statistic which reflects the "best" critical test of the hypothesis.

H_0: $\sigma^2 = \sigma_0^2$, where σ_0^2 is some specified value (e.g., H_0: $\sigma_0^2 = 225$)

Consider a sampling situation where we draw a random sample of size N, and calculate s^2, the sample variance on those data. Do our data give us reason to reject the null hypothesis?

H_1: $\sigma^2 \neq \sigma_o^2$ if two-tailed, non-directional

$\sigma^2 < \sigma_o^2$ if one-tailed (use left tail of test distribution)

$\sigma^2 > \sigma_o^2$ if two-tailed, (use right tail of test distribution)

Form the test statistic from the above algebra:

$$\frac{(N-1)s_Y^2}{\sigma_Y^2} = \chi_{N-1}^2$$

Example:
One-tailed, alpha = .05

H_0: $\sigma_0^2 = 225$ *and* $\sigma_0 = 15$ N=26

H_1: $\sigma_0^2 > 225$ $s^2 = 350$ (thus s= 18.7)

[(25)350]/225 = 38.89

Using Appendix, Table IV, we see that the Critical Value for χ_{25}^2 = **37.6525**

Thus our observation falls in the region of rejection and we **reject H_0.**

The sample data are deviant enough so that (with these df) we can reject the null.

Use table 4 for df up to 100. Above that, use the "large sample approximation" in Hays section 9.6, especially equation 9.6.3

Also examine section 9.5 to see that confidence intervals for s and s^2 are available, and based on the use of the chi-squared distribution.

Assumption: Normality of the random variable population distribution is more important here than for the location tests considered earlier (section 9.7 in Hays).

F-Distribution

The F-distribution is defined as the ratio of two chi-squared variables, each divided by their own respective df.

$$\frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2}$$

Thus each F has both numerator (v_1) and denominator (v_2) associated with it. As is the case for "t" and Chi-squared, F is a family of distributions. This is why tables of critical values are extensive combinations of numerator and denominator df.

Now consider the usage of F to test two-sample dispersion hypotheses.

The Two-Sample Dispersion Problem

H_0: $\sigma_1^2 = \sigma_2^2$ where σ_1^2 *and* σ_2^2 represent population variance of two populations

H_A: $\sigma_1^2 \neq \sigma_2^2$, yielding a non-directional, two-tailed test here in this illustration

Our task, once again is to find an appropriate test statistic.

Recall that $s^2 = \frac{\sigma_Y^2 \chi_{N-1}^2}{N-1}$ and more generally, $s^2 = \frac{\sigma^2 \chi_\nu^2}{\nu}$

So, then $s_1^2 = \frac{\sigma_1^2 \chi_{N_1-1}^2}{N_1-1}$ and $s_2^2 = \frac{\sigma_2^2 \chi_{N_2-1}^2}{N_2-1}$ when sample variances are calculated on two different samples, drawn at random.

If the null is true, then the following ratio is distributed as F, (when the random sampling and normality assumptions are met)

$\frac{s_1^2}{s_2^2}$ is distributed as $\frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2} = F_{(v1,v2)}$

$E(F_{v1,v2}) = v_2/(v_2-2)$

Thus the expected value of an F is close to 1.0, but not exactly. As denominator df rises, the expected value will be closer to one.

Now, if we draw two samples, at random, from the same population, or from two populations which have the same variance, then

$\frac{s_1^2}{s_2^2}$ is distributed as F with $v_1=N_1-1$, and $v_2=N_2-1$

This approximation holds well if the normality assumption is met.

Consider an illustration:

Sample	s^2	N
I	400 (sd=20)	11
II	144 (sd=12)	15

F(10,14)= 400/144=2.78

It is typical in this test to place the larger of the two values in the numerator and use only the upper tail of the F distribution. Even though it is technically a two-tailed test, this arrangement allows us to place the whole alpha in the upper tail and simplify the process. (See Hays section 9.9 for use of the F-tables and how to find lower tail probabilities if ever necessary)

From Table V, we see that the C.V. for an $F_{(10,14)} = 2.60$ when $\alpha = .05$

Conclusion: reject H_0 as a 2-tailed test with $\alpha = .05$

We conclude that the first sample was derived from a population which did not have the same variance as the population from which the second sample was derived.

Notice two things:
1. The ratio of the sd's does not have to exceed 2/1 in order for rejection of some null hypotheses.
2. Recognize how this might help us in deciding the appropriateness of the homogeneity of variance assumption for the two-sample location test we considered earlier.

From Table V, note that C.V. $F_{(1,15)} = 4.54$ and sqrt 4.54 = 2.13
- why interesting?

Relationships among the Distributions

1. Normal is parent for each and limiting form of t and chi-square.
2. "t" approximates std normal Z when df are large.
3. Chi-square is a sum of squared Z's
4. F is the ratio of two chi-squareds, each divided by their own df
5. $t^2_{v_2} = F_{(1,V_2)}$

Proof of point #5:

$t_{N-1} = \dfrac{\bar{Y} - \mu}{\sqrt{s^2 / N}}$ from the approximation permitted in the one-sample location test

$t_{N-1} = \dfrac{\left(\bar{Y} - \mu\right) / \sigma_Y}{\sqrt{s^2 / N\sigma_Y^2}}$ Divide numerator and denominator by pop s.d. $\left(\sigma_Y\right)$

$$t_{N-1} = \frac{\left(\bar{Y} - \mu\right)\sqrt{N} / \sigma_Y}{\sqrt{s^2 / \sigma_Y^2}} = \frac{\left(\bar{Y} - \mu\right) / \sigma_Y / \sqrt{N}}{\sqrt{s^2 / \sigma_Y^2}} = \frac{\left(\bar{Y} - \mu\right) / \sigma_M}{\sqrt{s^2 / \sigma_Y^2}}$$

Multiply both numerator and denominator by $\sqrt{N}$, and rearrange numerator to place the $\sqrt{N}$ term in the denominator of the numerator ratio.

Now square both sides:

$$t^2_{N-1} = \frac{\left(\left(\bar{Y} - \mu\right) / \sigma_M\right)^2}{s^2 / \sigma_Y^2} = \frac{Z_M^2}{s^2 / \sigma_Y^2} = \frac{\chi_1^2 / 1}{\chi^2_{N-1} / N - 1}$$

$= \mathbf{F}_{(\mathbf{1,N-1})}$

From definitions shown above, both the numerator and denominator are now chi-squared variables, each divided by their respective df.

It is also possible to derive this proof for the two-sample "t", but we won't here.

Considering these four primary probability density functions, and families of functions, it would help to summarize each of their respective

Expected Values
and
Variances (not shown for F)

and recognize their shapes as N or df becomes large.

Rely, in part, on section 9.11 in the textbook for this.