The x^2 (AKA chi-square) Distribution

Here's a new distribution that is somewhat helpful to understand for today. It is the sum of the squares of a bunch of independent standard Gaussians. Here's a good reference page for it.

NIST reference page

A few facts, which I got from that page:

- Mean = v = degrees of freedom
- Mode v-2 (for v>2)

Here are a pair of Extend models that can generate it for us. The first demonstrates the use of the TimeOut block and "R" connector for the Accumulator to reset an accumulator during a simulation. We also selectively histogram things. You may find this a useful technique in your projects.

- chi2a.mox
- chi2b.mox

The X^2 Statistical Test

Recently, we talked about how random numbers are generated. Today, we'll look at a statistical test that is used to test "count" data: how many of each of several different categories. This is called the X^2 test (pronounced chi-square) from the Greek letter that is associated with it (much as we say z-test or t-test). We'll also see the X^2 test can be used to test independence.

The X^2 test is typically used when we have "count" data: data in which we count the number of items or events or whatever, and we have some expectation of how many we will observe. The test determines whether the deviations of our observations from our expectations is attributable to chance, under the *assumption* that the deviations from expectation is *Gaussian*.

Aside

You'll note a problem with this: if the data are count data, the deviations from the expected value will not be *continuous* If you expect 4.7 sparrows, and you count 5 sparrows or 6 sparrows, the deviations are 0.3 and 1.3, but the probability of a deviation of 0.8 is *zero*. This is one reason that the chi-square test is considered inapplicable if the actual counts or expected counts are too small, because effects of the discontinuous deviations is not negligible. The usual rule of thumb is that an expected

count of at least 5 is okay; this is because the binomial is reasonably approximated by the Gaussian when *np* is at least 5.

Example

Let's take an example:

- Set up a simulation that collects 1000 uniform random numbers and puts them into 5 bins.
- If the numbers are really uniform (our null hypothesis), there should be 200 numbers in each bin; that is, the **expected** number is 200.
- It's unlikely that we'll see exactly 200 in each bin. But, is the value we see, called the **observed** value, okay?
- You'll notice that, under H0, the observation is **binomial**. Why? We have n=1000 trials, with p=0.2 chance of falling in this bin, and the observed value is the number of successes.
- For large n, the binomial distribution looks reasonably Gaussian, and the Gaussian is easier for the mathematical statisticians. So, they assumed that the **deviation** (**observed-expected**) is Gaussian.
- If we add up all the deviations, though, we get zero. So, they **square** the deviation, and then divide by the expected number, yielding a number that is, under H0, a standard Gaussian. (Mean zero, unit variance.)
- The **sum** of K of these has a **chi-square** distribution with K-1 degrees of freedom.

Here is the formula for a chi-square statistic:

$X^2 = \Sigma \text{ (observed-expected)}^2/\text{expected}$

Let's calculate the chi-square value for our model and test it. You can test the significance (and find the p-value):

- with Excel using the **chidist(x,dof)** function, or
- with table V on page 492 of your book,
- any other other the chi-square table of critical values.

In each case, you must use the appropriate number of *degrees of freedom* (dof).

In almost every case, dof is one less than the number of bins.

Why is the Degrees of Freedom like that?

Consider a chi-square test with just two bins. The deviation of the second is the same as the deviation of the first, just in the opposite direction. (That's why they add up to

zero.) So, you don't really have two independent samples from a Gaussian here (which is our null hypothesis); we have just one. In general, the deviation of the last bin is determined by the aggregate deviations of the others, so that we really only have n-1 samples from a Gaussian. That's why the degrees of freedom is n-1.

In some circumstances (namely, where the overall number of counts isn't fixed), it is appropriate to use n as the degrees of freedom rather than n-1. For example, you go out to the field and count bird species for one hour or one day or something. Your data is the number of each species. You can test whether that fits your expectation using n degrees of freedom.

Why not use the t-test?

If these deviations are assumed to be Gaussian, why not use the t-test or the z-test? That is, just use a t-test for each bin?

The answer is very important:

> **we want to do exactly one test, so that we can control the p-value.**

If we do multiple tests, we have problems with the p-value. Consider the following (extreme) example:

> Set up a chi-square table with 100 bins. Do a t-test for each one, at a significance level (p-value) of 0.05. Even if the null hypothesis is true, we expect to reject it 20 times. What, then, is our criteria for rejecting H0? What is the p-value for the test?

This reminds me of a story ...

Using Chi-Square to test Independence

The chi-square test is generally used to test for fit to a discrete distribution. For example, you could use it to see if some data fits a binomial or negative binomial or a Poisson.

The chi-square test also finds a lot of use in testing for independence, so much so that some people forget its original purpose. Let's see an example:

- Suppose we believe that big law firms are biased against women, in the sense that women are less likely to make partner. The other side argues that the firms are not biased and that making partner is **independent** of someone's sex or gender.

- Let's make up some data. Suppose this is the aggregate data for big Boston law firms:

	made partner		
	yes	**no**	**total**
male	90	35	
female	45	30	
total			

- We can do the rest of this in Excel:
- Calculate the **marginals**: the fraction that are male, female, made partner, and didn't. For example, fraction that are male is 125/200.
- Calculate the **probability** of landing in each cell, under the assumption of independence. For example, the probability of being a man who made partner is (125/200)*(135/200)
- Calculate the **expected** number in each cell, under the assumption of independence. Here is the general formula:

$E_{i,j} = (\text{RowTotal}_i * \text{ColTotal}_j)/\text{TableTotal}$

For example, the expected number of men who made partner is

$E_{men,partner} = (\text{RowTotal}_{men} * \text{ColTotal}_{partner})/\text{TableTotal}$
$E_{men,partner} = (125 * 135)/200$
$E_{men,partner} = 84.375$

- Calculate the chi-square statistic.
- The degrees of freedom is (R-1)*(C-1) where R is the number of rows and C the number of columns. So, for a 2x2 chi-square table (also called a contingency table), the number of degrees of freedom is just 1.
- Do you accept or reject H0? You can check the significance (p-value) using
  - chidist(x,dof),
  - with table V on page 492 of your book,
  - any other other the chi-square table of critical values.

Why is Degrees of Freedom Like That?

This is a generalization of the idea from the test-of-fit use of the chi-square, where we realized that only k-1 of the bins were free to vary. With an RxC 2D table, the use of the row and column margins means that each row is only contributing C-1 independent numbers and each column is only contributing R-1 independent numbers. So the degrees of freedom is the product of R-1 and C-1.

Other Resources

Here is a good tutorial on the chi-square test of independence

Serial Correlation

We can use this idea to test **serial correlation** of Extend's random number generator. Here's a model:

serial correlation

We'll spend some time exploring this and then test using chi-square.