Chi-Square

In this Statistics Appendix Lecture, we'll go over the Chi-Square Distribution and the Chi-Square Test.

Note: Before viewing this lecture, see the Hypothesis Testing Notebook Lecture.

Let's start by introducing the general idea of observed and theoretical frequencies, then later we'll approach the idea of the Chi-Sqaure Distribution and its definition. After that we'll do a qcuik example with Scipy on using the Chi-Square Test.

Supppose that you tossed a coin 100 times. Theoretically you would expect 50 tails and 50 heads, however it is pretty unlikely you get that result exactly. Then a question arises... how far off from you expected/theoretical frequency would you have to be in order to conclude that the observed result is statistically significant and is not just due to random variations.

We can begin to think about this question by defining an example set of possible events. We'll call them Events 1 through *k*. Each of these events has an expected (theoretical) frequency and an observed frequency. We can display this as a table:

Event	**Event 1**	**Event 2**	**Event 3**	**...**	**Event k**
Observed Frequency	o_1	o_2	o_3	...	o_k
Expected Frequency	e_1	e_2	e_3	...	e_k

Since we wanted to know whether observed frequencies differ significantly from the expected frequencies we'll have to define a term for a measure of discrepency.

We'll define this measure as Chi-Square, which will be the sum of the squared difference between the observed and expected frequency divided by the expected frequency for all events. To show this more clearly, this is mathematically written as:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \ldots + \frac{(o_k - e_k)^2}{e_k}$$

Which is the same as:

$$\chi^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j}$$

If the total frequency is N

$$\sum o_j = \sum e_j = N$$

Then we could rewrite the Chi-Square Formula to be:

$$\chi^2 = \sum \frac{o_j^2}{e_j^2} - N$$

We can now see that if the Chi Square value is equal to zero, then the observed and theoretical frequencies agree exactly. While if the Chi square value is greater than zero, they do not agree.

The sampling distribution of Chi Square is approximated very closely by the *Chi-Square distribution*

The Chi Square Distribution

The Chi-Square Distribution is related to the standard normal distribution. If a random variable Z, then Z^2 has the Chi Square distribution with one degree of freedom. This idea is best presented graphically in a video. I've embedded a video below which goes over this an a way better than this static iPython Notebook format.

Here is an excellent video explaining the basics of the Chi Square Distribution.

In [27]:
```
from IPython.display import YouTubeVideo
YouTubeVideo("hcDb12fsbBU")
```

Out[27]:

The Chi Square Test for Goodness of Fit

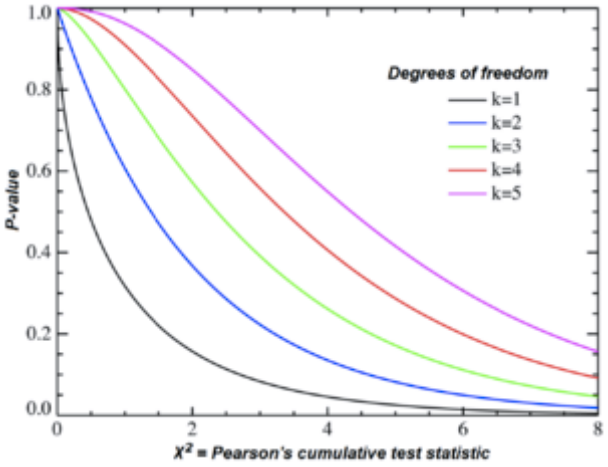
We can now use the Chi-Square test (http://stattrek.com/chi-square-test/goodness-of-fit.aspx?Tutorial=AP) can be used to determine how well a theoretical distribution fits an observed empirical distribution.

Scipy will basically be constructing and looking up this table for us:

In [35]:
```
url='http://upload.wikimedia.org/wikipedia/commons/thumb/8/8e/Chi-square_distribu

from IPython.display import Image
Image(url)
```

Out[35]:



Let's go ahead and do an example problem. Say you are at a casino and are in charge of monitoring a craps (http://en.wikipedia.org/wiki/Craps)(a dice game where two dice are rolled). You are suspcious that a player may have switched out the casino's dice for their own. How do we use the Chi-Square test to check whether or not this player is cheating?

You will need some observations in order to begin. You begin to keep track of this player's roll outcomes.You record the next 500 rolls taking note of the sum of the dice roll result and the number of times it occurs.

You record the following:

Sum of Dice Roll	2	3	4	5	6	7	8	9	10	11	12
Number of Times Observed	8	32	48	59	67	84	76	57	34	28	7

Now we also know the espected frequency of these sums for a fair dice. That frequency distribution looks like this:

Sum of Dice Roll	2	3	4	5	6	7	8	9	10	11	12
Expected Frequency	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Now we can calculated the expected number of rolls by multiplying the expected frequency with the total sum of the rolls (500 rolls).

In [19]:
```
# Check sum of the rolls
observed = [8,32,48,59,67,84,76,57,34,28,7]
roll_sum = sum(observed)
roll_sum
```

Out[19]:
```
500
```

In [16]:
```
# The expected frequency
freq = [1,2,3,4,5,6,5,4,3,2,1]

# Note use of float for python 2.7
possible_rolls = 1.0/36

freq = [possible_rolls*dice for dice in freq]

#Check
freq
```

Out[16]:
```
[0.027777777777777776,
 0.05555555555555555,
 0.08333333333333333,
 0.1111111111111111,
 0.1388888888888889,
 0.16666666666666666,
 0.1388888888888889,
 0.1111111111111111,
 0.08333333333333333,
 0.05555555555555555,
 0.027777777777777776]
```

Excellent, now let's multiply our frequency by the sum to get the expected number of rolls for each frequency.

In [23]:
```
expected = [roll_sum*f for f in freq]
expected
```

Out[23]:
```
[13.888888888888888,
 27.777777777777775,
 41.666666666666664,
 55.55555555555555,
 69.44444444444444,
 83.33333333333333,
 69.44444444444444,
 55.55555555555555,
 41.666666666666664,
 27.777777777777775,
 13.888888888888888]
```

We can now use Scipy to perform the Chi Square Test (http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.chisquare.html) by using chisquare.

In [26]:
```
from scipy import stats

chisq,p = stats.chisquare(observed,expected)

print 'The chi-squared test statistic is %.2f' %chisq
print 'The p-value for the test is %.2f' %p
```

```
The chi-squared test statistic is 9.89
The p-value for the test is 0.45
```

stats.chisquare returns two values, the chi-squared test statistic and the p-value of the test.

With such a high p-value, we have no reason to doubt the fairness of the dice.

That's it for the Chi-Square Distirbution and Test!

For more information, check out these links:

1. Wikipedia (http://en.wikipedia.org/wiki/Chi-squared_test)
2. Stat trek (http://stattrek.com/chi-square-test/independence.aspx)
3. Khan Academy (https://www.khanacademy.org/math/probability/statistics-inferential/chi-square/v/pearson-s-chi-square-test-goodness-of-fit)

In [ ]: