Understanding χ^2 (chi-squared)

In class, I showed a way of calculating χ^2 with slightly less complicated notation. Let's imagine the same setup as in class:

	var 2		
var 1	0	1	
0	a	b	a+b
1	c	d	c+d
	a+c	b+d	a+b+c+d

Here, a, b, c, and d are all frequencies - they're the counts of observations in each cell of the contingency table. We'll use the example from lecture to illustrate:

```
. tab mid4 dembin
```

```
           |        dembin
      mid4 |         0          1 |     Total
-----------+----------------------+----------
         0 |    10,698      3,979 |    14,677
         1 |     2,965      1,490 |     4,455
-----------+----------------------+----------
     Total |    13,663      5,469 |    19,132
```

Here,

$$\begin{aligned} a &= 10698 \\ b &= 3979 \\ c &= 2965 \\ d &= 1490 \end{aligned}$$

The first thing we need to do is find the expected value in each cell, call this E. The expected value is the *row marginal* (call this m_r multiplied by the *column marginal* call this m_c, divided by the total number of observations n. We are trying to calculate what we would have expected to see if the two variables really were statistically independent.

Let's start with the upper left-hand cell which has value 10698. Its row marginal value is 14677 and its column marginal value is 13663. For all of these cells, the total number of observations is 19132. So, for this cell, we can calculate the expected value:

$$\begin{aligned} E &= \frac{m_r \times m_c}{n} \\ &= \frac{14677 \times 13663}{19132} \\ &= \frac{200531851}{19132} \\ &= 10481.49 \end{aligned}$$

We can do this for the rest of the cells as well:

- Upper right-hand cell:

$$\begin{aligned} E &= \frac{m_r \times m_c}{n} \\ &= \frac{14677 \times 5469}{19132} \\ &= \frac{80268513}{19132} \\ &= 4195.51 \end{aligned}$$

- Lower left-hand cell:

$$\begin{aligned} E &= \frac{m_r \times m_c}{n} \\ &= \frac{4455 \times 13663}{19132} \\ &= \frac{60868665}{19132} \\ &= 3181.51 \end{aligned}$$

- Lower right-hand cell:

$$\begin{aligned} E &= \frac{m_r \times m_c}{n} \\ &= \frac{4455 \times 5469}{19132} \\ &= \frac{24364395}{19132} \\ &= 1273.49 \end{aligned}$$

Now, we can fill in the table:

Table 1: Observed and Expected Values

(a) Observed

	dembin	
mid4	0	1
0	10698	3979
1	2965	1490

(b) Expected

	dembin	
mid4	0	1
0	10481.49	4195.51
1	3181.51	1273.49

Now, we need to find $(Observed - Expected)^2$.

- Upper left-hand cell:

$$\begin{aligned}(O-E)^2 &= (10698-10487.49)^2 \\ &= 216.51^2 \\ &= 46876.58\end{aligned}$$

- Upper right-hand cell:

$$\begin{aligned}(O-E)^2 &= (3979-4195.51)^2 \\ &= -216.51^2 \\ &= 46876.58\end{aligned}$$

- Lower left-hand cell:

$$\begin{aligned}(O-E)^2 &= (2965-3181.51)^2 \\ &= -216.51^2 \\ &= 46876.58\end{aligned}$$

- Upper left-hand cell:

$$\begin{aligned}(O-E)^2 &= (1490-1273.49)^2 \\ &= 216.51^2 \\ &= 46876.58\end{aligned}$$

Notice here that these are all the same. This will always be the case in two-by-two tables, but may not be the case in tables of other dimensions. Now, we can calculate the rest of the formula $\frac{(O-E)^2}{E}$.

- Upper left-hand cell:

$$\begin{aligned}\frac{(O-E)^2}{E} &= \frac{46876.58}{10481.49} \\ &= 4.47\end{aligned}$$

- Upper right-hand cell:

$$\begin{aligned}\frac{(O-E)^2}{E} &= \frac{46876.58}{4195.51} \\ &= 11.17\end{aligned}$$

- Lower left-hand cell:

$$\begin{aligned}\frac{(O-E)^2}{E} &= \frac{46876.58}{3181.51} \\ &= 14.73\end{aligned}$$

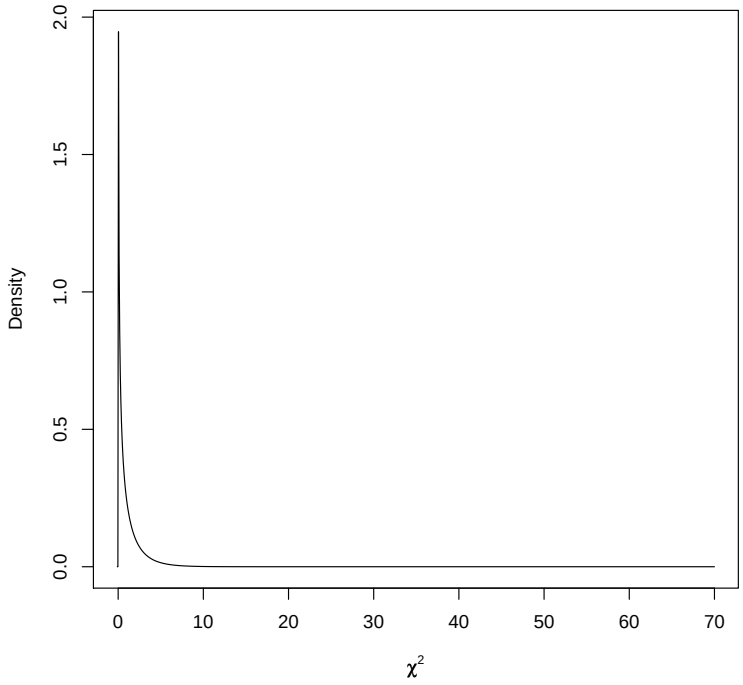
- Lower right-hand cell:

$$\begin{aligned}\frac{(O-E)^2}{E} &= \frac{46876.58}{1273.49} \\ &= 36.81\end{aligned}$$

Now, we simply add all of these individual χ^2 statistics to get the overall statistic.

$$\begin{aligned}\chi^2 &= \sum_{everycell} \frac{(O-E)^2}{E} \\ &= 4.47 + 11.17 + 14.73 + 36.81 \\ &= 67.18\end{aligned}$$

Now, we can compare this to a χ^2 distribution with one degree of freedom, which looks as follows



The probability of finding our test statistic (≈ 67) if the variables were independent is approximately zero. Now, we can have Stata tell us this directly, which we'll do in class, but it's worth knowing what's going on in the background:

```
. tab mid4 dembin, chi2

           |        dembin
      mid4 |         0          1 |     Total
-----------+----------------------+----------
         0 |    10,698      3,979 |    14,677
         1 |     2,965      1,490 |     4,455
-----------+----------------------+----------
     Total |    13,663      5,469 |    19,132

          Pearson chi2(1) =  67.1895   Pr = 0.000
```