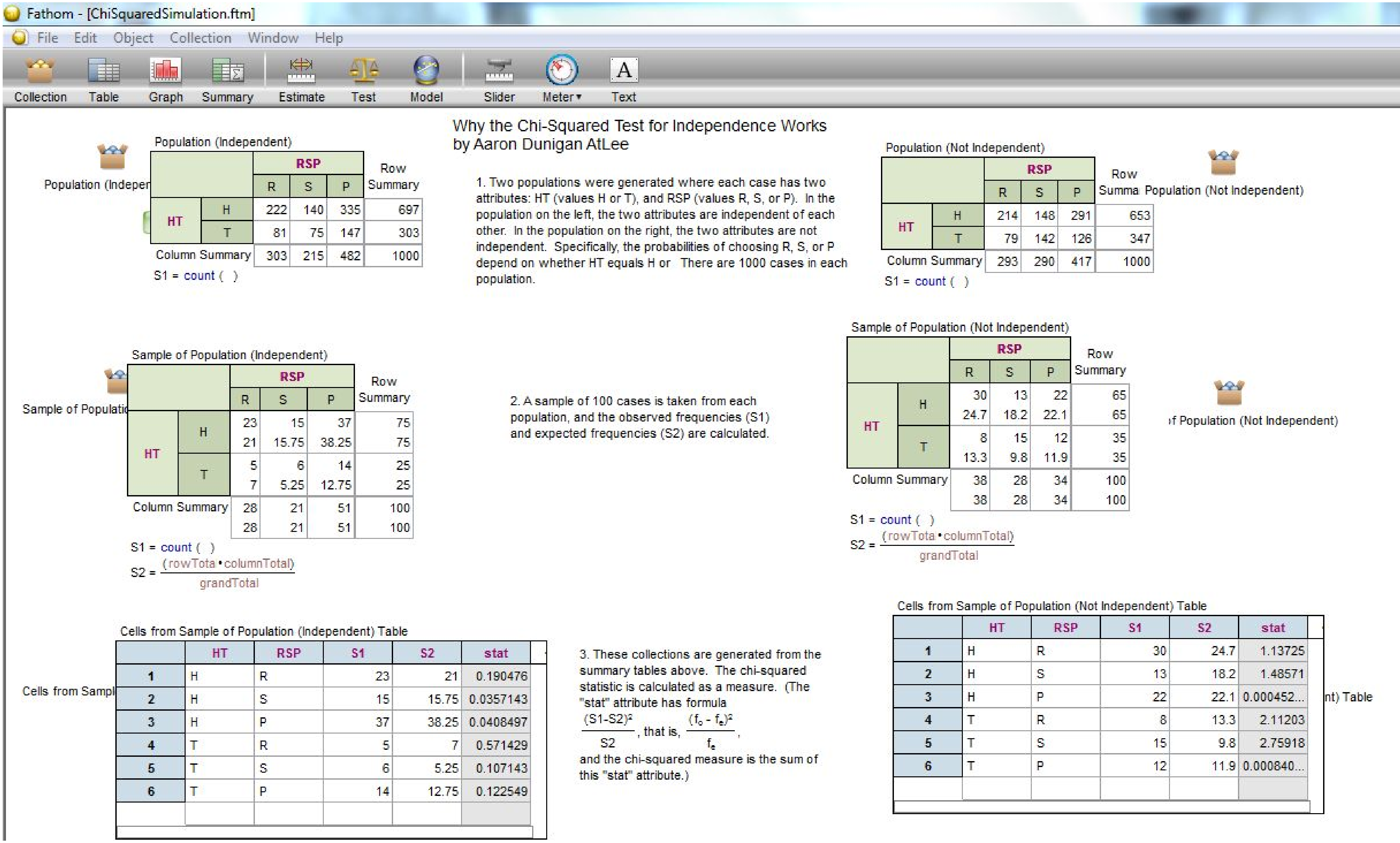


Why the Chi-Squared Test for Independence Works
by Aaron Dunigan AtLee

1. Two populations were generated where each case has two attributes: HT (values H or T), and RSP (values R, S, or P). In the population on the left, the two attributes are independent of each other. In the population on the right, the two attributes are not independent. Specifically, the probabilities of choosing R, S, or P depend on whether HT equals H or There are 1000 cases in each population.

Population (Independent)

		RSP			Row Summary
		R	S	P	
HT	H	222	140	335	697
	T	81	75	147	303
Column Summary		303	215	482	1000

S1 = count ( )

Population (Not Independent)

		RSP			Row Summa
		R	S	P	
HT	H	214	148	291	653
	T	79	142	126	347
Column Summary		293	290	417	1000

S1 = count ( )

2. A sample of 100 cases is taken from each population, and the observed frequencies (S1) and expected frequencies (S2) are calculated.

Sample of Population (Independent)

		RSP			Row Summary
		R	S	P	
HT	H	23 21	15 15.75	37 38.25	75 75
	T	5 7	6 5.25	14 12.75	25 25
Column Summary		28 28	21 21	51 51	100 100

S1 = count ( )

$S2 = \frac{(\text{rowTota} \cdot \text{columnTotal})}{\text{grandTotal}}$

Sample of Population (Not Independent)

		RSP			Row Summary
		R	S	P	
HT	H	30 24.7	13 18.2	22 22.1	65 65
	T	8 13.3	15 9.8	12 11.9	35 35
Column Summary		38 38	28 28	34 34	100 100

S1 = count ( )

$S2 = \frac{(\text{rowTotal} \cdot \text{columnTotal})}{\text{grandTotal}}$

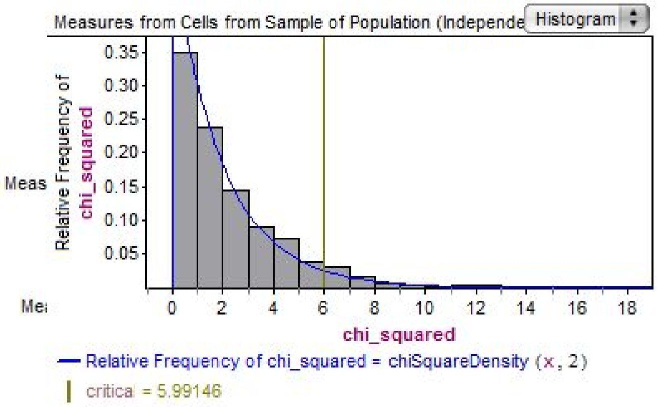
3. These collections are generated from the summary tables above. The chi-squared statistic is calculated as a measure. (The "stat" attribute has formula $\frac{(S1-S2)^2}{S2}$, that is, $\frac{(f_o - f_e)^2}{f_e}$, and the chi-squared measure is the sum of this "stat" attribute.)

Cells from Sample of Population (Independent) Table

	HT	RSP	S1	S2	stat
1	H	R	23	21	0.190476
2	H	S	15	15.75	0.0357143
3	H	P	37	38.25	0.0408497
4	T	R	5	7	0.571429
5	T	S	6	5.25	0.107143
6	T	P	14	12.75	0.122549

Cells from Sample of Population (Not Independent) Table

	HT	RSP	S1	S2	stat
1	H	R	30	24.7	1.13725
2	H	S	13	18.2	1.48571
3	H	P	22	22.1	0.000452...
4	T	R	8	13.3	2.11203
5	T	S	15	9.8	2.75918
6	T	P	12	11.9	0.000840...



Measures from Cells from Sample of Populatio...

	exceeds_crit		Row Summary
	true	false	
chi_squared	0.07	0.93	1

S1 = rowProportion

4. The sampling process was repeated 1000 times on each population (1000 samples of 100 cases each). For each sample, the chi-squared measure was recorded. On the left (independent), you can see that the distribution of chi-squared values from the samples matches the theoretical chi-square distribution (blue curve).

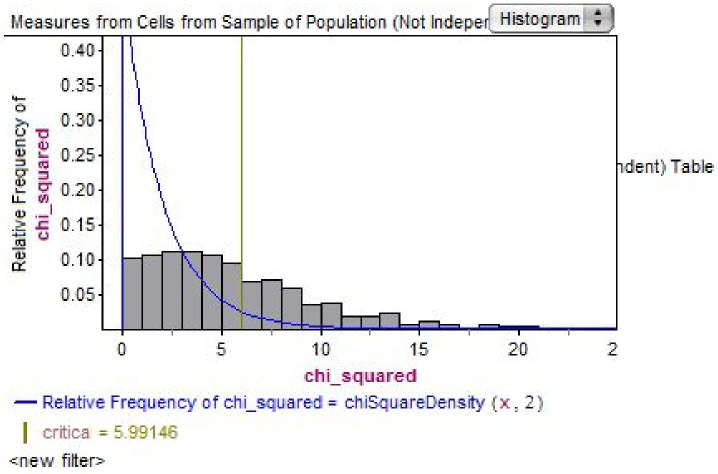
On the right (not independent), the distribution of sample chi-squared measures varies greatly from the theoretical distribution.

The gold vertical line shows the critical value (at the level of significance shown on the slider below). You can see that at 5% signficance, only about 5% of the samples (in the case of my simulation, 7%) from the independent population exceed the critical value. On the other hand, 37% of the samples of the non-independent population had chi-squared values that exceeded the critical value. (This proportion will vary depending on the nature of the inter-dependence of variables.)

What this shows is two things:
1. If the population exhibits independence, the probability of a random sample exceeding the critical value is only about 5%. So if the critical value is exceeded, there is a good chance that it came from a population that does not exhibit independence.

2. On the other hand, if there isn't independence in the original population, the chance of a chi-squared measure *below* the critical value is still quite high; thus, failing to exceed the critical value is not a good indicator that we actually have independence. Rather, it means we don't have enough evidence either way. This is why we say that we "fail to reject H_0," rather than saying that we "prove H_0."

▶ significance = 0.05 ▶ critical = 5.99



Measures from Cells from Sample of Population (No...

	exceeds_crit		Row Summary
	true	false	
chi_squared	0.37	0.63	1

S1 = rowProportion