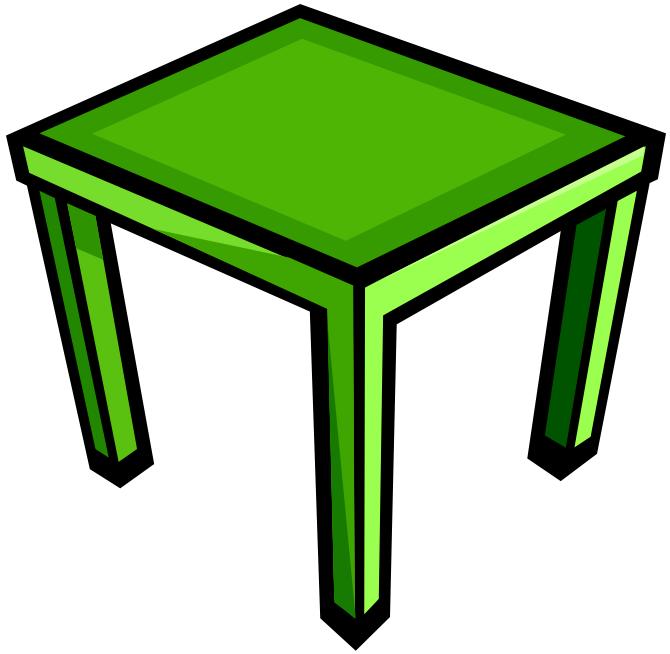
CONTINGENCY TABLES



Suppose we have two categorical (nonnumerical) variables such as gender (male/female) and political party (Republican/Democrat), and we want to determine if there is a relationship between these variables. Our null hypothesis, of course, will assume that there is no relationship, and we will set our level of significance to *alpha = .01*.

In particular, suppose we have a sample of 100 people, and the breakdown is as shown in the table below.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

The question now is if there is no relationship between gender and political party, then what numbers should we expect to see?

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Fortunately, we can answer that question.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

For example, 40% of our sample is male, and hence, we should expect 40% of the 70 democrats to be male.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

In other words, our expectation for the number of male democrats is [(70)(40)]/100.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Notice that this amounts to a column total times a row total divided by the grand total.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

$$\frac{70 \cdot 40}{100} = \frac{\text{column total} \times \text{row total}}{\text{grand total}}$$

We can find all of our expected values this way.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	(70*40)/100	(30*40)/100	40
F	(70*60)/100	(30*60)/100	60
	70	30	100

Or in other words,

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

A requirement now for continuing is that all the *expected values* be greater than or equal to 5. This requirement is met.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

The idea now is that if the difference between the observed values and the expected values is too large, then there is a significant difference.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

In particular, for each cell we will look at *observed minus expected*, square the difference to get rid of negatives, and then express the result as a fraction of the *expected value*.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

$$\frac{(O-E)^2}{E}$$

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

Next, we add it all up.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

$$\sum \frac{(O-E)^2}{E} = \frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} + \frac{(50-42)^2}{42} + \frac{(10-18)^2}{18}$$

$$= 12.6984127$$

The distribution that is used with this test is called the *chi-squared distribution*.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

$$\sum \frac{(O-E)^2}{E} = \frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} + \frac{(50-42)^2}{42} + \frac{(10-18)^2}{18}$$

$$= 12.6984127$$

The typical *chi-squared distribution* is asymmetrial, and *degrees of freedom = (rows -1) x (columns – 1)*.

Observed

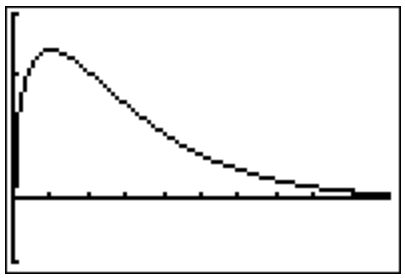
	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

$$\sum \frac{(O-E)^2}{E} = \frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} + \frac{(50-42)^2}{42} + \frac{(10-18)^2}{18}$$

$$= 12.6984127$$



Also, our test will be a *one-tail test* since we only reject the *null hypothesis* if the differences are too large.

Observed

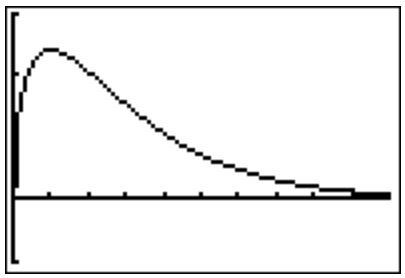
	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

$$\sum \frac{(O-E)^2}{E} = \frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} + \frac{(50-42)^2}{42} + \frac{(10-18)^2}{18}$$

$$= 12.6984127$$



To determine if this calculated value is too large we will use the *CHI-SQUARE TEST* on our calculator.

Observed

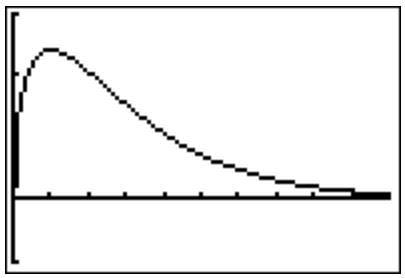
	D	R	
M	20	20	40
F	50	10	60
	70	30	100

Expected

	D	R	
M	28	12	40
F	42	18	60
	70	30	100

$$\sum \frac{(O-E)^2}{E} = \frac{(20-28)^2}{28} + \frac{(20-12)^2}{12} + \frac{(50-42)^2}{42} + \frac{(10-18)^2}{18}$$

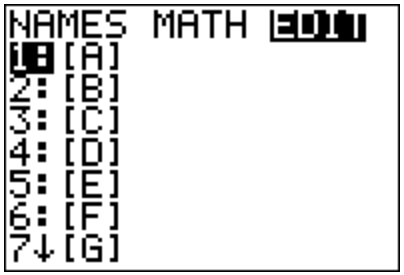
$$= 12.6984127$$

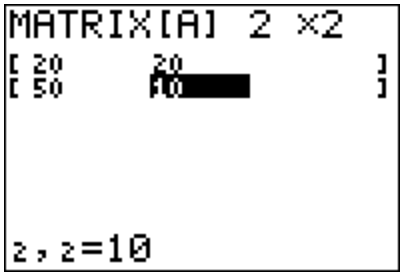


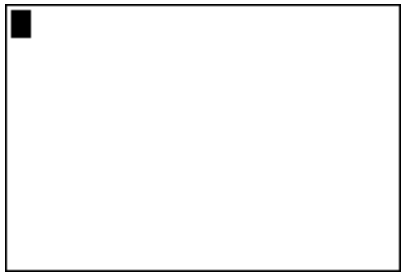
First, enter the observed values into a matrix in your calculator, and press *2nd QUIT* to return to the home screen.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100



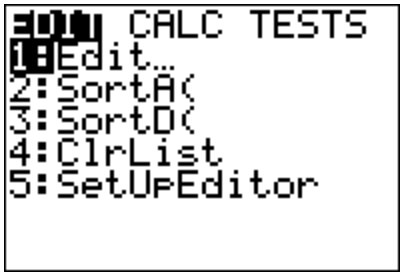


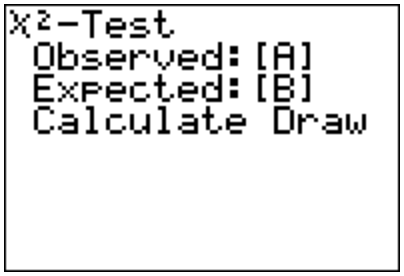


Now, go to *STAT* and then *TESTS*, and select the *CHI-SQUARE TEST*.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100

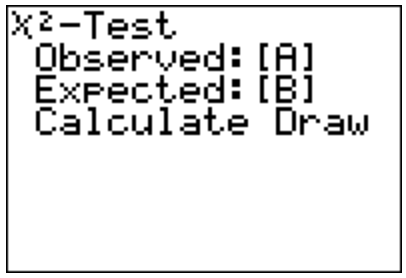


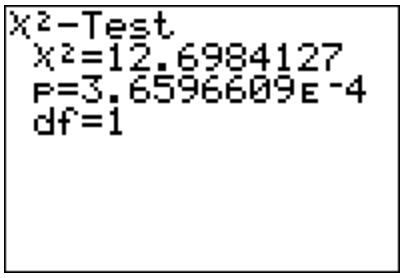


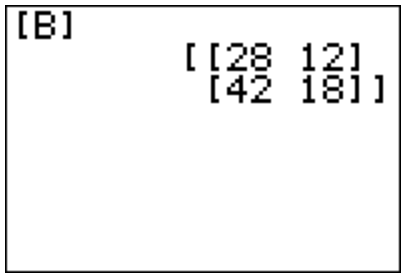
Your observed values should be stored in matrix *A*. Scroll down to calculate and hit *ENTER*, and your expected values will automatically be computed and stored in matrix
B.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100



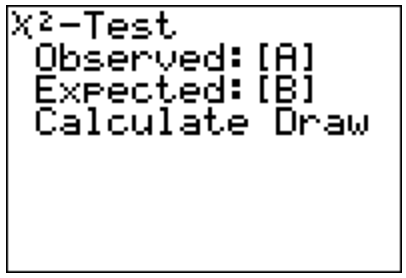


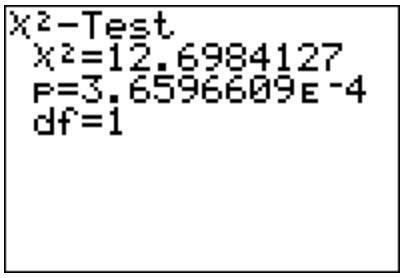


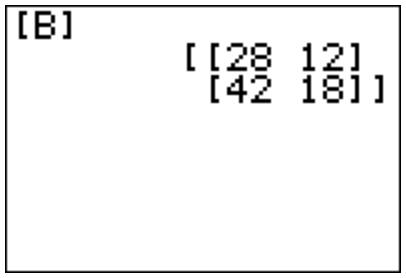
The low *P-value* (less than .01) indicates that we must reject the *null hypothesis*. In other words, there is evidence of a relationhip between the variables *gender* and *political party*.

Observed

	D	R	
M	20	20	40
F	50	10	60
	70	30	100







And here is our more formal hypothesis test.

H_0 : *Gender* and *Political Party* are independent

H_1 : *Gender* and *Political Party* are dependent

Test Statistic Formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

Degrees of Freedom: $df = 1$

Level of Significance: $\alpha = .01$

P-Value: $P = 0.0004$

Decision: Reject H_0

Sometimes we have just a single row or column, and we want to know if our frequencies match what is expected. In other words, we want to test for *goodness-of-fit*. This test is available on the TI-84 calculator, but not the TI-83. Thus, we'll just give you a demonstration of how it works. We'll use the same distribution and test statistic formula as before, and we will again require that the *expected values* be at least five.

Suppose we have a sample of numerical data of size 1000, and we suspect that the data is normally distributed. If this is so, then we expect there to be certain percentages present based on how many standard deviations we are from the mean.

INTERVAL	PERCENTAGE
more than 2 standard deviations below the mean	2%
1 to 2 standard devations below the mean	14%
1 or fewer standard deviations below the mean	34%
1 or fewer standard deviations above the mean	34%
1 to 2 standard deviations above the mean	14%
more than 2 standard deviations above the mean	2%

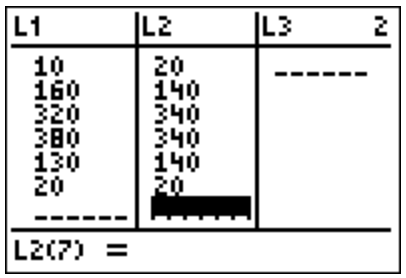
Now suppose our data is distributed as below.

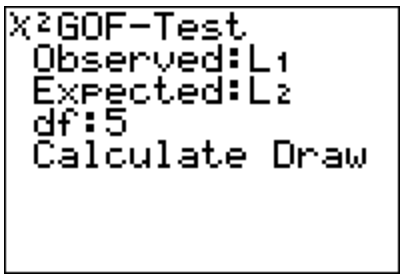
DEVIATIONS FROM THE MEAN	OBSERVED FREQUENCY	EXPECTED FREQUENCY
> 2 below	10	20
1 to 2 below	160	140
< 1 below	320	340
< 1 above	380	340
1 to 2 above	130	140
> 2 above	20	20

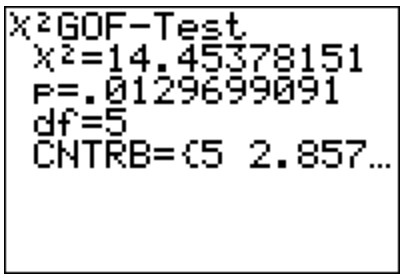
In our TI-84 calculator we enter our observed frequencies into *List 1*, our expected frequencies into *List 2*, and we select the *CHI SQUARE GOF TEST*. Also, since we have 6 categories, we have 5 degrees of freedom. Additionally, we will use *alpha = .01.*

DEVIATIONS FROM THE MEAN	OBSERVED FREQUENCY	EXPECTED FREQUENCY
> 2 below	10	20
1 to 2 below	160	140
< 1 below	320	340
< 1 above	380	340
1 to 2 above	130	140
> 2 above	20	20

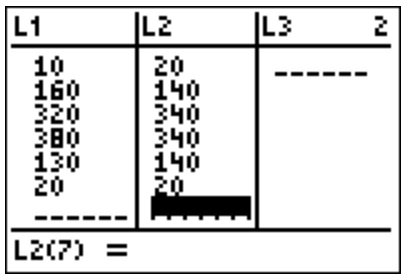
And here is the result.

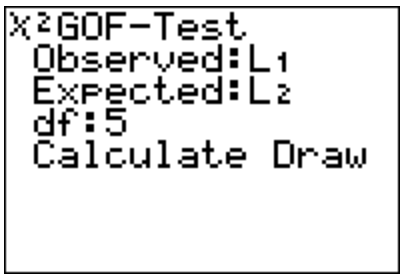


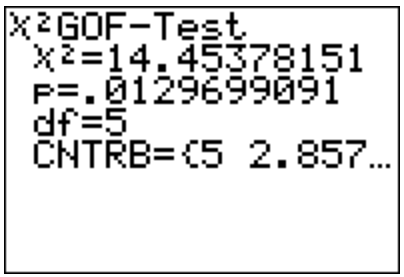




At the .01 level of significance, we cannot reject the *null hypothesis*.







And here is our more formal analysis.

H_0 : The data fit a normal distribution

H_1 : The data do not fit a normal distribution

Test Statistic Formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

Degrees of Freedom: $df = 5$

Level of Significance: $\alpha = .01$

P-Value: $P = 0.0130$

Decision: Fail to reject H_0