Chi-Squared Test

Chi-Squared Test

The Chi-Squared test is used to determine if your data is consistent with an assumed distribution. It is used to test

- Whether a die is fair (each number has equal probability)
- Whether a distribution is Normal (vs. Poisson or geometric)

For example, suppose you want to test whether the following Matlab code generates a uniform distribution

```
d6 = ceil(rand*6);
```

To run a Chi-Squared test to see if this really is a fair die, do the following:

- Divide the results into M bins (6 bins in this case: numbers 0 .. 5)
- Collect n data points.
- Count how many times the data fell into each of the M bins
- Compute the Chi-Squared total for each bin as

$$\chi^2 = \left(\frac{(np-N)^2}{np} \right)$$

  - *np is the expected number of times data should fall into each bin*
  - *N is the actual number of times data fell into each bin*
- Use a Chi-Squared table to convert the resulting Chi-Squared score to a probability. Note that the degrees of freedom is equal to the number of bins minus one.

Example: n = 120 die rolls

```
RESULT = zeros(1,6);

for i=1:120
   d6 = ceil(rand*6);
   RESULT(d6) = RESULT(d6) + 1;
end

RESULT =  18    27    25    19    14    17
```

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	20	18	0.2
2	1/6	20	27	2.45
3	1/6	20	25	1.25
4	1/6	20	19	0.05
5	1/6	20	14	1.8
6	1/6	20	17	0.45
			Total:	**6.2**

From a Chi-Squared table with 5 degrees of freedom, a Chi-Squared total of 6.2 corresponds to a probability of about 1%. This tells you that, based upon this data, there is only a 1% chance that the die is loaded. The code we've been using appears to generate a fair die.

Chi-Squared Table

Probability of rejecting the null hypothesis

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41

Chi-Squared table. With 5 degrees of freedom (df) and $\chi^2 = 6.2$, the probability is between 90% and 10%

You can also go to StatTrek.com to get the same result (www.StatTrek.com)

- Enter a value for degrees of freedom.
- Enter a value for one, and only one, of the remaining unshaded text boxes.
- Click the **Calculate** button to compute values for the other text boxes.

Degrees of freedom	5
Chi-square critical value (CV)	6.2
$P(X^2 < 6.2)$	0.71
$P(X^2 > 6.2)$	0.29

The probability the die is loaded is 0.71 (www.StatTrek.com)

This might seem like a problem with the *rand* function in Matlab. It's not - it just reflects chance with rolling a small number of dice.

If you repeat with 1,000,000 rolls of the dice:

```
RESULT = zeros(1,6);

for i=1:1e6
   d6 = ceil(rand*6);
   RESULT(d6) = RESULT(d6) + 1;
end

RESULT =     166220     166399     166933     167052     166500     166896

Chi2 = sum( (RESULT - 166666.666).^2) / 166666.666

Chi2 =    3.4257
```

There is only a 37% chance the die is loaded (this is a good result: too small tells you that the data is fudged - stay tuned for this....)

Example 2: Loaded Die

Suppose instead you had a loaded die:

- 90% of the time, the die is fair (all results have equal probability)
- 10% of the time, the result is always a 6.

Can you detect that the die is fair after 120 rolls?

Code:

```
RESULT = zeros(1,6);

for i=1:1000
    if(rand < 0.1)
        d6 = 6;
    else
       d6 = ceil(rand*6);
    end
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT =    24    18    10    17    21    30

Chi2 = sum( (RESULT - np).^2) / np

Chi2 =   11.5000
```

To check if this die is loaded, again collect data and compute the Chi-Squared score. Running this code 300 times results in the following (note: each time you do this you'll get different results. It's random)

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	20	24	0.8
2	1/6	20	18	0.2
3	1/6	20	10	5
4	1/6	20	17	0.45
5	1/6	20	21	0.05
6	1/6	20	30	5
			Total:	11.5

A Chi-Squared table allows you to convert the Chi-Squared score to a probability

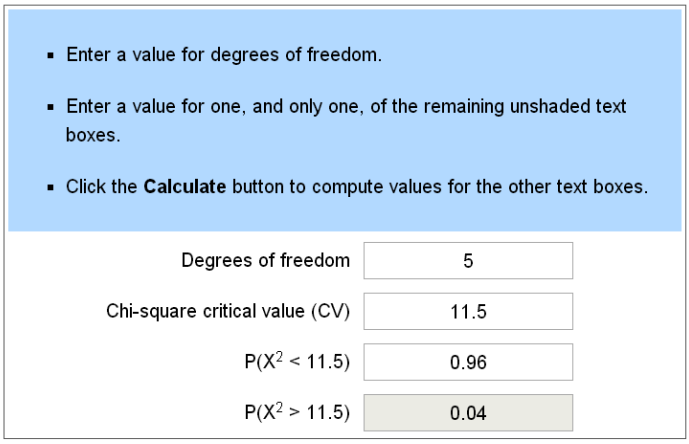
Chi-Squared Table

Probability of rejecting the null hypothesis

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41

Chi-Squared table. With 5 degrees of freedom (df) and χ^2=11.5, the probability is about 95%

Form the Chi-Squared table, the probability is between 10% and 90%. From StatTrek, it's actually 0.62



The probability the die is loaded is 0.96 (www.stattrek.com)

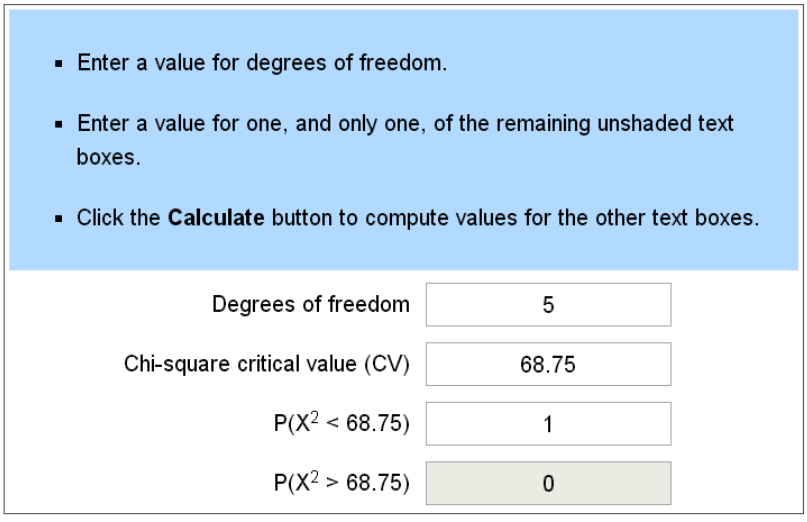
Based upon this data, there is a 95% chance that this die is loaded. That's still not enough to accuse someone of cheating - but you're suspicious.

Repeat for 1200 rolls:

```
RESULT = zeros(1,6);
N = 1200;
p = 1/6;

for i=1:N
    if(rand < 0.1)
        d6 = 6;
    else
       d6 = ceil(rand*6);
    end
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT =   186   180   160   173   197   304

Chi2 = sum( (RESULT - N*p) .^ 2 ) / (N*p)

Chi2 =   68.7500
```

Using a Chi-Squared table with 5 degrees of freedom, the odds that the die is loaded is more than p = 0.9995 (off the chart);

Loaded dice are difficulty to spot unless you have a LOT of data. By that time, you're probably broke. But then, you'll know why.



Probability the die is loaded is more than 0.9995 (rounded to 1)

Example 3: How loaded is too loaded?

Suppose you want to load a die so that there is only a 5% chance that someone will detect the die is loaded after 120 rolls.

How loaded can you make the die?

To solve, go backwards. The Chi-Squared score for 95% and 5 degrees of freedom is 11.1.

Suppose you get x too many 6's and all the other numbers come up at their expected frequency.

Set up a Chi-Squared table

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	20	20 - x/5	$\left(\frac{(x/5)^2}{20}\right)$
2	1/6	20	20 - x/5	$\left(\frac{(x/5)^2}{20}\right)$
3	1/6	20	20 - x/5	$\left(\frac{(x/5)^2}{20}\right)$
4	1/6	20	20 - x/5	$\left(\frac{(x/5)^2}{20}\right)$
5	1/6	20	20 - x/5	$\left(\frac{(x/5)^2}{20}\right)$
6	1/6	20	20 + x	$\left(\frac{x^2}{20}\right)$
			Total:	$\left(\frac{1.2x^2}{20}\right) = 11.5$

From this, you can get away with an extra 13.84 sixes with the customer only being 95% certain he/she is being cheated.

The loading is then

$$\left(\frac{13.84}{120}\right) = 0.115$$

You can load the die so that 11.5% of the time you always get a six.

Note:

- If you get too greedy, the customer will notice.
- It's hard to tell if a die is loaded unless you make lots and lots of rolls.

Example 4: Fudging Data

Chi-Squared tests can also detect if data was fudged. Take for example a Chi-Squared table for 1-5 degrees of freedom.

- If the Chi-Squared score is too large (say, 16.75 for 5 degrees of freedom), you can be 99.5% certain that the data does not come from the distribution assumed (you are 99.5% certain you can reject the null hypothesis.)
- If the chi-squared score is a moderate value (0.55 to 15.09), you cannot accept or reject the null hypothesis.
- If, however, the Chi-Squared score is too good (less than 0.41), you can be suspicious that the data was forged. It is possible you got really good data - sometimes you get lucky. If this happens over and over again, you can be almost certain that the data was faked.
-

Chi-Squared Table

Probability of rejecting the null hypothesis

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41

For example, suppose instead of rolling a fair die 1000 times I

- Only roll a die 100 times, and then
- Add 150 to the sum of each die roll so that it *looks* like I rolled the dice 1000 times.

You can spot this fake data by the data fitting the null hypothesis *too* well.

In Matlab:

```
RESULT = 150 * ones(1,6);

for i=1:100
    d6 = ceil(rand*6);
    RESULT(d6) = RESULT(d6) + 1;
end

RESULT

sum( ( (RESULT - 166.666).^2) / 166.666)
```

The chi-sqared table then looks like:

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	166.67	168	0.01
2	1/6	166.67	168	0.01
3	1/6	166.67	159	0.35
4	1/6	166.67	169	0.03
5	1/6	166.67	169	0.03
6	1/6	166.67	167	0
			Total:	0.44

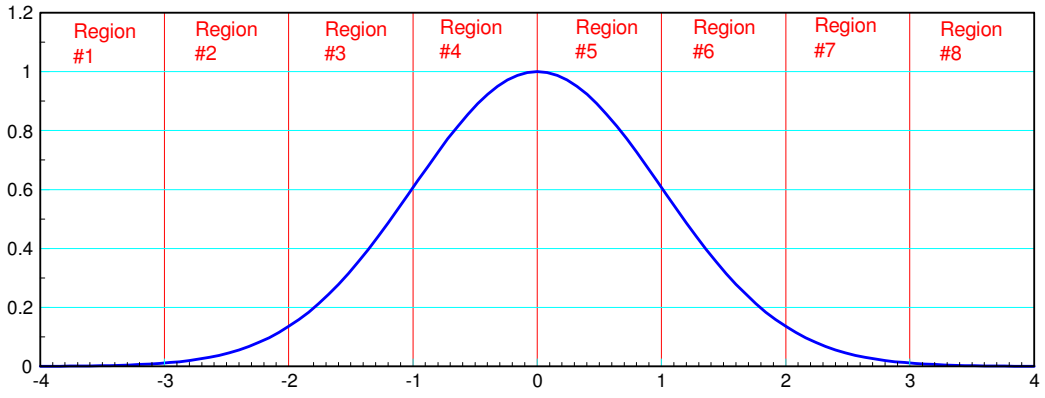
A chi-squared score of 0.44 corresponds to a probability of 0.5%

- The odds against getting such good data are 200 : 1 against.
- Most likely the data was faked.

Chi-Squared with Continuous Distributions

You can also use Chi-Squared tables with continuous distributions. In this case, the procedure is to

- Split the continuous variable into N distinct regions (many ways to do this)
- Calculate the probability that any given data point will fall into each region,
- Calculate the expected number of observations you should have in each region,
- Compare the expected number of observations (np) to the actual number of observations (N) in each region (i.e. the chi-squared scored), then
- Convert the chi-squared score into a probability.



To do a Chi-Squared test with a continuous function, split the data into X regions. Then proceed as before:
Compare the expected unmber of times data lands in each region (np) vs. the actual number of times it does (N)

For example, in the lecture on the central limit theorem, an approximation for a standard normal distribution is

```
X = sum( rand(12,1) ) - 6
```

This is how some computers generate standard normal random variables:

- You take a uniform distribution over the interval of (0, 1) (mean = 1/2, variance = 1/12)
- Add twelve of these together (mean = 6, variance = 1)
- Subtract six (mean = 0, variance = 1)

This looks like a standard normal variable. Is it?

The answer of course is no - I can see the code.

Can you detect that it is not a standard normal variable?

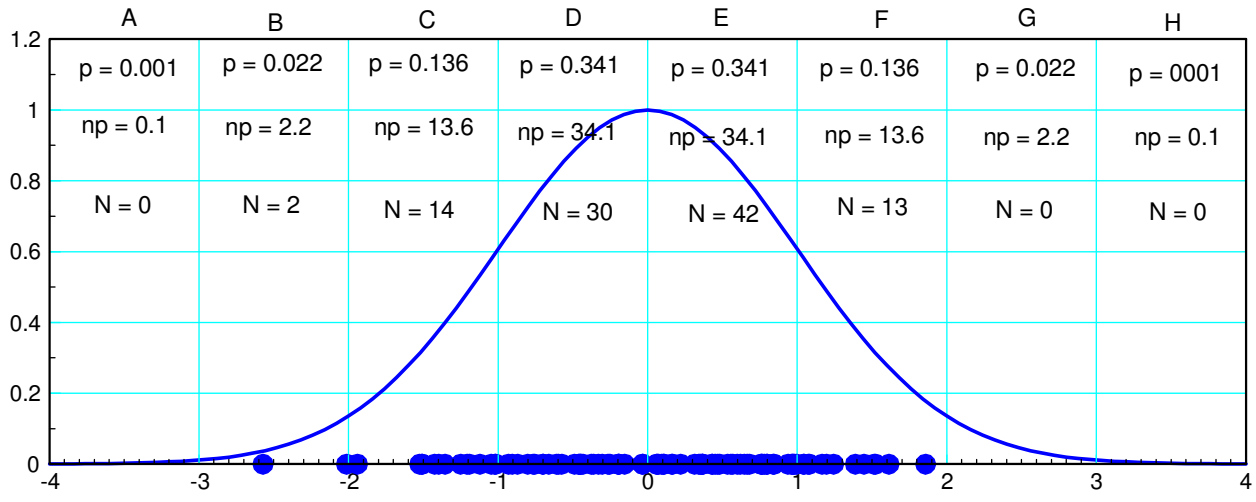
That is harder to do

Example: Generate 100 random numbers

```
X = [];

for i=1:100
   X = [X ; sum( rand(12,1) ) - 6];
   end
```

To do a Chi-Squared test,

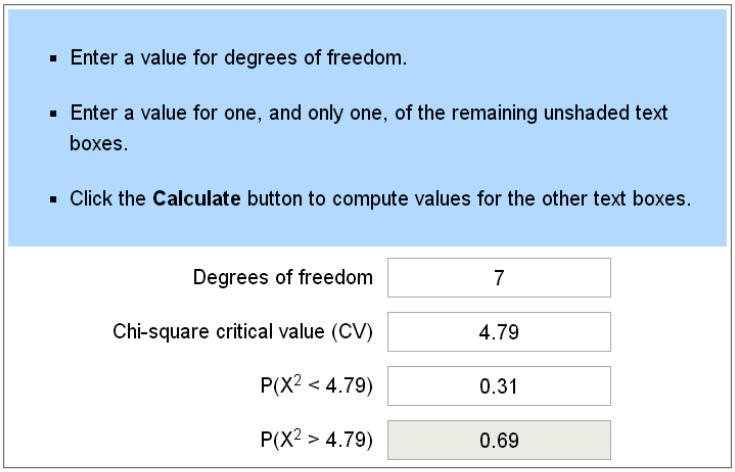
- Split the X axis into 8 regions (A..H) (this is somewhat arbitrary).
- Compute the probability of each region (p) and the expected frequency (np)
- Count how many times X fell into each region (N)
- From this, create a Chi-Squared table



Normal curve split into eight regions (A..H) along with the expected and actual frequency of each region

Region (bin)	p	np	N	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
A	0.001	0.1	0	0.1
B	0.022	2.2	2	0.0182
C	0.138	13.8	14	0.0029
D	0.341	34.1	30	0.493
E	0.341	34.1	42	1.8302
F	0.138	13.8	13	0.0464
G	0.022	2.2	0	2.2
H	0.001	0.1	0	0.1
			Total:	**4.7906**

A Chi-Squared table converts this number to a probability



With 100 random numbers, you cannot conclude that this is not a standard normal distribution.

Repeat with 100,000 random numbers.

```
BIN = zeros(20,1);

for i=1:1e5
   X = sum( rand(12,1) ) - 6;
   BIN(floor(X) + 10) = BIN(floor(X) + 10) + 1;
   end

N = [1:20]' - 10;
```

Region (bin)	p	np	N	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
A	0.0013	132	97	**9.2803**
B	0.0214	2,140	2,085	1.4136
C	0.1359	13,591	13,751	1.8836
D	0.3413	34,134	34,067	0.1315
E	0.3413	34,134	33,845	2.4469
F	0.1359	13,591	13,895	6.7998
G	0.0214	2,140	2,168	0.3664
H	0.0013	132	88	**14.6667**
			Total:	**36.9886**

With 100,000 random numbers, you can tell that this isn't a standard normal curve. Note that it's the tails that tell you this: there just aren't enough numbers that are more than 3 standard deviations away from the mean.

Chi-Squared Table

Probability of rejecting the null hypothesis
http://people.richland.edu/james/lecture/m170/tbl-chi.html

df	99.5%	99%	97.5%	95%	90%	10%	5%	2.5%	1%	0.5%
1	7.88	6.64	5.02	3.84	2.71	0.02	0	0	0	0
2	10.6	9.21	7.38	5.99	4.61	0.21	0.1	0.05	0.02	0.01
3	12.84	11.35	9.35	7.82	6.25	0.58	0.35	0.22	0.12	0.07
4	14.86	13.28	11.14	9.49	7.78	1.06	0.71	0.48	0.3	0.21
5	16.75	15.09	12.83	11.07	9.24	1.61	1.15	0.83	0.55	0.41
6	18.55	16.81	14.45	12.59	10.65	2.2	1.64	1.24	0.87	0.68
7	20.28	18.48	16.01	14.07	12.02	2.83	2.17	1.69	1.24	0.99
8	21.96	20.09	17.54	15.51	13.36	3.49	2.73	2.18	1.65	1.34
9	23.59	21.67	19.02	16.92	14.68	4.17	3.33	2.7	2.09	1.74
10	25.19	23.21	20.48	18.31	15.99	4.87	3.94	3.25	2.56	2.16
11	26.76	24.73	21.92	19.68	17.28	5.58	4.58	3.82	3.05	2.6
12	28.3	26.22	23.34	21.03	18.55	6.3	5.23	4.4	3.57	3.07
13	29.82	27.69	24.74	22.36	19.81	7.04	5.89	5.01	4.11	3.57
14	31.32	29.14	26.12	23.69	21.06	7.79	6.57	5.63	4.66	4.08
15	32.8	30.58	27.49	25	22.31	8.55	7.26	6.26	5.23	4.6
16	34.27	32	28.85	26.3	23.54	9.31	7.96	6.91	5.81	5.14
17	35.72	33.41	30.19	27.59	24.77	10.09	8.67	7.56	6.41	5.7
18	37.16	34.81	31.53	28.87	25.99	10.87	9.39	8.23	7.02	6.27
19	38.58	36.19	32.85	30.14	27.2	11.65	10.12	8.91	7.63	6.84
20	40	37.57	34.17	31.41	28.41	12.44	10.85	9.59	8.26	7.43
30	53.67	50.89	46.98	43.77	40.26	20.6	18.49	16.79	14.95	13.79
40	66.77	63.69	59.34	55.76	51.81	29.05	26.51	24.43	22.16	20.71
50	79.49	76.15	71.42	67.51	63.17	37.69	34.76	32.36	29.71	27.99
60	91.95	88.38	83.3	79.08	74.4	46.46	43.19	40.48	37.49	35.53
70	104.22	100.43	95.02	90.53	85.53	55.33	51.74	48.76	45.44	43.28
80	116.32	112.33	106.63	101.88	96.58	64.28	60.39	57.15	53.54	51.17
90	128.3	124.12	118.14	113.15	107.57	73.29	69.13	65.65	61.75	59.2
100	140.17	135.81	129.56	124.34	118.5	82.36	77.93	74.22	70.07	67.33