Chi-Squared Test

ECE 376 Embedded Systems

Jake Glower - Lecture #15

Please visit Bison Academy for corresponding
lecture notes, homework sets, and solutions

Statistics

- Every time you roll a die, you get a different result.
- Every time you run an experiment, you get different results.

Statistics is a branch of mathematics which allow you to analyze such random events.

With statistics, you can answer such questions as

- Is the 6-sided die biased? (do some numbers come up too often?)
- What is the 90% confidence interval for the energy in a AA battery?
- Does a lid significantly increase the thermal resistance of a hot cup of water?

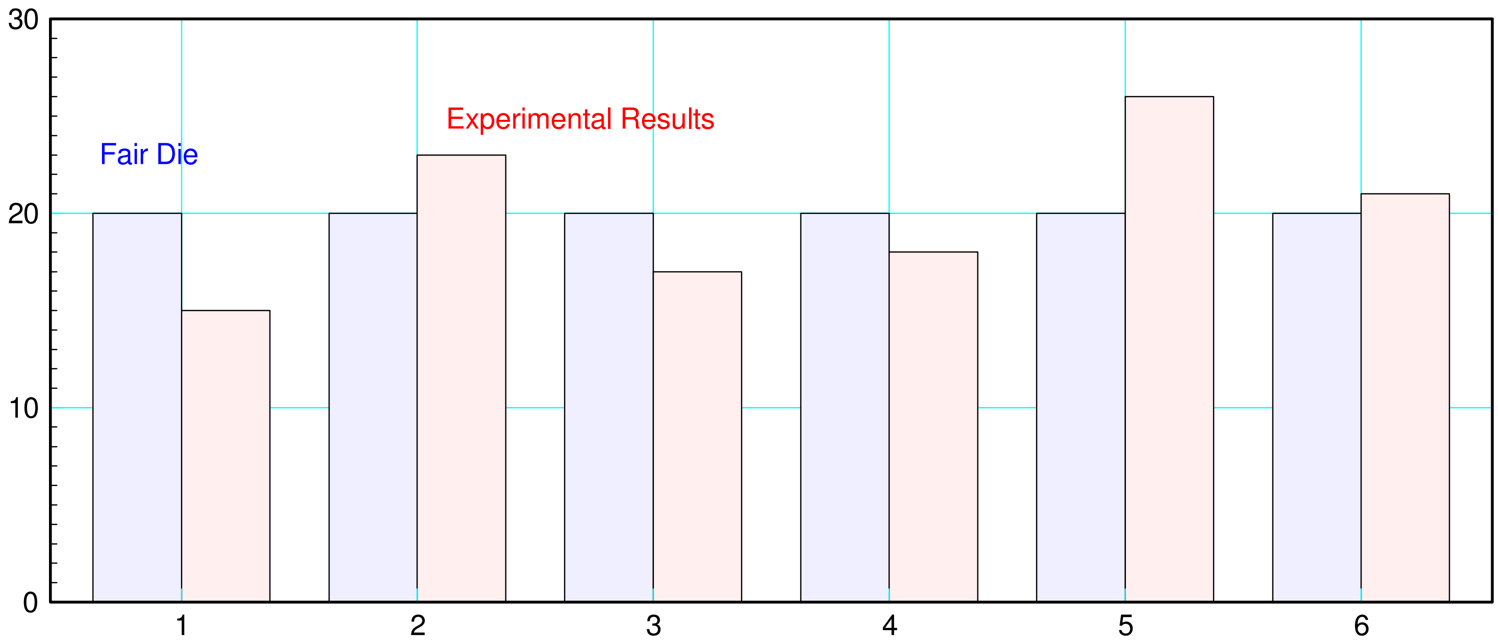
The next two lectures provide a brief overview of statistics and how to take data that we collected in our last lecture and analyze that data.

Chi-Squared Test

Is your data is consistent with an assumed distribution?

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

Example: Roll a 6-sided die 120 times



Procedure

i) Collect data.

ii) Splint the data into M bins.

- {1} {2} {3} {4} {5} {6}
- {1,2,3} {4,5} {6}

iii) Compute the Chi-Squared value

$$\chi^2 = \sum \left(\frac{(np_i - N_i)^2}{np_i} \right)$$

where

- np is the expected frequency of data falling into bin #i, and
- Ni is the actual frequency of data falling into bin #i

iv) Convert the Chi-Squared value to a probability

- Chi-Squared table (or StatTrek).

Chi-Squred Table

- The degrees of freedom are the number of bins minus one
- The number in the table is the Chi-Squred value
- The numbers on the top give you the probability of rejecting the null hypothesis

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

Interpreting the Results

Large χ^2 Score:

- The data is inconsistent with your assumed distribution
- The die is probably loaded

Small χ^2 Score:

- The data is *too* good
- The data was probably fudged

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

Example 1: Fair Die

Does this code produce a fair die?

```
while(1) {
   while(!RB0);
   while(RB0) DIE = (DIE + 1) % 6;
   DIE += 1;
   LCD_Move(1,0);  LCD_Out(DIE, 1, 0);
   SCI_Out(DIE, 1, 0);
   SCI_CRLF();
   }
```

Experiment:

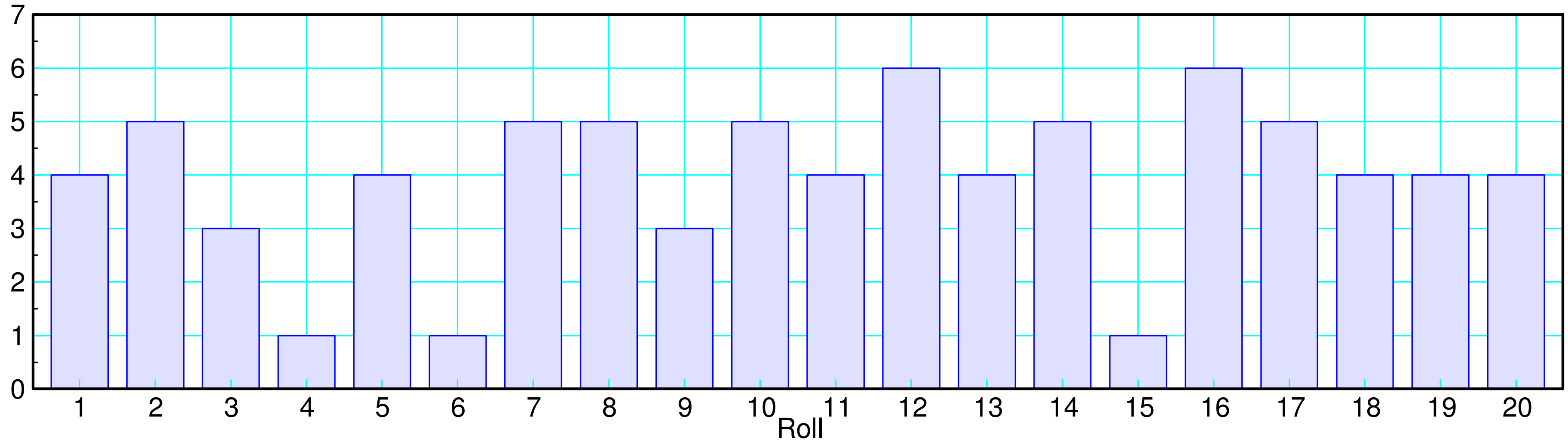
- Divide the results into M bins (6 bins in this case: numbers 1 .. 6)
- Collect n data points.
- Count how many times the data fell into each of the M bins
- Compute the Chi-Squared total for each bin as

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

  - *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 = 129 die rolls

Number	1	2	3	4	5	6
Frequency	23	16	22	24	29	15



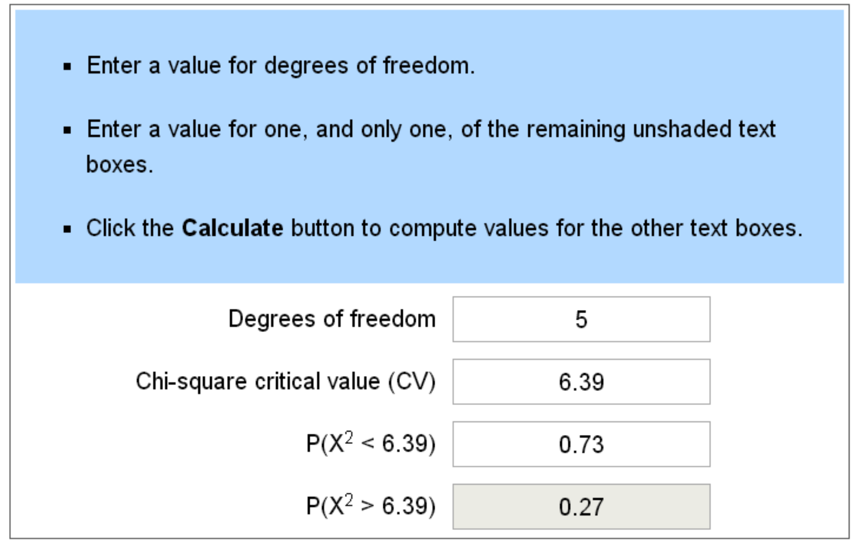
Compare expected vs. actual frequency

- Compute the χ^2 score

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	;;$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	21.5	23	0.1
2	1/6	21.5	16	1.41
3	1/6	21.5	22	0.01
4	1/6	21.5	24	0.29
5	1/6	21.5	29	2.62
6	1/6	21.5	15	1.97
			Total:	**6.4**

Convert χ^2 to a probability

- Use a Chi-Squared table
- 5 degrees of freedom (6 bins)
- $\chi^2 = 6.39$ means p = 73%
- I am 73% certain this is a loaded die
- (no conclusion)



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

Example 2: 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 100 rolls?

Code:

```
while(1) {
   while(!RB0);
   while(RB0) {
      DIE = (DIE + 1) % 6;
      X = (X+1) % 101;
      }
   if(X < 10) DIE = 6;
   else DIE += 1;
   LCD_Move(1,0);  LCD_Out(DIE, 1, 0);
   SCI_Out(DIE, 1, 0);
   SCI_CRLF();
   }
```

Roll the dice 100 times

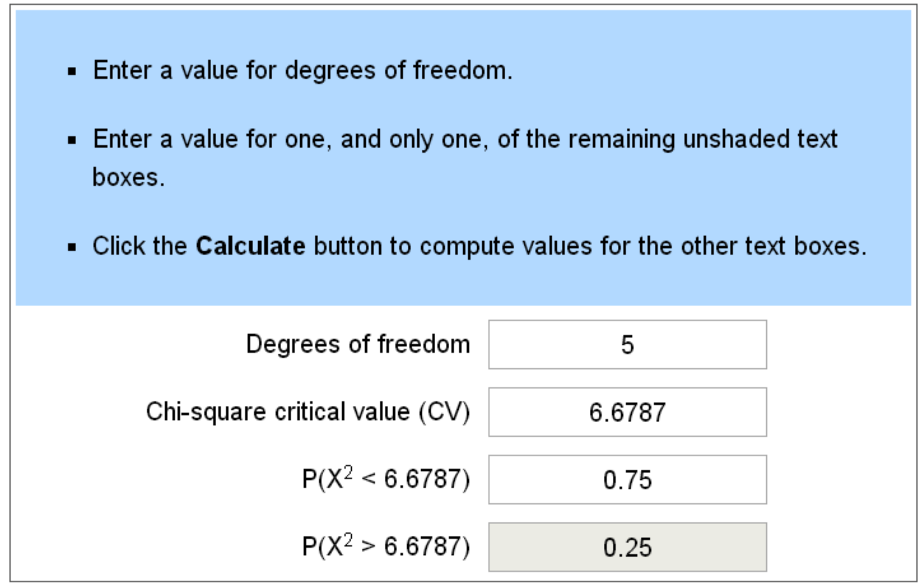
Number	1	2	3	4	5	6
Frequency	17	14	14	14	15	26

Compute the Chi-Squared value

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	16.67	17	0.01
2	1/6	16.67	14	0.43
3	1/6	16.67	14	0.43
4	1/6	16.67	14	0.43
5	1/6	16.67	15	0.17
6	1/6	16.67	26	5.22
			Total:	**6.68**

Use a Chi-Squared table (or StatTrek) to convert this back to a probability:

- p = 0.75
- I am 75% certain that this is a loaded die
- (no conclusion)



Note:

- It is hard to detect that a die is loaded with only 100 rolls

Repeat for 348 rolls:

Die Roll (bin)	p theoretical probability	np expected frequency	N actual frequency	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	58	55	0.16
2	1/6	58	55	0.16
3	1/6	58	58	0
4	1/6	58	43	3.88
5	1/6	58	57	0.02
6	1/6	58	80	8.34
			Total:	**12.55**

Now you can start to detect that the die is loaded with a probability of 97.5%:

- With enough data, you can detect that the die is loaded
- You're also probably broke at this point...

Example 3: How loaded is too loaded?

- Load a die
- 5% chance of detection after 120 rolls
- "detect" means p(loaded) = 95% ($\chi^2 = 11.1$)

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:	$\left(\frac{1.2x^2}{20}\right) = 11.5$

Result:

- You can get away with an extra 13.84 sixes

The loading is then 11.5%

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

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.
- This is what Alan Turing was referring to in the movie "The Imitation Game"
    - *Sink too many German subs and they'll know you cracked their code*
    - *Chi-squared tests tell you what "too many" means*

Example 4: Fudging Data

Chi-Squared tests can also detect if data was fudged

- If the Chi-Squared score is too large (16.75) the die is probably loaded
- If it's too small (less than 0.41), the data is probably fudged. It's too good.

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

Fudging Data Example

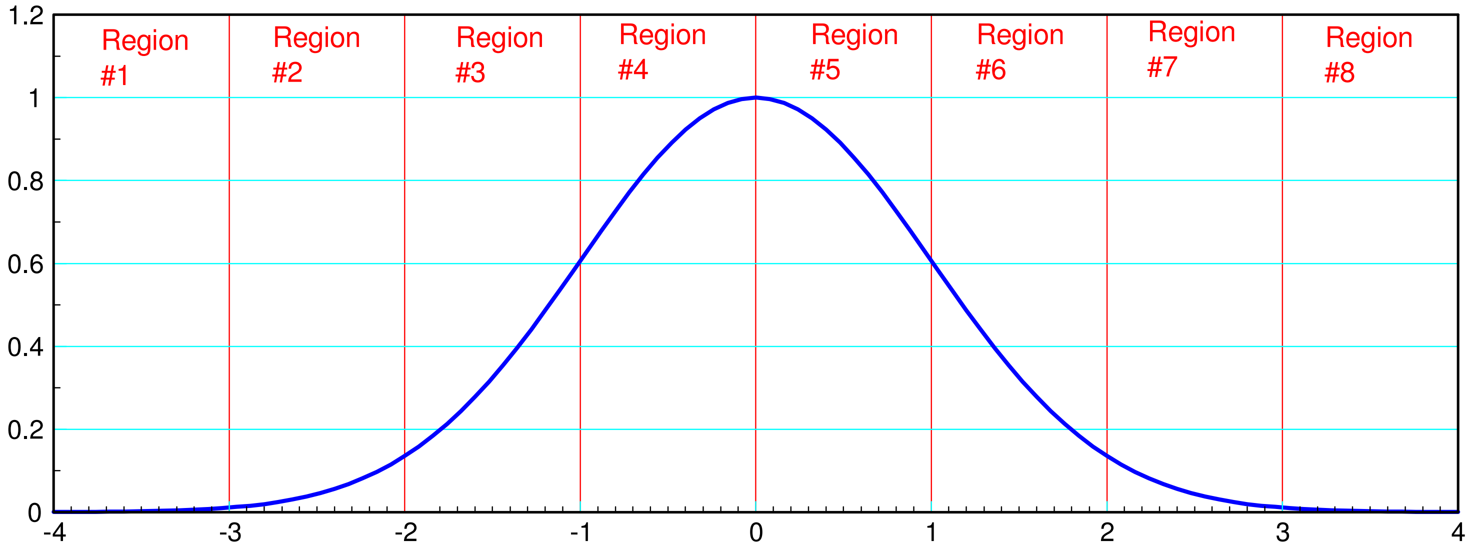
- Roll a die 129 times
- Add 200 to each result
- It *looks* like I rolled the dice 1329 times.
- $p(\chi^2 = 0.62) = 0.01$
  - *The odds against getting such good data are 100 : 1 against.*
  - *Most likely the data was faked.*

Die Roll	p	np	N	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
1	1/6	221.5	223	0.01
2	1/6	221.5	216	0.14
3	1/6	221.5	222	0
4	1/6	221.5	224	0.03
5	1/6	221.5	229	0.25
6	1/6	221.5	215	0.19
			Total:	**0.62**

Chi-Squared with Continuous Distributions

Also works with continuous distributions

- Split the continuous variable into N distinct regions / bins (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 (N)
- Convert the chi-squared score into a probability.



Example: Is this a Normal distribution?

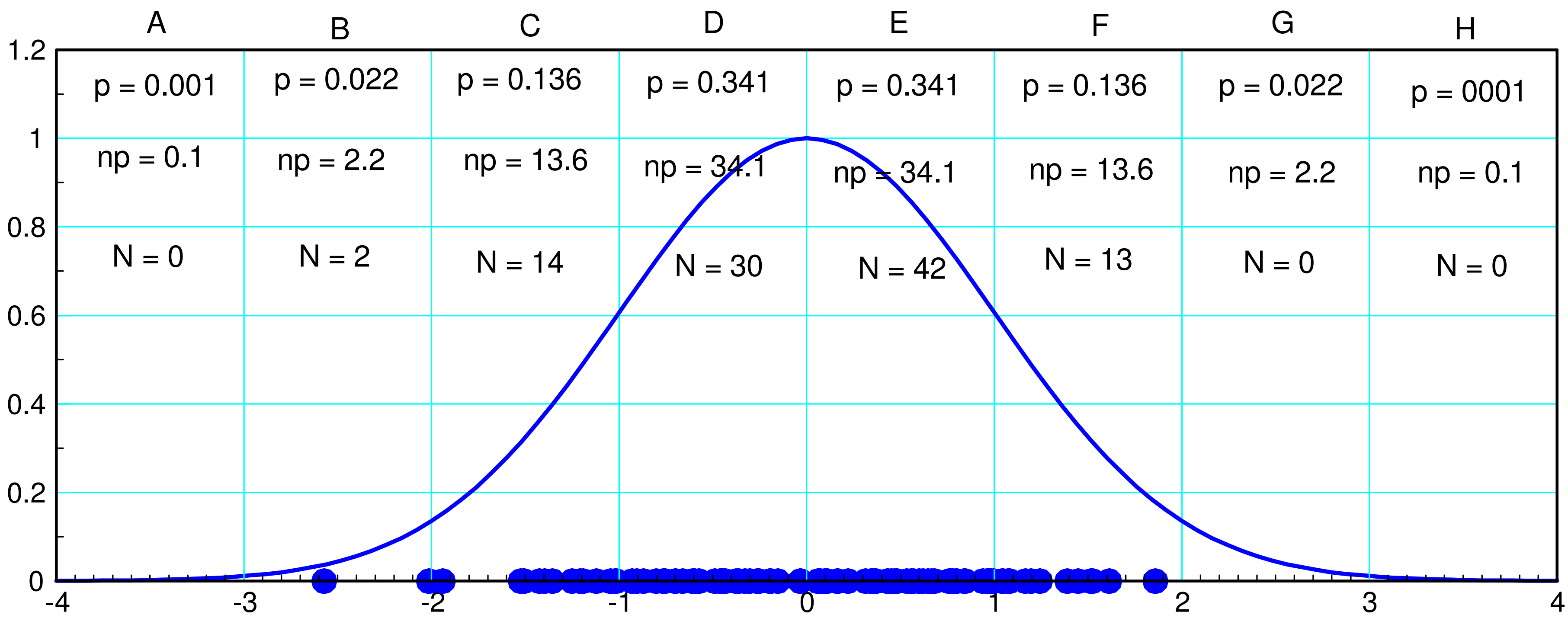
```
X = sum( rand(12,1) ) - 6
```

- Generate 100 random numbers

```
X = [];

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

- 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


A
B
C
D
E
F
G
H
p = 0.001
p = 0.022
p = 0.136
p = 0.341
p = 0.341
p = 0.136
p = 0.022
p = 0001
np = 0.1
np = 2.2
np = 13.6
np = 34.1
np = 34.1
np = 13.6
np = 2.2
np = 0.1
N = 0
N = 2
N = 14
N = 30
N = 42
N = 13
N = 0
N = 0
1.2
1
0.8
0.6
0.4
0.2
0
-4
-3
-2
-1
0
1
2
3
4

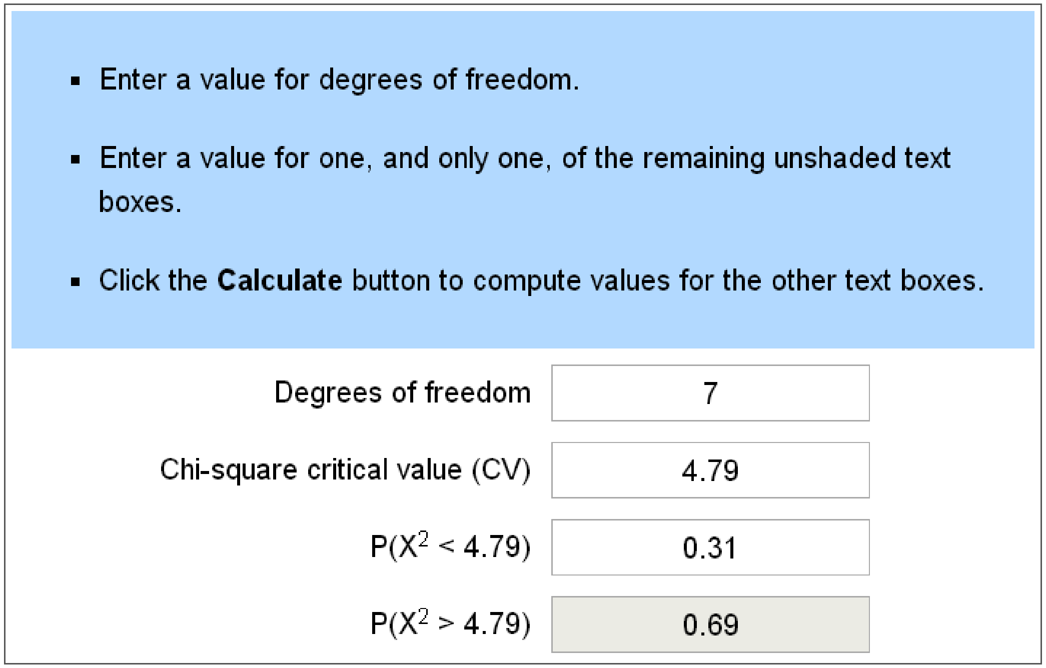
Compute the Chi-Squared score

$$\chi^2 = 4.79$$

Region (bin)	p	np	N	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
A	0	0.1	0	0.1
B	0.02	2.2	2	0.02
C	0.14	13.8	14	0
D	0.34	34.1	30	0.49
E	0.34	34.1	42	1.83
F	0.14	13.8	13	0.05
G	0.02	2.2	0	2.2
H	0	0.1	0	0.1
			Total:	4.79

Convert to a probability

- p = 0.31
- 31% chance this is not a normal distribution
- no conclusion



Note: With enough data you can detect the difference

- 100,000 numbers
- The information is in the tails

Region (bin)	p	np	N	$\chi^2 = \left(\frac{(np-N)^2}{np}\right)$
A	0	132	97	9.28
B	0.02	2,140	2,085	1.41
C	0.14	13,591	13,751	1.88
D	0.34	34,134	34,067	0.13
E	0.34	34,134	33,845	2.45
F	0.14	13,591	13,895	6.8
G	0.02	2,140	2,168	0.37
H	0	132	88	14.67
			Total:	36.99

Summary

A chi-squared test is a test of a distribution

- Is your data consistent with the assumed distribution.

With it, you can

- Detect whether a die is fair or loaded,
- Calculate how much you can "cheat" without getting caught,
- Detect if someone fudged their data,