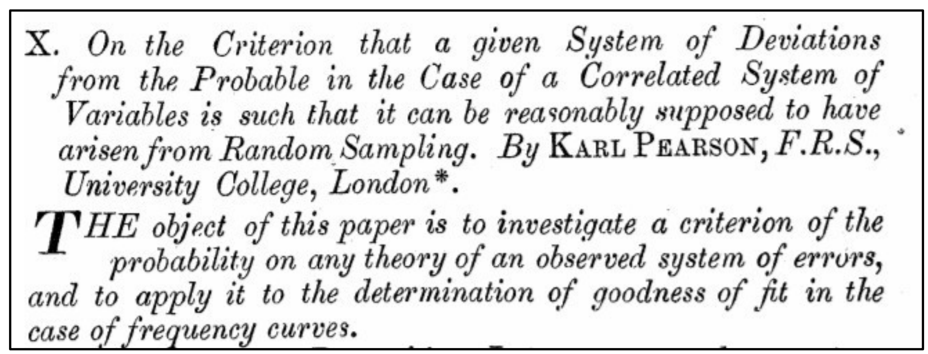
CHI-SQUARE TESTS

HISTORY

- In 1900 Karl Pearson wrote a paper
  "On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling."
- The key conceptis "goodness of fit"



X. *On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling.* By KARL PEARSON, *F.R.S., University College, London**.

THE object of this paper is to investigate a criterion of the probability on any theory of an observed system of errors, and to apply it to the determination of goodness of fit in the case of frequency curves.

Source: http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf

CHI-SQUARE OR CHI-SQUARED?

- You'll see it written both ways
- The Wikipedia article about it is titled "Pearson's Chi-squared Test"
- While Pearson does use the χ^2 notation in his paper he never assigns a term to it
- Since χ^2 is a single statistic, it's proper to use chi-square, notchi-squared

CHI-SQUARE TEST

- A Chi-square Test (also written χ^2) is used to determine the probability of an observed frequency of events given an expected frequency.

CHI-SQUARE TEST

- For example, if we flip a coin 18 times and observe that it comes up heads 12 times, can we say that this is due to chance,

  or do we assume that our coin is biased?

CHI-SQUARE TEST

- The chi-square formula considers the sum of square distances between observed values O and expected values E, divided by each expected value:

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

CHI-SQUARE TEST

- For our coin example, we had heads come up 12 times out of 18 flips, with an expected frequency of 9 heads (half of 18).
- This means that tails came up 6 times, with an expected frequency of 9 tails.

CHI-SQUARE TEST

- Our calculation becomes:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \overset{\textit{heads}}{\frac{(12 - 9)^2}{9}} + \overset{\textit{tails}}{\frac{(6 - 9)^2}{9}} = 2.0$$

CHI-SQUARE TEST

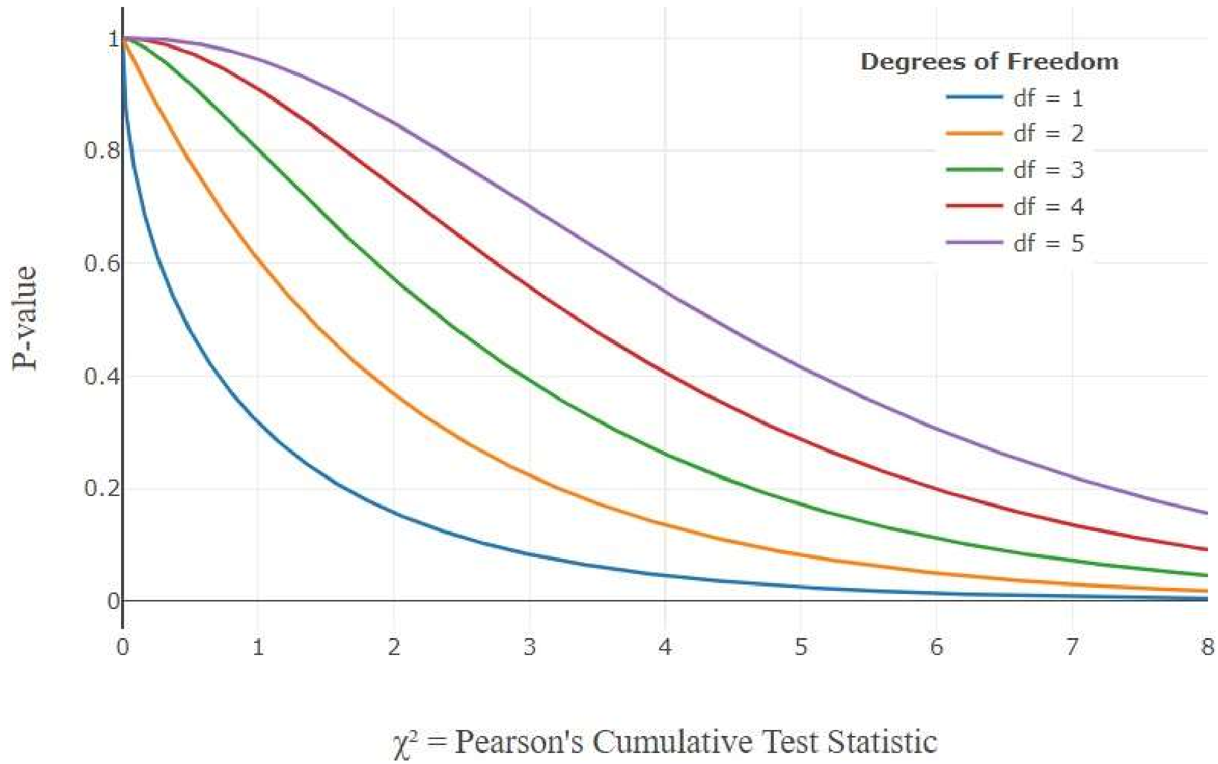
- So what does a value of 2.0 represent?

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \underbrace{\frac{(12 - 9)^2}{9}}_{} + \frac{(6 - 9)^2}{9} = 2.0$$

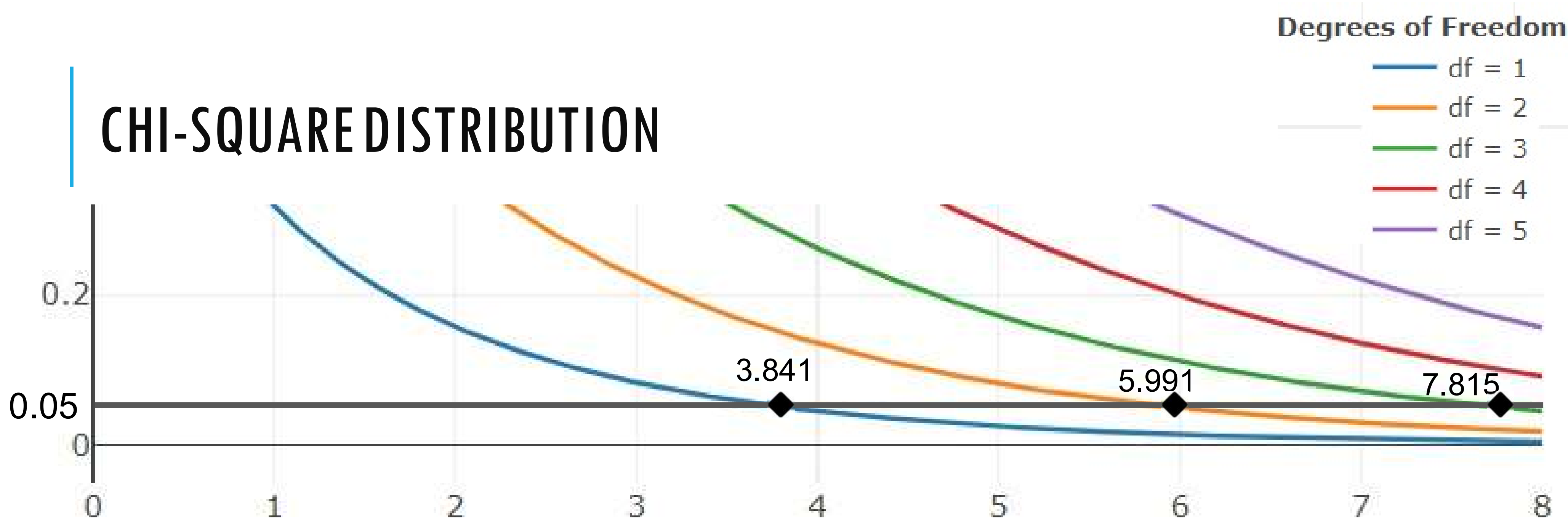
heads tails

CHI-SQUARE DISTRIBUTION

- Chi-square distribution, showing x^2 on the x-axis and P-value on the y-axis



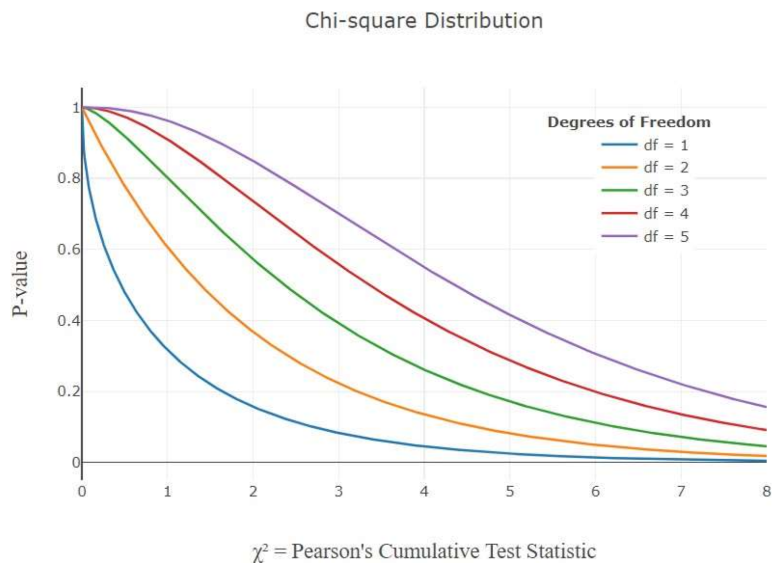
CHI-SQUARE DISTRIBUTION



Chi-square Critical Values						
df	**0.15**	**0.10**	**0.05**	**0.01**	**0.005**	**0.001**
1	2.072	2.706	3.841	6.635	7.879	10.828
2	3.794	4.605	5.991	9.210	10.597	13.816
3	5.317	6.251	7.815	11.345	12.838	16.266

CHI-SQUARE DISTRIBUTION

- A low x^2 value means a high correlation between the observed values and the expected values.



CHI-SQUARE TABLE

- Our coin example had a x^2 value of 2.0
- Degrees of freedom was (2-1) or 1
- Our critical value with 95% confidence is:

Chi-square Critical Values						
df	**0.15**	**0.10**	**0.05**	**0.01**	**0.005**	**0.001**
1	2.072	2.706	3.841	6.635	7.879	10.828
2	3.794	4.605	5.991	9.210	10.597	13.816
3	5.317	6.251	7.815	11.345	12.838	16.266

CHI-SQUARE CONCLUSIONS

- Our null hypothesis was that 12 heads in 18 flips was statistically reasonable for a fair coin, with 95% probability.
- Since $\chi^2 = 2.0$ **and** $\chi^2_{critical} = 3.841$,

$$\chi^2 < \chi^2_{critical}$$

We fail to reject the null hypothesis

CHI-SQUARE EXAMPLE

- A company runs six identical servers to support its IT infrastructure.
- Logically, the failure rate should be the same across all servers.
- Based on the following data, can we assume that the servers fail at the same rate?

CHI-SQUARE EXAMPLE

- First, let’s state some assumptions:
  1. If a server fails, it does NOT affect the probability of that server failing again, or of other servers failing
  2. A server either fails or it doesn’t – there are no “degrees of failure” to consider

CHI-SQUARE EXAMPLE

- Record observations:

Server Failures

Server	Observed
A	46
B	36
C	52
D	26
E	42
F	38

CHI-SQUARE EXAMPLE

- Add up the observations:

Server Failures

Server	Observed
A	46
B	36
C	52
D	26
E	42
F	38
Σ	**240**

CHI-SQUARE EXAMPLE

- Calculate expected values:

Server Failures

Server	Observed	Expected
A	46	40
B	36	40
C	52	40
D	26	40
E	42	40
F	38	40
Σ	**240**	

Since we expect each server to have the same probability of failure, divide the number of observations by the number of servers to get an expected failure rate of $240 \div 6 = 40$ for each server.

CHI-SQUARE EXAMPLE

- Consider thechi-square formula:

Server Failures

Server	Observed	Expected
A	46	40
B	36	40
C	52	40
D	26	40
E	42	40
F	38	40
Σ	**240**	

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

CHI-SQUARE EXAMPLE

- Subtract, then square, then divide:

Server Failures

Server	Observed	Expected	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
A	46	40	6	36	0.9
B	36	40	-4	16	0.4
C	52	40	12	144	3.6
D	26	40	-14	196	4.9
E	42	40	2	4	0.1
F	38	40	-2	4	0.1
Σ	**240**				

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

CHI-SQUARE EXAMPLE

- Add the last column:

Server Failures

Server	Observed	Expected	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
A	46	40	6	36	0.9
B	36	40	-4	16	0.4
C	52	40	12	144	3.6
D	26	40	-14	196	4.9
E	42	40	2	4	0.1
F	38	40	-2	4	0.1
Σ	**240**			Σ	**10.0**

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

$$\chi^2 = 10.0$$

CHI-SQUARE EXAMPLE

- Lookup our critical value:

Server Failures

Server	Observed
A	46
B	36
C	52
D	26
E	42
F	38

$$\alpha = 0.05$$
$$df = (6 - 1) = 5$$

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

$$\chi^2 = 10.0$$

CHI-SQUARE TABLE

Chi-square Critical Values						
	Probability of exceeding the critical value					
df	**0.15**	**0.10**	**0.05**	**0.01**	**0.005**	**0.001**
1	2.072	2.706	3.841	6.635	7.879	10.828
2	3.794	4.605	5.991	9.210	10.597	13.816
3	5.317	6.251	7.815	11.345	12.838	16.266
4	6.745	7.779	9.488	13.277	14.860	18.467
5	8.115	9.236	11.070	15.086	16.750	20.515

USING EXCEL

- To look up a critical value in Excel, for a 95% confidence level and 5 degrees of freedom:

  =CHISQ.INV.RT(0.05,5)

  11.0705

USING PYTHON

- To look up a critical value in Python, for a 95% confidence level and 5 degrees of freedom:

```
>>> from scipy.stats import chi2
>>> chi2.isf(0.05,5)
11.070497693516353
```

CHI-SQUARE EXAMPLE

- Lookup our critical value:

Server Failures

Server	Observed
A	46
B	36
C	52
D	26
E	42
F	38

$$\alpha = 0.05$$
$$df = (6-1) = 5$$
$$\chi^2_{critical} = 11.070$$

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

$$\chi^2 = 10.0$$

Since $\chi^2 < \chi^2_{critical}$
we fail to reject the null hypothesis

Even though Server C failed twice as many times as Server D, the results show that this can happen at least 5% of the time!

WHEN NOT TO USE CHI-SQUARE

- Chi-square statistics don’t perform well for small expected frequencies.
- Each cell should have a value greater than 5