χ^2 Distribution

(chi-square)

x^2 -Distribution

Suppose you modeled a situation using a probability distribution and have an expectation of how things will shape up in the long run.

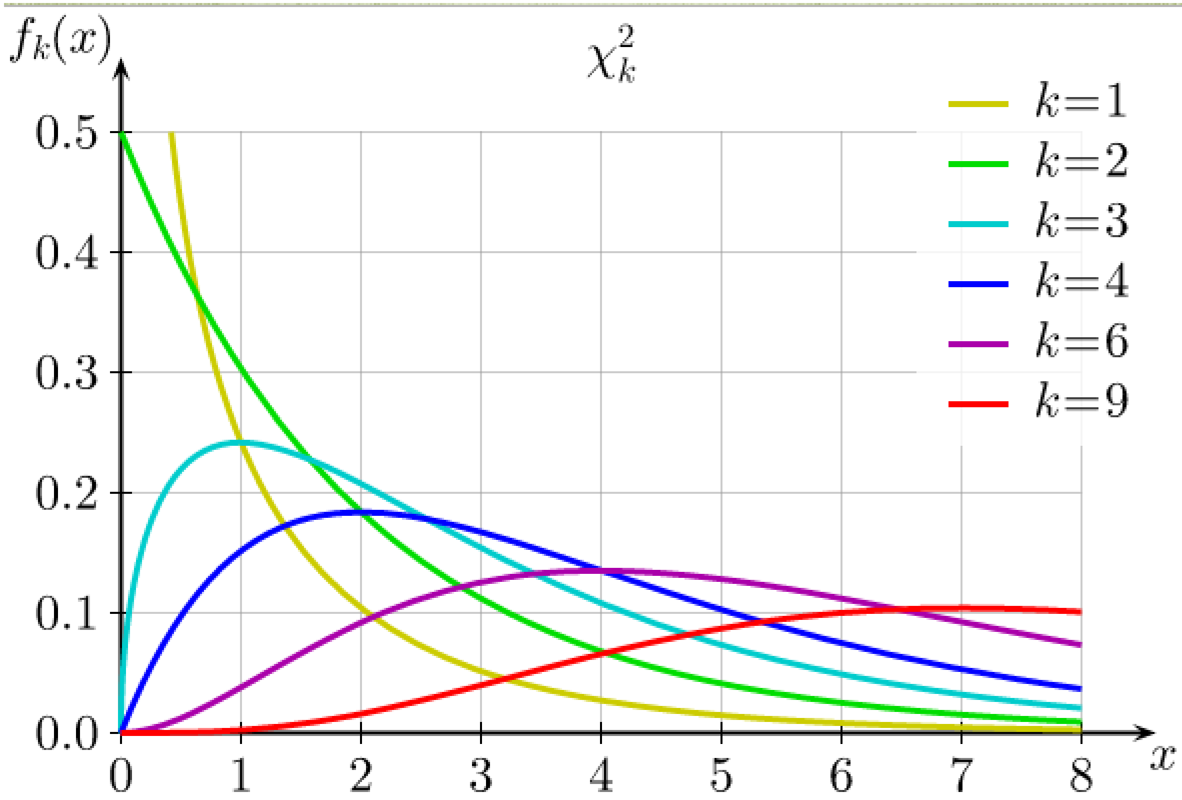
But what if, the frequency of observed data is different from expected data? We are not just comparing the mean of observed vs expected. We are talking about the entire distribution.

How would you know if the difference is due to normal fluctuations or if your model was incorrect?

Chi Square Distribution

- Let X ~ N(0,1)

- How does a distribution of x^2 look like?
- How does the distribution of sum of squares of two random picks looks like? i.e $x_1^2 + x_2^2$


$f_k(x)$
χ^2_k
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
x

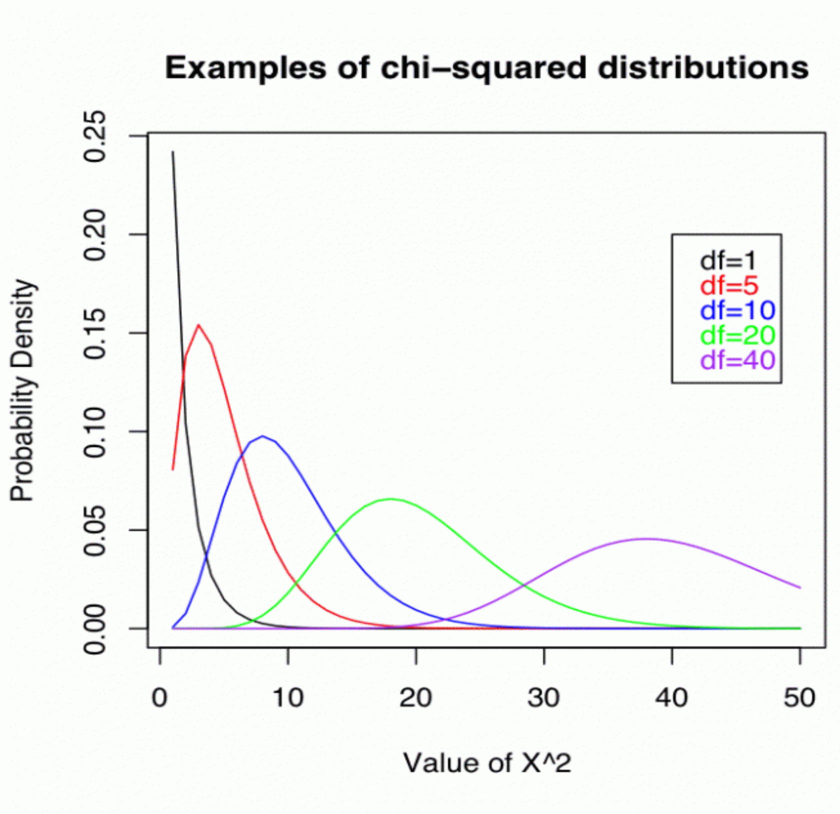
χ^2 distribution

Recall; $z = \frac{X-\mu}{\sigma}$

$z^2 = \frac{(X-\mu)^2}{\sigma^2}$

$z^2 = \chi^2_{(1)}$

χ^2 distribution is a distribution of squared deviates

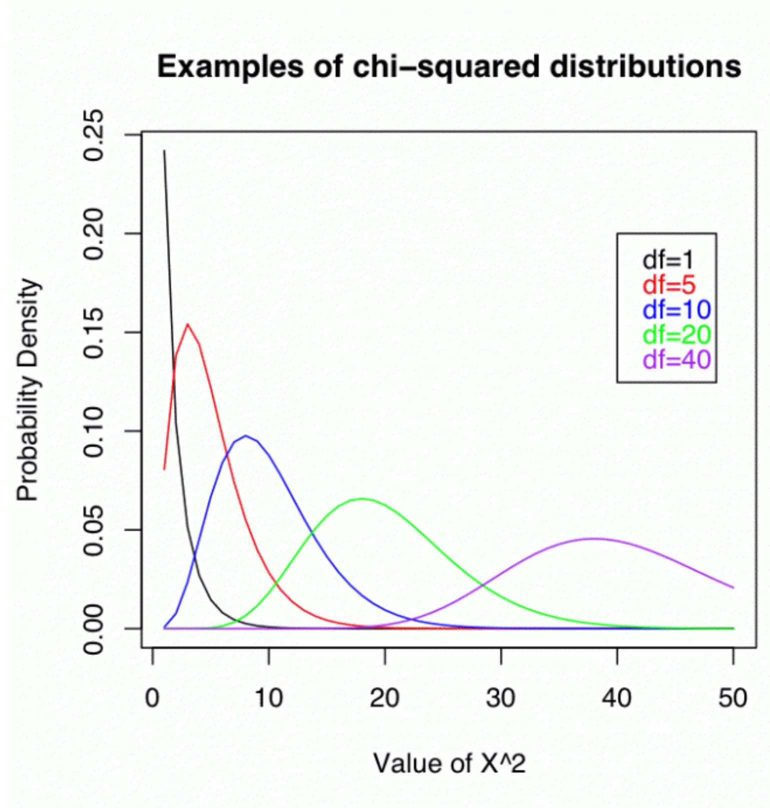


The shape depends on number of squared deviates added together.

χ^2 distribution

$\chi^2 \sim \chi^2_v$, where v represents the degree of freedom.

When v is greater than 2, the shape Of the distribution is skewed Positively gradually becoming approximately normal for large v.



Properties of χ^2 random variable

- A X^2 random variable takes values between 0 and 8.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- Mode of a χ^2 distribution is ν - 2.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads

wider.

- As ν increases, distribution tends to normal.

Attention check

Let us say you are running a casino and the slot machines are causing you headaches. You had designed them with the following expected probability distribution, with X being the net gain from each game played.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

You collected some statistics and found the following frequency of peoples' winnings.

x	-2	23	48	73	98
Frequency	965	10	9	9	7

You want to compare the actual frequency with the expected frequency.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

Are these differences significant and if they are, is it just pure chance?

χ^2 test to the rescue

χ^2 distribution uses a test statistic to look at the difference between the expected and the actual, and then returns a probability of getting observed frequencies as extreme.

$X^2 = \sum \frac{(O - E)^2}{E}$, where O is the observed frequency and E the expected frequency.

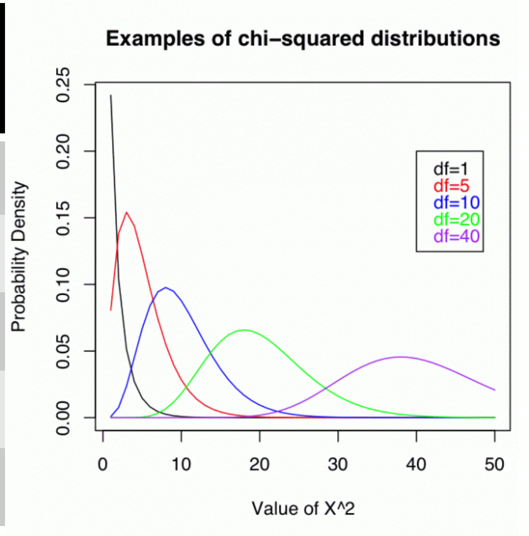
x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$$X^2 = 38.272$$

Is this high?

To find this, we need to look at the χ^2 distribution.

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1



In the above case, we had 5 frequencies to calculate. However

observed frequency (**RESTRICTION**), calculating 4 would give t

5th. Therefore, there are 5-1=4 degrees of freedom.

ν = **(number of classes) – (number of restrictions)**, or

ν = **(number of classes) – 1 – (number of parameters being estimated from sample data)**

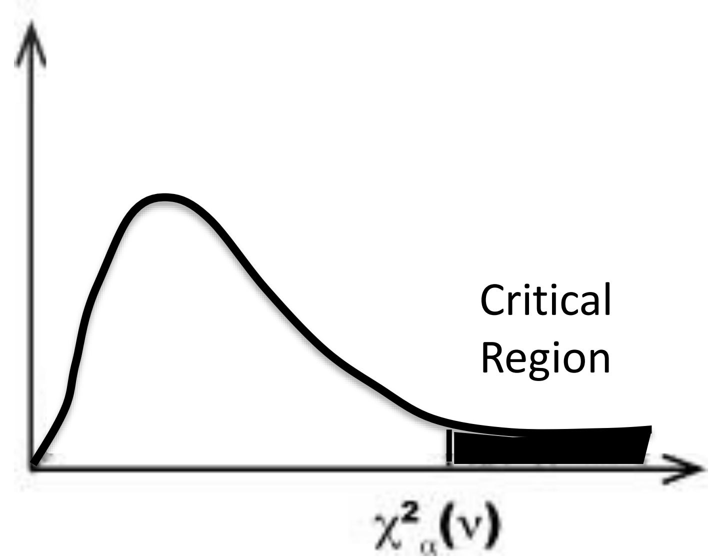
How do we know the Significance of the difference?

One-tailed test using the upper tail of the distribution as the critical region.

A test at significance level α is written as

$\chi^2_\alpha(v)$. The critical region is to its right

Higher the value of the test statistic, the bigger the difference between observed and expected frequencies.



Reference

www.khanacademy.com