χ^2 Distribution

(chi-square)

x^2 - Distribution

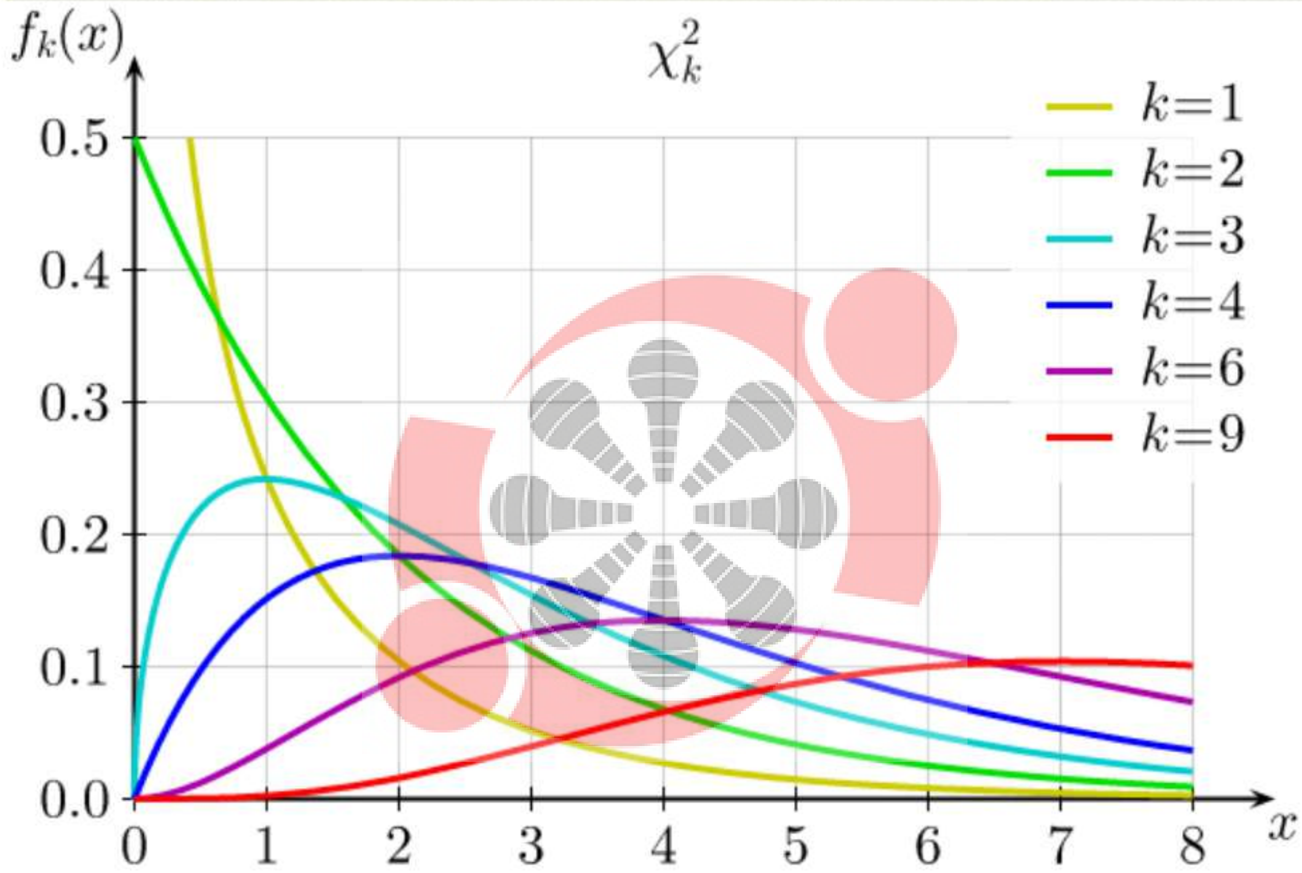
Suppose you modeled a situation using a probability distribution and have an expectation of how things will shape up in the long run.

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

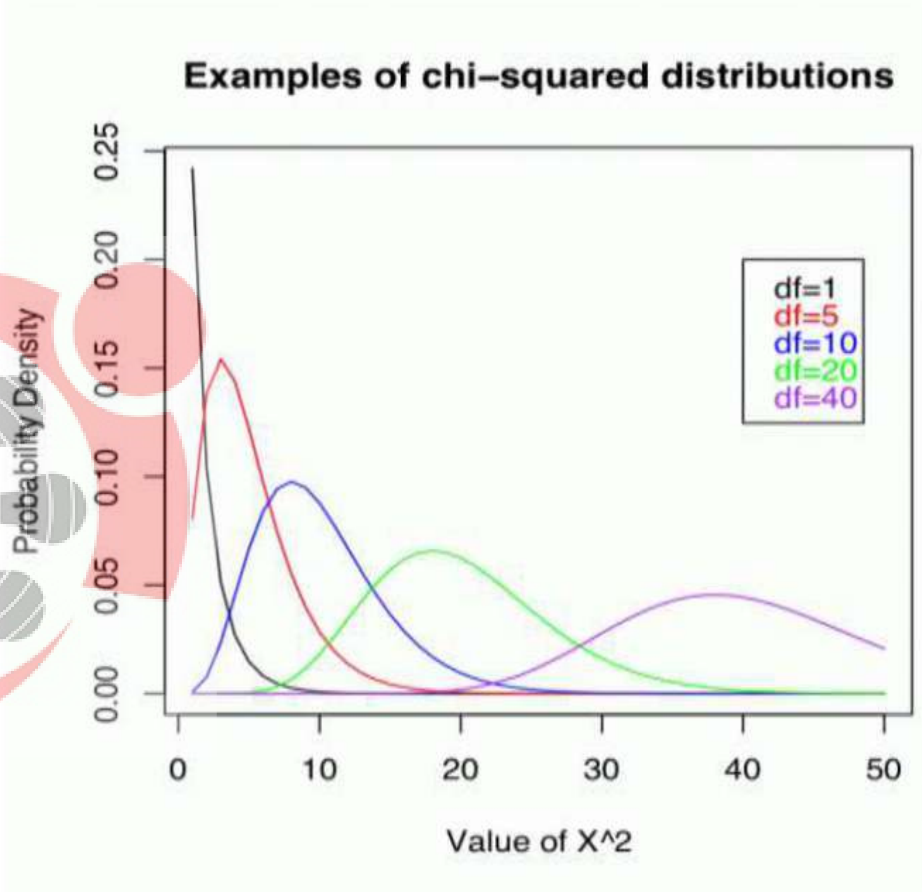
χ^2 distribution

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

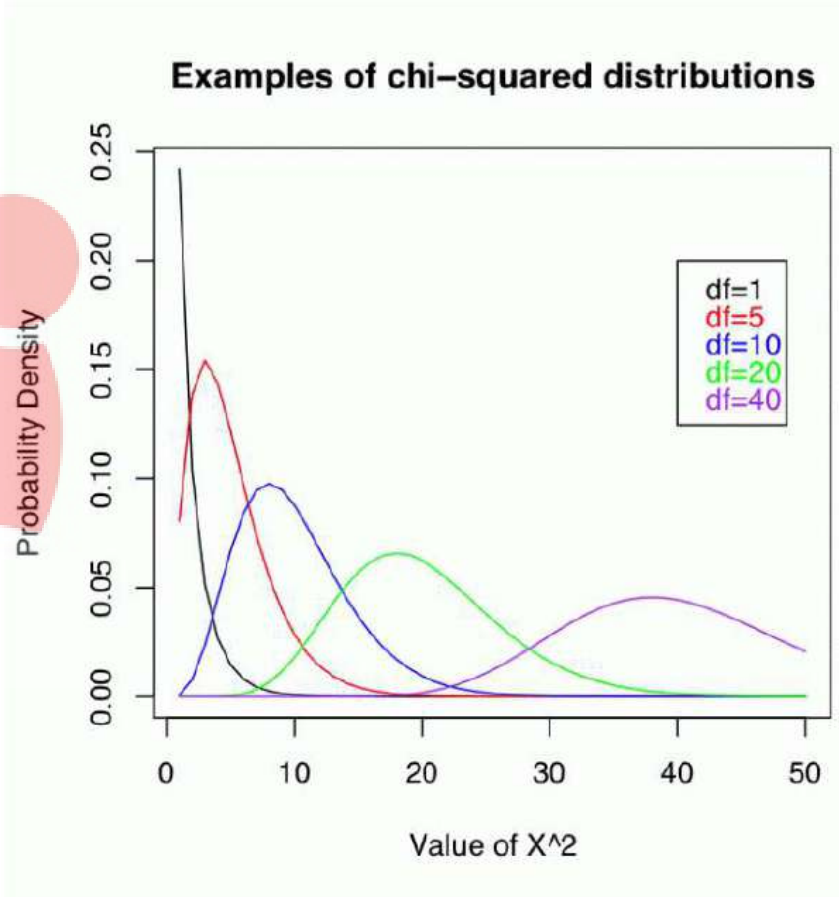


The shape depends on number of squared deviates added together.

χ^2 distribution

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

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



Properties of χ^2 random variable

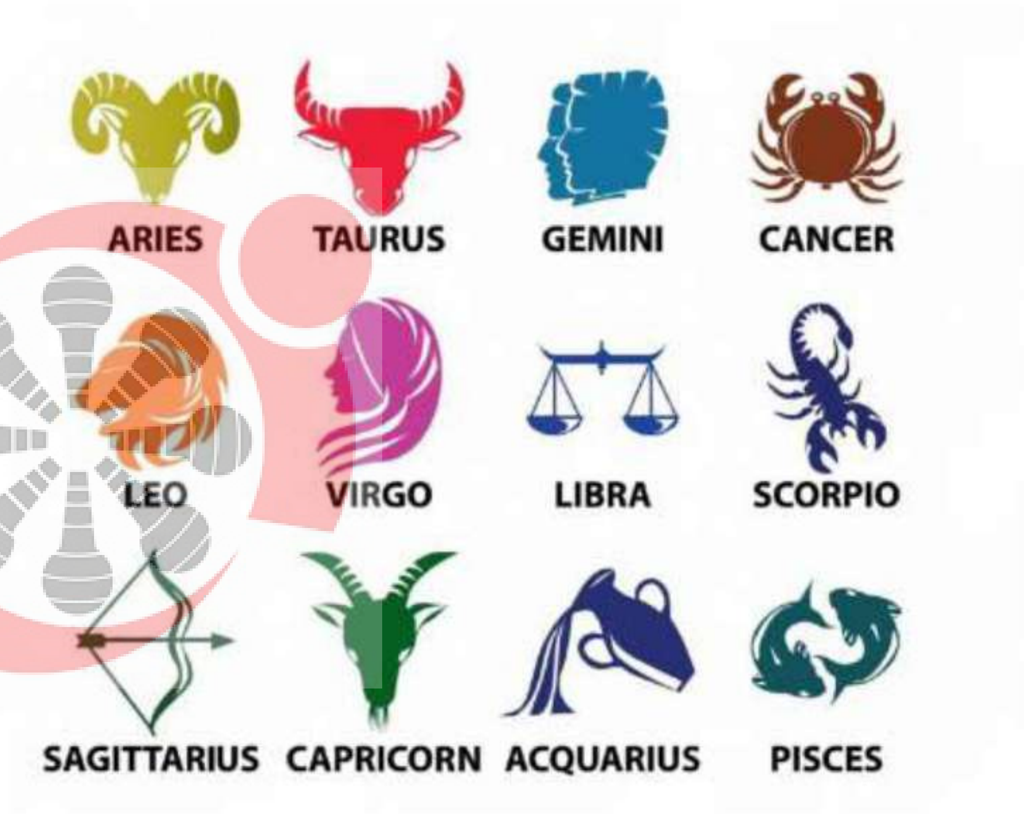
- A X^2 random variable takes values between 0 and 8.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- Mode of a χ^2 distribution is ν - 2.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads

wider.

- As ν increases, distribution tends to normal.

The Zodiac

256 visual artists were surveyed to find out their zodiac sign. The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23). Test the hypothesis that zodiac signs are evenly distributed across visual artists.



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

Category	Observed	Expected
Aries	29	21.333
Taurus	24	21.333
Gemini	22	21.333
Cancer	19	21.333
Leo	21	21.333
Virgo	18	21.333
Libra	19	21.333
Scorpio	20	21.333
Sagittarius	23	21.333
Capricorn	18	21.333
Aquarius	20	21.333
Pisces	23	21.333

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.

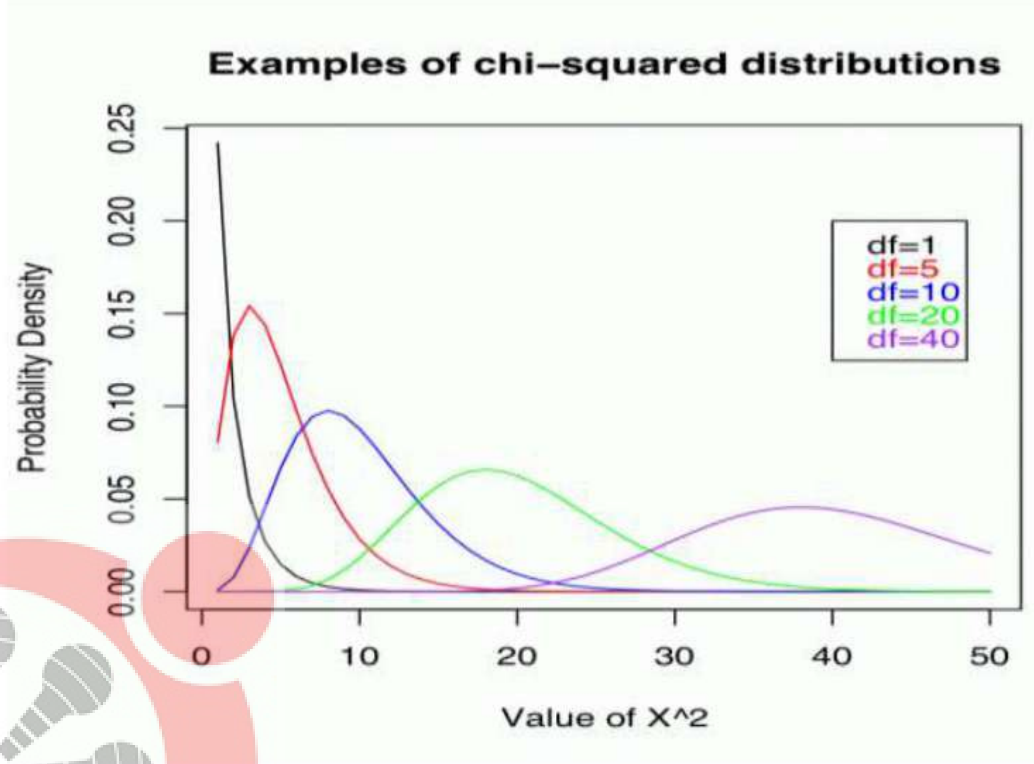
Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457

$X^2 = 5.094017203$

Is this high?

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

Category	Observed	Expected
Aries	29	21.333
Taurus	24	21.333
Gemini	22	21.333
Cancer	19	21.333
Leo	21	21.333
Virgo	18	21.333
Libra	19	21.333
Scorpio	20	21.333
Sagittarius	23	21.333
Capricorn	18	21.333
Aquarius	20	21.333
Pisces	23	21.333



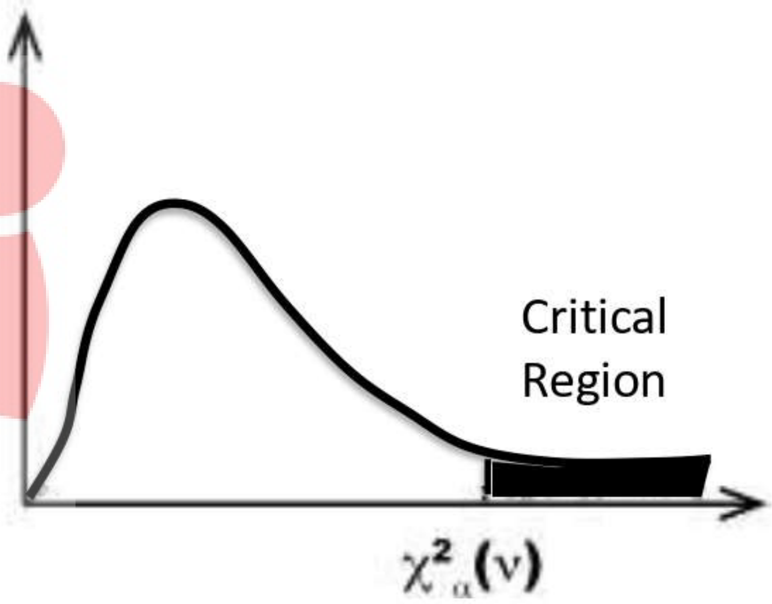
In the above case, we had 12 frequencies to calculate. However, observed frequency (RESTRICTION), calculating 11 would give the 12th. Therefore, there are 12-1 = 11 degrees of freedom.

υ = (number of classes) – (number of restrictions), or
υ = (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.

References

- https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/