Chi-Square test

- A test to measure the differences between what is observed and what is expected according to an assumed hypothesis.

Chi-Square test

- A test to measure the differences between what is observed and what is expected according to an assumed hypothesis.
- While the exact form depends on the problem, it is usually a great approximation to calculate χ^2 as

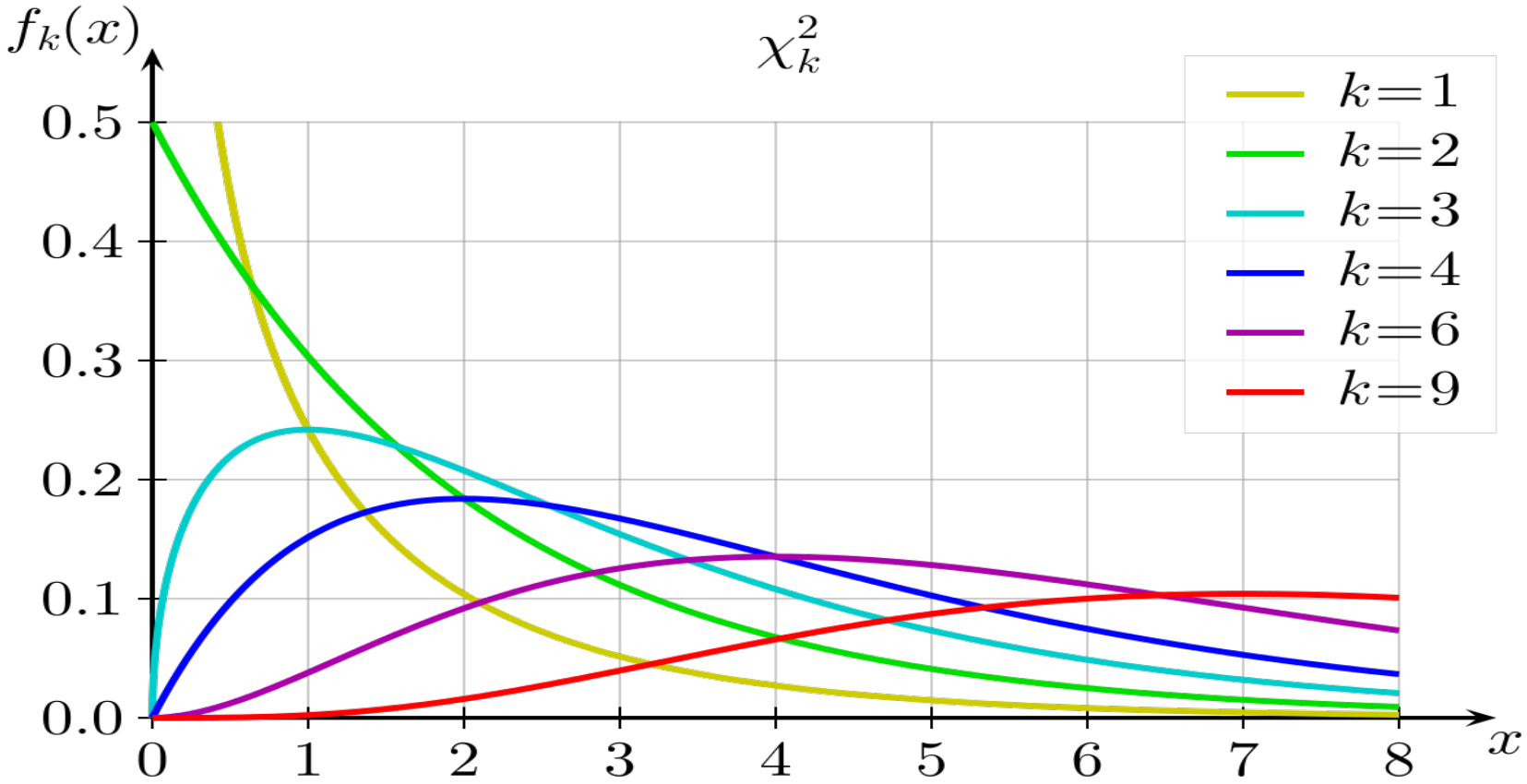
$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

where

- O_i stands for an observed value
- E_i is its expected value should the null hypothesis be true.

Chi-Square random variables

- A $\chi^2(k)$ variable
  is the sum of squares of k independent standard normal variables.



The Chi square test can be used when

1. the data is in the form of frequencies,
2. the frequency data must have a precise numerical value and must be organised into categories or groups,
3. the expected frequency in any cell of any given table must at least equal 5. Counts are discrete but the χ^2 approximation relies on the counts being roughly normally distributed.

Three main uses of Chi square

1. to test for a difference between two or more proportions,
2. to determine whether two variables are independent or related,
3. to determine whether the frequencies of a distribution are the same as the hypothesised frequencies.

Steps to follow in doing a Chi square test

1. determine the specific probability distribution to compare the data,
2. estimate or hypothesise the value of each parameter of the selected probabilities,
3. determine the theoretical probability in each category using the selected probability distribution
4. use the Chi square test statistic to test whether the selected probability is a good fit to the data.

Chi-Square test - Example

- We want to see if the proportion of people who experiences stroke differs among the smoker and non-smoker.

	Smoker	Non-smoker
Stroke	15	35
No stroke	8	42
Total	23	77

Chi-Square test - Example

- We want to see if the proportion of people who experiences stroke differs among the smoker and non-smoker.

	Smoker	Non-smoker
Stroke	15	35
No stroke	8	42
Total	23	77

- Is the difference between two datasets is only due to the randomness inherent to sampling?
- H_0: no difference between two groups

Chi-Square test - Example

- We want to see if the proportion of people who experiences stroke differs among the smoker and non-smoker.

	Smoker	Non-smoker
Stroke	15	35
No stroke	8	42
Total	23	77

- Is the difference between two datasets is only due to the randomness inherent to sampling?
- H_0: no difference between two groups
- If H_0 is true, we can ignore the smoker/non-smoker label and estimate the proportion of people in the population at large by dividing the total number of stroke by the total number of people.

Chi-Square test - Example

	Smoker	Non-smoker
Stroke	15	35
No stroke	8	42
Total	23	77

- The pooled proportion for the item of interest

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{15 + 35}{23 + 77} = 0.5$$

Chi-Square test - Example

- Let tabulate the expected counts below the actual counts in each cell.

	Smoker	Non-smoker
Stroke	15 11.5	35 38.5
No stroke	8 11.5	42 38.5
Total	23	77

Chi-Square test - Example

- Let tabulate the expected counts below the actual counts in each cell.

	Smoker	Non-smoker
Stroke	15 11.5	35 38.5
No stroke	8 11.5	42 38.5
Total	23	77

- We can be confident in claiming a real difference between the smoker and non-smoker if the observed counts deviate “extremely" from the expected ones.

Chi-Square test - Example

- The statistic for this sort of hypothesis test is

$$\sum_{\text{cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

- The degree of freedom is $rc - r - c + 1 = (r-1)(c-1)$

f_0	f_e	$(f_o - f_e)^2/f_e$
15	11.5	1.0652
35	38.5	0.3182
8	11.5	1.0652
42	38.5	0.3182
		$\chi^2_{\text{stat}} = 2.7668$

- $\chi^2_{0.05,1} = 3.841$

Chi-Square test for the difference in the proportion between 2 independent population

- Hypotheses:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

- Test Statistics: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ where f_o is the observed frequency and f_e is the expected frequency.
  - $$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
  - f_e for items of interest in group 1 $= \bar{p} \times n_1$
  - f_e for remaining items in group 1 $= (1 - \bar{p}) \times n_1$
  - f_e for items of interest in group 2 $= \bar{p} \times n_2$
  - f_e for remainning items in group 2 $= (1 - \bar{p}) \times n_2$
- Decision Rule: Reject H_0 if $\chi^2 > \chi^2_{critical} = \chi^2_{\alpha,1}$

McNemar test

- Testing the difference between two proportions when data are collected from repeated measures or matched samples.
- Hypotheses:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

- Test Statistics: $\chi^2 = \frac{(B-C)^2}{B+C}$
- Decision Rule: Reject H_0 if $\chi^2 > \chi^2_{critical}$

Test for independence between two categorical variables

- Independence tests test whether two attributes in a given population are related.
- To test the Independence of the two attributes A and B, we look for deviations from what's expected if A and B are independent.
- Hypotheses:

$$H_0 : \text{two variables are independent}$$
$$H_1 : \text{two variables are dependent}$$

- Test Statistics: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$ where

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{total observations}}$$

- Decision Rule: Reject H_0 if $\chi^2 > \chi^2_{critical} = \chi^2_{\alpha,(r-1)(c-1)}$

Goodness of fit test

- A goodness of fit test is used to test if sample data fits a distribution from a certain population.
- Hypotheses:

  H_0 : any observed effect is due solely to random chance.

  H_1 : there is a significant effect not due to chance alone.

- Test Statistics: $\chi^2 = \sum_{\text{all categories}} \frac{(f_o - f_e)^2}{f_e}$

  where $f_e =$ Probability of the category under the distribution under $H_0 \times$ Sample size.
- Decision Rule: Reject H_0 if $\chi^2 > \chi^2_{critical}$
- Note that every constraint and every point estimate of a parameter using experimental data reduces the degrees of freedom by 1.

Decision Analysis

- A decision analysis is a general approach that helps decision makers make intelligence choice.
- A decision analysis problem typically involves:
  - states of nature;
  - alternatives;
  - payoffs.
- Three categories of decision-making:
  - decision making under certainty,
  - decision making under uncertainty
  - decision making under risk

Three commonly used methods under uncertainty

1. Maximim : find the worse possible payoff for each alternative and then choose the maximum or best payoff of those minimums selected under each decision alternative.
2. Maximax: find the best possible payoff for each alternative and then choose the alternative that yields the maximum best possible payoff.
3. Minimax regret: minimises the maximum regret

Under risk

1. Expected monetary value (EMV): compute the expected monetary payoff for each alternative and choosing the alternative with the largest payoff.
2. Expected opportunity loss (EOL). An opportunity loss is the difference between the payoff for the decision you made and the payoff you would have received if you had made the best decision.
3. The decision selected by the maximum EMV approach is always the same as the decision selected by the minimum EOL approach.

EVPI

- The expected value of perfect information (EVPI) is the difference between the payoff that would occur if the decision maker knew which states of nature would occur and the expected monetary payoff without perfect information about the states of nature.
- It is the maximum amount a decision maker would pay for additional information.
- The value of EVPI is equal to the value of minimum EOL.