Chi square

LING572 Advanced Statistical Methods for NLP
January 23, 2020

Chi square

- An example: is having a masters degree a good feature for predicting footwear preference?
  - A: MS (binary)
  - B: footwear preference

- Bivariate tabular analysis:
  - Is there a relationship between two random variables A and B in the data?
  - How strong is the relationship?
  - What is the direction of the relationship?

Raw frequencies

	Sandal	Sneaker	Leather shoe	Boots	Others
MS	6	17	13	9	5
no-MS	13	5	7	16	9

Feature: has a masters degree/not

Classes: {Sandal, Sneaker, ….}

Two distributions

Observed distribution (O):

	Sandal	Sneaker	Leather	Boot	Others	Total
MS	6	17	13	9	5	50
no-MS	13	5	7	16	9	50
Total	19	22	20	25	14	100

Expected distribution (E):

	Sandal	Sneaker	Leather	Boot	Others	Total
MS						50
no-MS						50
Total	19	22	20	25	14	100

Two distributions

Observed distribution (O):

	Sandal	Sneaker	Leather	Boot	Others	Total
MS	6	17	13	9	5	50
no-MS	13	5	7	16	9	50
Total	19	22	20	25	14	100

Expected distribution (E):

	Sandal	Sneaker	Leather	Boot	Others	Total
MS	9.5	11	10	12.5	7	50
no-MS	9.5	11	10	12.5	7	50
Total	19	22	20	25	14	100

Chi square

- Expected value = row total * column total / table total

  = P(row value) * P(column value) * table total

- $$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- $\chi^2 = (6\text{-}9.5)^2/9.5 + (17\text{-}11)^2/11 + \ldots.$

  = 14.026

Calculating χ^2

- Fill out a contingency table of the observed values ➔ O
- Compute the row totals and column totals
- Calculate expected value for each cell assuming no association ➔ E
- Compute chi square: $(O - E)^2/E$

When r=2 and c=2

O =

	$\bar{c}_i$	c_i	total
$\bar{t}_k$	a	b	a+b
t_k	c	d	c+d
total	a+c	b+d	N

E =

	$\bar{c}_i$	c_i	total
$\bar{t}_k$	$\frac{(a+c)(a+b)}{N}$	$\frac{(b+d)(a+b)}{N}$	a+b
t_k	$\frac{(a+c)(c+d)}{N}$	$\frac{(b+d)(c+d)}{N}$	c+d
total	a+c	b+d	N

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(ad - bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

χ^2 test

Basic idea

- Null hypothesis (the tested hypothesis): no relation exists between two random variables.
- Calculate the probability of having the observation with that χ^2 value, assuming the hypothesis is true.
- If the probability is too small, reject the hypothesis.

Requirements

- The events are assumed to be independent and have the same distribution.
- The outcomes of each event must be mutually exclusive.
- At least 5 observations per cell.
- Collect raw frequencies, not percentages

Degree of freedom

- Degree of freedom df = (r – 1) (c – 1)

  r: # of rows c: # of columns

- In this ex: df=(2-1)(5-1)=4

χ^2 distribution table

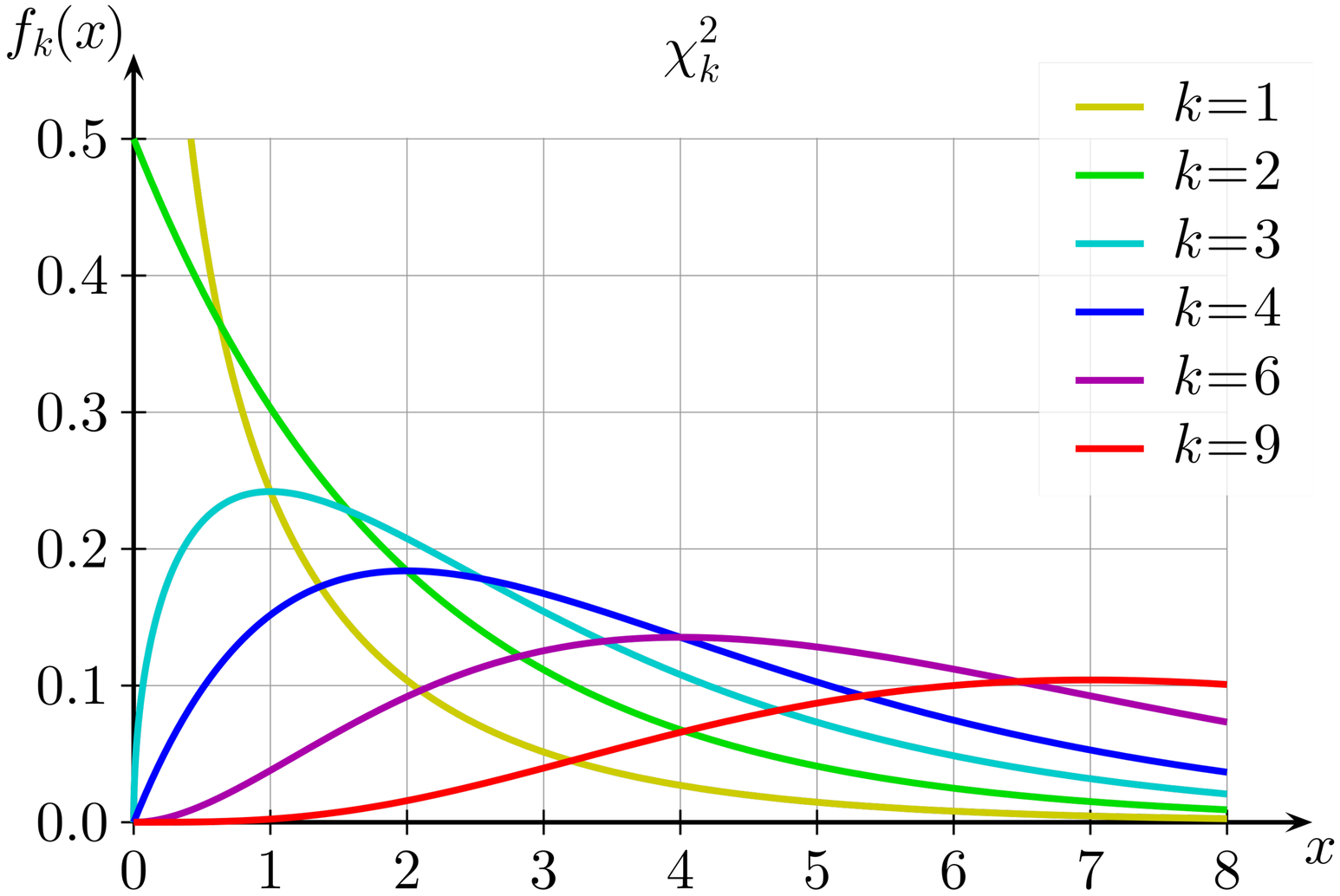
	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
…					

df=4 and 14.026 > 13.277

➔ $p<0.01$

➔there is a significant relation

χ^2 distribution



source

χ^2 to P Calculator

http://vassarstats.net/newcs.html

scipy.stats.chi2_contingency

Steps of χ^2 test

- Select significance level p_0
- Calculate χ^2

- Compute the degrees of freedom

  df = (r-1)(c-1)

- Calculate p given χ^2 value (or get the χ^2_0 for p_0)

- if $p < p_0$ (or if $\chi^2 > \chi^2_0$)

  then reject the null hypothesis.

Summary of χ^2 test

- A very common method for determining whether two random variables are independent

- Many good tutorials online
  - Ex: http://en.wikipedia.org/wiki/Chi-square_distribution
  - https://www.khanacademy.org/math/ap-statistics/chi-square-tests/chi-square-tests-two-way-tables/v/chi-square-test-homogeneity

Applying to Text Classification

- Exercise: is ‘bad’ a good feature for predicting sentiment?
  - Is sentiment *independent* from ‘bad’ or not?
  - What are counts in this table?
    - Number of documents

	bad=1	bad=0	Total
positive	13	185	
negative	212	28	
Total			

Additional slides

χ^2 example

- Shared Task Evaluation:
  - Topic Detection and Tracking (aka TDT)

- Sub-task: Topic Tracking Task
  - Given a small number of exemplar documents (1-4)
    - Define a topic
    - Create a model that allows tracking of the topic
      - I.e. find all subsequent documents on this topic
  - Exemplars: 1-4 newswire articles
    - 300-600 words each

Challenges

- Many news articles look alike
  - Create a profile (feature representation)
  - Find terms that are strongly associated with current topic

- Not all documents are labeled
  - Only a small subset belong to topics of interest
    - Differentiate from other topics AND ‘background’

Approach

- X^2 feature selection:
  - Assume terms have binary representation
    - Positive class term occurrences from exemplar docs
    - Negative class term occurrences from
      - other class exemplars, ‘earlier’ uncategorized docs
  - Compute X^2 for terms
    - Retain terms with highest X^2 scores
    - Keep top N terms
- Create one feature set per topic to be tracked

Tracking Approach

- Build vector space model
  - Feature weighting: tf*idf
  - Distance measure: Cosine similarity

- Select documents scoring above threshold

- Result: Improved retrieval