Metody probabilistyczne i statystyka, 2022
informatyka algorytmiczna, WIiT PWr

7-Statistical Inference 2

chapter 10 from Byron

Testing a distribution

Previously discussed:

learning unknown parameters while distribution was known

Problem:

How to check that a source A has probability distribution D?

A crucial question!

- **understanding social networks**
- **anomaly detection in cybersecurity**
- **correlations (therapy versus mortality rate …)**
- **…**

Chi Square test

given a data sample and a hypothetic distribution with density *f*

Chi Square test

Goal: testing whether a source is distributed according to a hypothesis H_0

Method:

- ❑ **split the population to *N* bins**
- ❑ **count how many samples fall into each bin**
- ❑ **compute the expected value for each bin under H_0**
- ❑ **calculate statistics (what function?)**
- ❑ **make a decision**

Chi Square test

Statistics used:

$$\chi^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}.$$

Obs(k) **= number of samples in bin *k***
Exp(k) **= expected number of samples in bin *k***

One sided statistics: result ≤ thereshold → accept H_0
result > thereshold → reject H_0

rejection region: $R = [\chi^2_\alpha, +\infty),$

Chi Square background

Pearson's Theorem

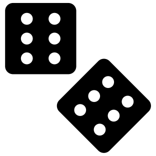
χ^2 distribution for N bins converges to the chi-square distribution with N-1 degrees of freedom

Rule of thumb:

each bin should contain at least 5 samples

Chi Square application example

testing whether a die is unbiased:



- ❑ **6 bins corresponding to 6 possible outcomes**
- ❑ **90 samples**
- ❑ **Exp(i)=90/6=15**
- ❑ **Counts observed: 20,15,12,17,9,17**
- ❑ **Statistics:**

$$\chi^2_{\text{obs}} = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}$$

$$= \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(9-15)^2}{15} + \frac{(17-15)^2}{15} = 5.2.$$

Chi Square application example

interpretation

ν (d.f.)	α, the right-tail probability													
	.999	.995	.99	.975	.95	.90	.80	.20	.10	.05	.025	.01	.005	.001
1	0.00	0.00	0.00	0.00	0.00	0.02	0.06	1.64	2.71	3.84	5.02	6.63	7.88	10.8
2	0.00	0.01	0.02	0.05	0.10	0.21	0.45	3.22	4.61	5.99	7.38	9.21	10.6	13.8
3	0.02	0.07	0.11	0.22	0.35	0.58	1.01	4.64	6.25	7.81	9.35	11.3	12.8	16.3
4	0.09	0.21	0.30	0.48	0.71	1.06	1.65	5.99	7.78	9.49	11.1	13.3	14.9	18.5
5	0.21	0.41	0.55	0.83	1.15	1.61	2.34	7.29	9.24	11.1	12.8	15.1	16.7	20.5

5.2 belongs here

Chi Square test for independence

Goal: test the hypothesis that two parameters of the sample are stochastically independent

Application: eliminating false claims about dependence

example:
„eating cabbage influences the cholesterol level in blood"

Chi Square test for independence

	B_1	B_2	$\cdots$	B_m	row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

$$\widehat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \quad \widehat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \quad \widehat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

Chi Square test for independence

$$\widehat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \quad \widehat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \quad \widehat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

$$\widetilde{P}\{x \in A_i \cap B_j\} = \left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right)$$

$$\widehat{\mathrm{Exp}}(i,j) = n\left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$$

$$\chi^2_{\mathrm{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left\{Obs(i,j) - \widehat{\mathrm{Exp}}(i,j)\right\}^2}{\widehat{\mathrm{Exp}}(i,j)}.$$

Chi Square test for independence

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left\{Obs(i,j) - \widehat{\text{Exp}}(i,j)\right\}^2}{\widehat{\text{Exp}}(i,j)}.$$

Number of degrees of freedom for Chi-square distribution:

$$km - (k + m - 1) = (k-1)(m-1)$$

motivation: ***k*** **equations for computing** $n_{i\cdot}$

m **equations for computing** $n_{\cdot j}$

but only ***m+k-1*** **independent**

Chi Square test for independence

… and look into tables (input to a program)