

Lab 1: Revision

12.03.2015

1. χ^2 test – dependency between nominal variables.

(a) Hypotheses:

H0: X and Y are independent

H1: X and Y are dependent

(b) Formula for calculating value of statistics:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where O is observed value and E is expected value. Expected values are calculated using formula:

$$E = \frac{(\text{suminrow}) * (\text{sumincolumn})}{\text{totalsum}}$$

(c) Number of degrees of freedom is: $df = (k - 1)(l - 1)$.

(d) Find critical value for given significance level α and make final decision.

2. Chi square test help us to decide whether there exist dependencies between nominal variables. In order to calculate strength of this dependency we can use:

n – number of observations

k, l – number of different values of random variables

(a) Φ -Yul coefficient

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

(b) V-Cramer coefficient

$$V = \sqrt{\frac{\chi^2}{n * \min(k - 1; l - 1)}}$$

(c) Pearson's contingency coefficient

3. Entropy - expected surprisal.

$$Ent(X) = - \sum_i p(x_i) \log_b(p(x_i))$$

Figure 1: Entropy of a discrete random variable X

$b = 2$ - units of Ent = bits

$$Ent(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$

Figure 2: Conditional entropy