Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Pearson's χ^2 tests [chi-squared]

- **Test of independence**
  Are two variables, expressed in a contingency table, independent of each other?
- **Goodness-of-fit test**
  Does an observed frequency distribution differ from a hypothesized theoretical probability distribution?
- **Test of homogeneity**
  Does two observed frequency distributions of the same categorical variable come from populations with different probability distributions?
  - works with a 2-way contingency table similarly as the independence test

Sum of k independent standard normal variables

Let $Z_i \sim \mathrm{N}(0, 1)$ be independent variables with standard normal distribution.

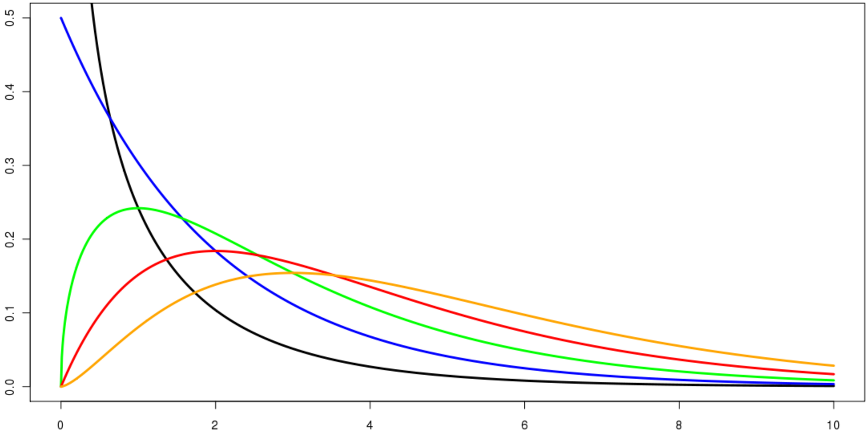
Then what is the distribution of $\sum_{i=1}^{k} Z_i^2$?

```
show.sum.Z.square <- function(k) {
  # shows the empirical distribution of the sum of
  # k independent standard normal variables
  # mean = k, variance = 2k

  sum.Z2 = 0
  for(i in 1:k){ sum.Z2 = sum.Z2 + rnorm(10^6)^2 }

  cat("Sample statistics:\n")
  print(summary(sum.Z2))
  cat("\nSample variance: ", var(sum.Z2), "\n")
  plot(cut(sum.Z2, 200))
}
```

χ^2 distribution – density



Chi-Squared test of independence

A test of independence assesses whether observations on two variables, expressed in a contingency table, are independent of each other.

χ^2 independence test

We observe two categorical variables. $O_{i,j}$ are the observed frequencies arranged in an contingency table. Expectations $E_{i,j}$ can be computed using estimated marginal probabilities. Pearson's χ^2 test is based on the following formula for Pearson's cumulative test statistic

$$\boxed{X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}}$$

Pearson's cumulative test statistic X^2 has approximately χ^2_{df} distribution, where the degrees of freedom is

$$df = (Rows - 1) \times \ (Cols - 1)$$

χ^2 independence test

Then we compare the test statistic with
χ^2 critical value $\chi^2_k(\alpha)$, which is defined by

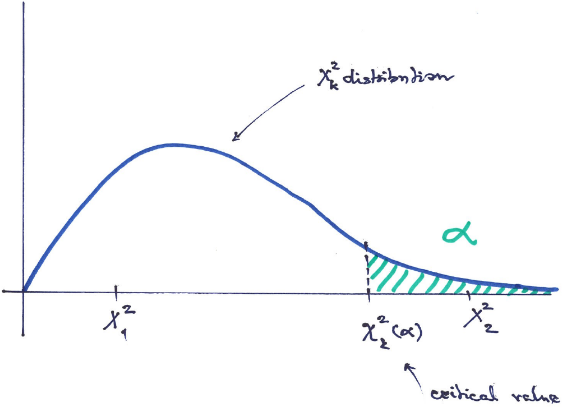
$$\boxed{\Pr\left\{X^2 > \chi^2_k(\alpha)\right\} = \alpha}$$

Practical note
χ^2 critical value can be computed as a quantile.

```
> qchisq( (1-alpha), df=k )
```

TODO: Get familiar with functions `{p|d|q}chisq()` available in R.

Critical value $\chi^2_k(\alpha)$ of the χ^2 distribution χ^2_k



X_1^2, X_2^2 ... statistics

$X_1^2 < \chi^2_k(\alpha) \Rightarrow$ p-value $> \alpha$

$X_2^2 > \chi^2_k(\alpha) \Rightarrow$ p-value $< \alpha$

Chi-Squared Goodness of Fit Test

The Chi-Squared Goodness of Fit Test is a test for comparing a theoretical distribution with the observed data from a sample.

χ^2 Goodness-of-fit test

Example 1
Rolling a die – after 600 rolls you got the following distribution

1	2	3	4	5	6
95	108	101	85	110	101

Question: Is the die fair? = Does it have the uniform distribution?

Example 2
Our hypothesis is that our classifier accuracy is 78 %. However, a test on 100 randomly chosen instances gives the following result

correct	error
81	19

Question: Should we reject the hypothesis?

χ^2 Goodness-of-fit test

Pearson's χ^2 goodness-of-fit test is based on the following formula for Pearson's cumulative test statistic

$$\boxed{X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}}$$

If the observed variables O_i have multinomial distribution, then Pearson's cumulative test statistic X^2 has approximately χ^2_{m-1} distribution.

χ^2 Goodness-of-fit test — example

Example based on real data

SENSES	estimated probabilities	test set observations
cord	9.2%	37
division	8.9%	51
formation	8.1%	52
phone	10.6%	44
product	53.5%	268
text	9.8%	48

```
> x = c(37, 51, 52, 44, 268, 48)
> p = c(9.2, 8.9, 8.1, 10.6, 53.5, 9.8)/100
```

Examination requirements

Chi-square tests

- Theory and practical use
- Independence test
- Goodness-of-fit test