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 in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Lecture 9 – Statistical tests II

χ^2-tests

- **Goodness-of-fit test**
- **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 $\sum_{i=1}^{k} Z_i^2$?

Sum of k independent standard normal variables

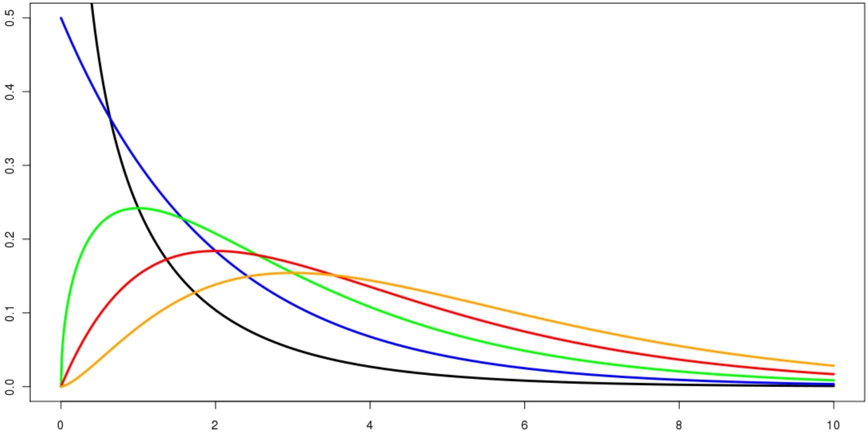
Let $Z_i \sim \mathrm{N}(0,1)$ be independent variables with standard normal distribution.

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

  Z <- rnorm(10^6); sum.Z.square <- Z^2
  while(k > 1) {
    Z <- rnorm(10^6); sum.Z.square <- sum.Z.square + Z^2
    k <- k-1
  }
  print(summary(sum.Z.square))
  plot(cut(sum.Z.square, 200))
}
```

χ^2 distribution – density



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

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?

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

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 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 χ^2_{m-1} distribution.

χ^2 Goodness-of-fit test

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

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

χ^2 Goodness-of-fit test

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

χ^2 Goodness-of-fit test

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
```