Probability Theory & Statistics

Innopolis University, BS-I,II

Spring Semester 2016-17

Lecturer: Nikolay Shilov

Part I

INTRO TO TEST STATISTICS

What for “Pearson’s chi-squared test”?

- Chi-squared tests a hypothesis that the observed frequency distribution of events is consistent with a particular distribution.
- Example: an ordinary die is “fair” – all six outcomes occur equally – i.e. frequency is consistent with uniform distribution.

Chi-squared test in brief

- Hypothesis H_0: (accidental) sampling

$$X_n = (x_1, \ldots x_n)$$

is consistent with distribution F with given confidence $p \in [0,1]$ (or significance $q=1-p$).

(Here consistency means that values $(x_1, \ldots x_n)$ are generated by testing some random variable X with distribution F.)

Chi-squared test in brief (cont.)

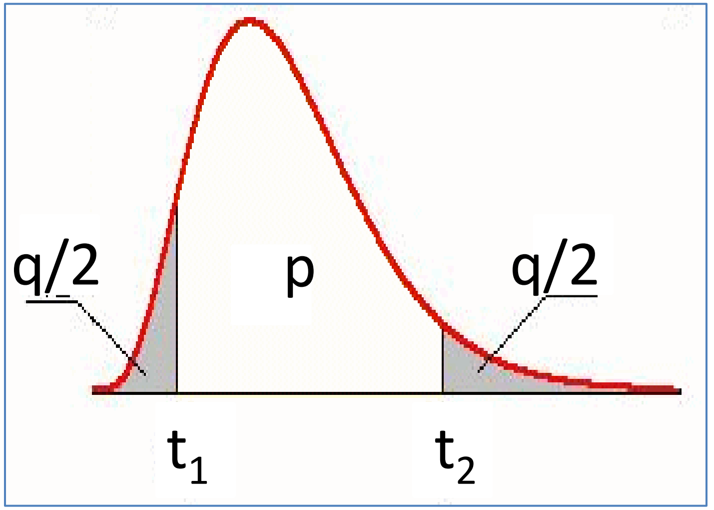
- Let $(a,b) \subseteq R^{\infty}$ be range of a random variable with distribution F;
- select freedom degree (d.f.) $k>1$ and split (a,b) onto $(k+1)$ disjoint events – intervals $(a_i, a_{i+1}]$ where $i \in [1..k]$, $a_1=a$, $a_{k+1}=b$;

Chi-squared test in brief (cont.)

- for each $i \in [1..k]$ let
  - n_i be number of $(x_1,\ldots x_n)$ within $(a_i, a_{i+1}]$: $n_i = |\{ x \in X_n : a_i < x \leq a_{i+1}\}|$;
  - $p_i = F(a_{i+1}) - F(a_i)$ the probability of the event $(a_i, a_{i+1}]$ according to H_0;
- compute statistic $\chi^2 = \sum_{i \in [1..k]} (n_i - n*p_i)^2/(n*p_i)$;

Chi-squared test in brief (cont.)

- define t_1 and t_2 according to significance q using Pearson's distribution chi-squared with k freedom degrees χ_k^2;



$\chi_1{}^2$ for k=1,…5

0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99
0,0002	0,0010	0,0039	0,0158	0,0642	0,1485	0,2750	0,4549	0,7083	1,0742	1,6424	2,7055	3,8415	5,0239	6,6349
0,0201	0,0506	0,1026	0,2107	0,4463	0,7133	1,0217	1,3863	1,8326	2,4079	3,2189	4,6052	5,9915	7,3778	9,2103
0,1148	0,2158	0,3518	0,5844	1,0052	1,4237	1,8692	2,3660	2,9462	3,6649	4,6416	6,2514	7,8147	9,3484	11,3449
0,2971	0,4844	0,7107	1,0636	1,6488	2,1947	2,7528	3,3567	4,0446	4,8784	5,9886	7,7794	9,4877	11,1433	13,2767
0,5543	0,8312	1,1455	1,6103	2,3425	2,9999	3,6555	4,3515	5,1319	6,0644	7,2893	9,2364	11,0705	12,8325	15,0863

Chi-squared test in brief (cont.)

- Conclusion:
  - if $\chi^2 \leq t_1$ then hypothesis H_0 holds (with significance level q);
  - if $t_1 < \chi^2 < t_2$ then hypothesis H_0 *may* hold;
  - if $t_2 \leq \chi^2$ then hypothesis H_0 is refuted.

Part II

EXAMPLES

Innopolis data

You know that information about Innopolis (https://ru.wikipedia.org/wiki/Иннополис) is very much incomplete.

Male-female ratio

- I would like
  - to check hypothesis that the ration of male and female of Innopolis residents is 50-50 with significance 0.05
  - using shuttle statistics (like in lecture for week 10) that gives 54 male and 46 female.

Using chi-squared test

- Since we have 2 events (male/female) then f.d. is 1 and I should use χ_1^2 for

  $\chi^2 = \Sigma_{i\in[1..k]} (n_i - n*p_i)^2/(n*p_i) =$

  $= (54\text{-}50)^2/50 + (46\text{-}50)^2/50 = 0.64.$
- Sorry, shuttle statistics can neither confirm nor refute the hypothesis.

(Pseudo-)random numbers

- Assume that a pseudo-random numbers algorithm generates n values in [0,1]
- and a hypothesis stating that they are uniformly distributed with high confidence has been confirmed
- then this generator is not very random (since values are too much uniformly distributed).

Part III

STATISTICS GLOSSARY: TERMS AND NOTATION

Hypothesis

- (Statistical) hypothesis: a statement about the discribution.
- Simple hypothesis: any hypothesis which specifies the distribution exactly.
- Composite hypothesis: any hypothesis which does not specify the distribution completely.

Hypothesis (cont.)

- Null hypothesis (H_0): usually a simple hypothesis one would like to prove.
- Alternative hypothesis (H_1): a hypothesis (often composite) opposite to the null hypothesis.

Test statistic

- Statistic: a value calculated from a (accidental) sample (often to summarize the sample for comparison purposes).
- Statistical test: a procedure whose inputs are (accidental) samples and hypothesis and whose result is hypothesis acceptance or refutation of hypothesis.

Test regions

- Region of acceptance: the set of values of the test statistic for which we fail to reject the null hypothesis.
- Region of rejection / Critical region: the set of values of the test statistic for which the null hypothesis is rejected.
- Critical value: the threshold value delimiting the regions of acceptance and rejection for the test statistic.

The End