Statistics

Week 8: Chi-squared Test (Chapter 9)

ESD, SUTD

Term 5, 2017

Outline

Goodness of fit – discrete

We can use the chi-squared test to determine whether a specified distribution fits our data!

We first look at an example involving a discrete distribution.

Example – number of passengers per car

Does the data suggest that the numbers of passengers follow a Poisson distribution? Use $\alpha = 0.01$.

Number	0	1	2	3	4
Frequency	678	227	56	28	8

Recall the Poisson distribution has the pmf

$$p(i) = e^{-\lambda} \frac{\lambda^i}{i!}.$$

The (unknown) parameter λ is, conveniently, also the mean.

Discrete example

Refer to the *Excel* file.

- We can *estimate* λ from sample mean, $\hat{\lambda} = 0.456$.
- Use the `exp` and `fact` commands to compute $e_i = n\,p(i)$ (except for e_4).
- **Rule 1**: every time we estimate a parameter, we lose 1 extra degree of freedom.
- **Rule 2**: make sure that each $e_i \geq 5$, in particular, no e_i should be < 1. Try to combine small e_i's with adjacent ones (and do the same to the corresponding observed values).

Final result: $\chi^2 = 72.2$, $\chi^2_{2,\,0.99} = 9.21$.

Continuous example

See *Excel* for an example testing whether some waiting times can be modeled by an exponential distribution (`expon.dist`).

- Since the mean is $1/\lambda$, we can estimate λ by $1/(\text{sample mean})$.
- The sample mean calculation can be simplified if we had all the data values.
- There are different ways to group the data, e. g. another way is to make the expected probability constant across the categories.

Final result: $\chi^2 = 6.09, \quad \chi^2_{4,\,0.95} = 9.49.$

Two-way tables

The χ^2 statistic can also be computed for two-way tables, to test if the two variables involved are **independent**.
(For more information, see textbook Section 9.4.)

See *Excel* example for Income vs Job Satisfaction, using $\alpha = 0.05$.

- If the null hypothesis (that the variables are independent) is true, then the expected number for each cell in the table is (row sum) $\times$ (column sum) / (grand sum)
- χ^2 is calculated using the same formula, but with a double sum.
- For an $n \times m$ table, the degree of freedom is $(n-1)(m-1)$.

Final result: $\chi^2 = 12.0$, $\chi^2_{9,\,0.95} = 16.9$.

Further notes

- You may find the chi-squared test very useful for your project.
- The chi-squared test tends to work well when n is very large (several hundreds, or several thousands).
- A weakness of the chi-squared test is that different groupings of the data may result in different conclusions.

  (The K-S test overcomes this problem.)