Data Mining Theory

DM-03: Supplement for Independence Test

Summary of Various Tests(3)

6. 2x2 cross tabulation → Is there significant association between 2 axes ?
   - Fisher's exact test
   - Accurate than χ^2 for small frequency and highly skewed
   - odds, p_val = ss.fisher_exact(DataFrame)

7. General cross tabulation→ Is there significant association between 2 axes ?
   - χ^2 test (Independence Test)
   - Each frequency should be 5 or more
   - chi2, p_val, dof, expected = ss. chi2_contingency(DataFrame)

cross tabulation

8. Observed frequency (x_obs) and theoretical frequency (x_exp) → Are they significantly different ?
   - χ^2 test (Test of Goodness of fit)
   - Each frequency should be 5 or more
   - chi2, p_val = ss. chisquare(x_obs, f_exp=x_exp)

If testing the categories to be equally likely, f_exp= can be omitted

dm-03-assign1

- shop_sales.csv in dm-04 contains sample data of some stores of an enterprise, including prefecture (Prefs), store type (franchised / directly managed), sales of a product (before a campaign (Sales1) and after a campaign (Sales2). Create a notebook to conduct the following tests, and submit the results and ipynb / html files. In all cases, the significance level is 5%.

1. Regarding all data, could we conclude that sales are increased comparing Sales2 after the campaign with Sale1 before the campaign.
2. Could we conclude that sales in Sales2 between Tokyo and Saitama are significantly different.
3. Create a cross tabulation with prefectures and store types, and determine if there is an association between prefectures and store types (using chi-squared test)
4. (Adv) May we conclude that the ratio of the number of stores in each prefecture is consistent to the ratio of population of each prefecture (M stands for 1 million, Tokyo:14.0M, Saitama:7.3M, Chiba:6.3M, Kanagawa:9.2M) ?

```
ct = pd.crosstab(df['Pref'], df['Type'])
display(ct)
```

Type Pref	direct	franchise
Chiba	13	15
Kanagawa	28	30
Saitama	24	25
Tokyo	34	31

```
chi2, p, dof, expected = ss.chi2_contingency(ct)
print(chi2, p, dof, expected)
```

```
0.35072743456882416 0.9502219446727902 3 [[13.86  14.14 ]
 [28.71  29.29 ]
 [24.255 24.745]
 [32.175 32.825]]
```

expected

Type Pref	direct	franchise
Chiba	13.86	14.14
Kanagawa	28.71	29.29
Saitama	24.255	24.745
Tokyo	32.175	32.825

Expected table if "Pref" column and "Type" column are independent

How to obtain "expected"

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.contingency.expected_freq.html#scipy.stats.contingency.expected_freq

```
ct = pd.crosstab(df['Pref'],
display(ct)
```

Type / Pref	direct	franchise
Chiba	13	15
Kanagawa	28	30
Saitama	24	25
Tokyo	34	31
	99	101

If the Type column is independent of the Prep column, then the distribution should be proportional to the total number of direct/franchise, regardless of the value of the Prep.

proportionally divided (99 : 101)

28 → 13.86, 14.14
58 → 28.71, 29.29
49 → 24.255, 24.745
65 → 32.175, 32.825

Type / Pref	direct	franchise
Chiba	13.86	14.14
Kanagawa	28.71	29.29
Saitama	24.255	24.745
Tokyo	32.175	32.825

suppl. on expected frequency: computing elements by assuming rows & columns are independent

Prefs &Types are independent → P(Pref ∩ Type)=P(Pref)P(Type)

✓ e.g. P(Chiba ∩ direct)=P(Chiba)P(direct)
P(Chiba), P(direct) can be computed by the marginal totals of the row & col,

$$P(\text{Chiba}) = 28/200, P(\text{direct}) = 99/200$$

Since the element is frequency(not probability), the all total count is multiplied;
(28/200)*(99/200)*200=28*99/200=13.86

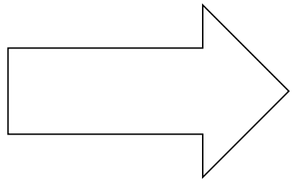
Generally, (i,j)-element is

$$\frac{T_i}{N} \times \frac{T_j}{N} \times N = \frac{T_i T_j}{N}$$

with T_i :i-th row's total, T_j:j-th column's total, N: all count

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare_print.html

Pref	direct	franchise	All
Chiba	???		28
Kanagawa	?		58
Saitama			49
Tokyo			65
All	99	101	200

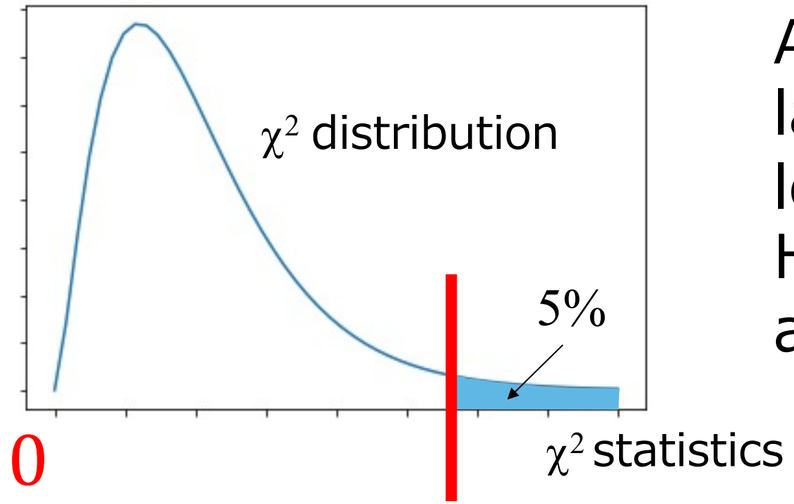


Pref	direct	franchise	All
Chiba	13.860	14.140	28
Kanagawa	28.710	29.290	58
Saitama	24.255	24.745	49
Tokyo	32.175	32.825	65
All	99	101	200

Test

chi2, p_val, dof, expected = ss. chi2_contingency(DataFrame)

- If the Pref and Type columns are independent, it is known that the chi2 statistic, which can be calculated from the deviation from the actual cross-tabulations (4 rows and 2 columns) and the expected table, follows a chi-square distribution of (4-1) * (2-1) = 3 degrees of freedom
  - χ^2 statistics: For each cell in the table, calculate the square of the difference between the value of expected table and that of the actual cross-tabulation divided by the value of actual cross-tabulation, and then add them together for all cells in the table.



As with the t (Welch) test, if the χ^2 statistics is largrer than the critical value (red line) at the 5% level of significance in the χ^2 distribution, we reject H0 (Prep and Type columns are independent) and accept H1 (Prep and Type columns are related).