# Analysis of Categorical Data

**Dr. Supaporn Erjongmanee**

Department of Computer Engineering
Kasetsart University
fengspe@ku.ac.th



1

## Outline

- Analysis of Categorical Data
  - Introduction
  - Homogeneity test
  - Independence test



2

## Introduction

- A study of data in categories
- Case: Population $I$ of interest; Each population is separated into $J$ categories
  - Example: 3 department stores vs. 5 payment methods (cash, check, store credit card, Visa, Mastercard)
- Homogeneity (Hypothesis) Test
  - Proportions of all categories in each population are the same





## Introduction (cont.)

- In general, data are put in the table
- Let $n_{ij}$ = number of samples in (i,j) category
- Table contains $\{n_{ij}\}$'s is called two-way contingency table

| | 1 | 2 | … | j | … | $J$ |
|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | … | $n_{1j}$ | … | $n_{1J}$ |
| 2 | $n_{21}$ | | | | | |
| … | … | | | | | |
| i | $n_{i1}$ | | | $n_{ij}$ | | |
| … | … | | | | | |
| $I$ | $n_{I1}$ | | | | | $n_{IJ}$ |

## Outline

- Analysis of Categorical Data
  - Introduction
  - Homogeneity test
  - Independence test





## Homogeneity Test

- Population *I* of interest; Each population is separated into *J* categories

- Let
  - $n_{ij}$ = number of samples in (i,j) category
  - $n_j$ = number of samples in j category = $\sum_i n_{ij}$
  - $n_i$ = number of samples in i population = $\sum_j n_{ij}$
  - n = number of all samples = $\sum_i \sum_j n_{ij}$
  - $p_{ij}$ = proportions of samples in (i,j) category
- Hypothesis test
  - Null hypothesis ($H_0$): $\boxed{p_{1j} = p_{2j} = \ldots = p_{Ij}}$
    - Proportion of samples in j category for each population is the same
  - Alternative hypothesis ($H_a$): $H_0$ is not true

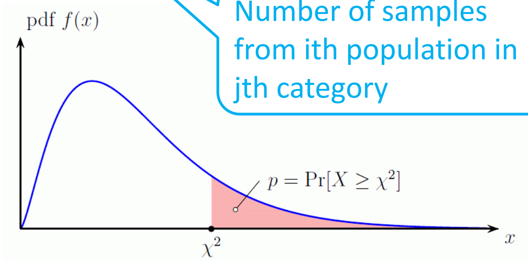
## Homogeneity Test (cont.)

P(samples in category j)

- Let $\hat{e}_{ij}$ = expected number of samples = $n_i p_j = n_i \frac{n_j}{n}$

Number of samples from ith population in jth category

- Test statistic
  - $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$
- Rejection region
  - $\chi^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$




- In each row i, there are J cells but $n_i = \sum_j n_{ij}$ is fixed. Hence, d.f.per row = J-1. There are I rows. Thus, sum of d.f. from all rows = I(J-1)
- In addition, we estimate $p_1, p_2, \ldots, p_J$ with $\sum_i p_i = 1$. There are J-1 parameters to estimate.
- At the end, resulting d.f. = I(J-1) - (J-1) = (I-1)(J-1)



7

## Example

- A can food company have three product sizes; each size is produced at different production lines
- Test in nonconformity of cans at significance level 0.5
  - Blemish, Crack, Improper pull tab location, Missing pull tab, Others

| | | Nonconformity | | | | | |
|---|---|---|---|---|---|---|---|
| | | Blemish | Crack | Location | Missing | Others | Sample size |
| Production line | 1 | 34 | 65 | 17 | 21 | 13 | 150 |
| | 2 | 23 | 52 | 25 | 19 | 6 | 125 |
| | 3 | 32 | 28 | 16 | 14 | 10 | 100 |
| | Total | 89 | 145 | 58 | 54 | 29 | 375 |

$n_i$

$n_J$



8

## Example (cont.)

- Hypothesis
  - $H_0$: All production lines are homogeneous in term of nonconformity categories (Blemish, Crack, Improper pull tab location, Missing pull tab, Others)
    - I = number of production lines = 3
    - J = types of nonconformity = 5
    - That is we test whether $\boxed{p_{1j} = p_{2j} = p_{3j} \text{ for } j = 1, 2, \ldots, 5}$
  - $H_a$: Production lines are not homogeneous



## Example (cont.)

- Find $\hat{e}_{ij}$ = expected number of samples = $\boxed{n_i \frac{n_j}{n}}$

| | | $\hat{e}_{ij}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | Blemish | Crack | Location | Missing | Others | Sample size |
| Production line | 1 | $\frac{150(89)}{375}$ =35.60 | $\frac{150(145)}{375}$ =58.00 | $\frac{150(58)}{375}$ =23.20 | $\frac{150(54)}{375}$ =21.60 | $\frac{150(29)}{375}$ =11.60 | 150 |
| | 2 | $\frac{125(89)}{375}$ = 29.67 | 48.33 | 19.33 | 18.00 | 9.67 | 125 |
| | 3 | $\frac{100(89)}{375}$ = 23.73 | 38.7 | 15.47 | 14.40 | 7.73 | 100 |
| | Total | 89 | 145 | 58 | 54 | 29 | 375 |

$n_i$

$n_j$

## Example (cont.)

- Find test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$

| | | $\frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$ | | | | |
|---|---|---|---|---|---|---|
| | | Blemish | Crack | Location | Missing | Others |
| Production line | 1 | $\frac{(34-35.60)^2}{35.60}$ = 0.072 | $\frac{(65-58.00)^2}{58.00}$ = 0.845 | $\frac{(17-23.20)^2}{23.20}$ = 1.657 | $\frac{(21-21.60)^2}{21.60}$ = 0.017 | $\frac{(13-11.60)^2}{11.60}$ = 0.169 |
| | 2 | $\frac{(23-29.67)^2}{29.67}$ =1.498 | 0.278 | 1.661 | 0.056 | 1.391 |
| | 3 | $\frac{(32-23.73)^2}{23.73}$ = 2.879 | 2.943 | 0.018 | 0.011 | 0.664 |

- Test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} = 14.159$



11

## Example (cont.)

- Test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} = 14.159$
- Find rejection region:
  - Degree of freedom = (I-1) (J-1) = (3-1)(5-1) = (2)(4) = 8
  - $\chi^2_{0.05,8}$= 15.507
- Thus, we do not reject hypothesis at α = 0.05
- At significance level = 0.05, all production lines are homogeneous in term of nonconformity categories



12

## Example (cont.)

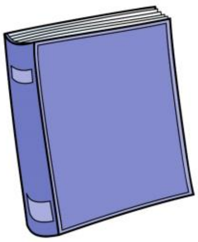
- Test statistic = $\sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 14.159$
- Find p-value
  - Degree of freedom = (I-1) (J-1) = (3-1)(5-1) = (2)(4) = 8
  - P-Value = 0.077
- Thus, we do not reject hypothesis since p-value > α = 0.05
- At significance level = 0.05, all production lines are homogeneous in term of nonconformity categories





## Example 2

- Compare two books whether they were written by the same author or not



- How to compare these two books?

*Image Source: http://www.clipartpanda.com/categories/school-book-clipart*

## Example 2

- Compare whether the frequencies of words in three of Austen's works are the same

| Word | Sense and Sensibility | Emma | Sandition |
|---|---|---|---|
| a | 147 | 186 | 101 |
| an | 25 | 26 | 11 |
| this | 32 | 39 | 15 |
| that | 94 | 105 | 37 |
| with | 59 | 74 | 28 |
| without | 18 | 10 | 10 |

- Test homogeneity



## Outline

- Analysis of Categorical Data
  - Introduction
  - Homogeneity test
  - Independence test

## Introduction

- A study of data in categories
- Case: Single population with two factors; One factor with *I* categories, and the other factor with *J* categories
  - Example: One department store, 6 departments (male clothes, female clothes, children, cosmetics, shoes, grocery) vs. 5 payment methods (cash, check, store credit card, Visa, Mastercard)
- Independence Test
  - Two factors occur independently



17

## Introduction (cont.)

- In general, data are put in the table
- Let $n_{ij}$ = number of samples in (i,j) category
- Table contains $\{n_{ij}\}$'s is called two-way contingency table

| | 1 | 2 | … | j | … | *J* |
|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | … | $n_{1j}$ | … | $n_{1J}$ |
| 2 | $n_{21}$ | | | | | |
| … | … | | | | | |
| i | $n_{i1}$ | | | $n_{ij}$ | | |
| … | … | | | | | |
| *I* | $n_{I1}$ | | | | | $n_{IJ}$ |



18

## Independence Test

- Single population with <u>two factors</u>; One factor with $I$ categories, and the other factor with $J$ categories

- Let
  - $n_{ij}$ = number of samples in (i,j) category
  - $n_j$ = number of samples in j category = $\sum_i n_{ij}$
  - $n_i$ = number of samples in i category = $\sum_j n_{ij}$
  - n = number of all samples = $\sum_i \sum_j n_{ij}$
  - $p_{ij}$ = proportions of samples in (i,j) category
- Hypothesis test
  - Null hypothesis ($H_0$): $\boxed{p_{ij} = p_i\, p_j}$
    - Proportion of samples in categories i and j are independent
  - Alternative hypothesis ($H_a$): $H_0$ is not true

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

19

## Independence Test (cont.)

If two factors are independent, $p_{ij} = p_i p_j$

- Let $\hat{e}_{ij}$ = expected number of samples = $np_{ij} = np_i p_j = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n}$

Derivation of $\hat{e}_{ij}$ is different from Homogeneity test

Same $\hat{e}_{ij}$ as Homogeneity Test

- Test statistic
  - $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$
- Rejection region
  - $\chi^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

20

## Example

- Study of gasoline station condition and aggressiveness in gasoline pricing
- Two factors: gasoline station condition (modern, standard, sub-standard) vs. aggressiveness in pricing (aggressive, neutral, nonaggressive)
- Test whether two factors are independent of each other at significance level = 0.01

| | | Aggressiveness in pricing | | | |
|---|---|---|---|---|---|
| | | Aggressive | Neutral | Non Aggressive | Sample Size |
| Condition | Substandard | 24 | 15 | 17 | 56 |
| | Standard | 52 | 73 | 80 | 205 |
| | Modern | 58 | 86 | 36 | 180 |
| | Total | 134 | 174 | 133 | 441 |



## Example (cont.)

- Hypothesis
  - $H_0$: Gasoline station condition and aggressiveness in pricing are independent
    - I = number of conditions = 3
    - J = levels of pricing aggressiveness = 3
    - We test on $p_{ij} = p_i \, p_j$
  - $H_a$: Gasoline station condition and aggressiveness in pricing are not independent

## Example (cont.)

- Find $\hat{e}_{ij}$ = expected number of samples = $\boxed{\frac{n_i n_j}{n}}$

| | | $\hat{e}_{ij}$ | | | |
|---|---|---|---|---|---|
| | | Aggressive | Neutral | Non Aggressive | Sample Size |
| Condition | Substandard | $\frac{56(134)}{441}$ =17.02 | $\frac{56(174)}{441}$ =22.10 | $\frac{56(133)}{441}$ =16.89 | 56 |
| | Standard | $\frac{205(134)}{441}$ =62.29 | 80.88 | 61.83 | 205 |
| | Modern | $\frac{180(134)}{441}$ =54.69 | 71.02 | 54.29 | 180 |
| | Total | 134 | 174 | 133 | 441 |

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

23

## Example (cont.)

- Find test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$

| | | $\frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$ | | |
|---|---|---|---|---|
| | | Aggressive | Neutral | Non Aggressive |
| Condition | Substandard | $\frac{(24-17.02)^2}{17.02}$ = 2.867 | $\frac{(15-22.10)^2}{22.10}$ = 2.278 | $\frac{(17-16.89)^2}{16.89}$ = 0.001 |
| | Standard | $\frac{(52-62.29)^2}{62.29}$ = 1.700 | 0.769 | 5.343 |
| | Modern | $\frac{(58-54.69)^2}{54.69}$ = 0.200 | 3.160 | 6.160 |

- Test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$ $\boxed{= 22.476}$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

24

## Example (cont.)

- Test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} = \boxed{22.476}$
- Given α = 0.01, find p-value
  - Degree of freedom = (I-1) (J-1) = (3-1)(3-1) = 4
  - P-value 0.00016
- P-value < α = 0.01 => Null hypothesis is rejected
- Gasoline station condition and aggressiveness in pricing are dependent

$\chi^2_{0.01,4}$=13.277

```python
from scipy.stats import chi2
1-chi2.cdf(22.476,4)
0.0001611050155756466
```

Supaporn Erjongmanee
fengspe@ku.ac.th
Analysis of Categorical Data
Slide 25
Department of Computer Engineering
Kasetsart University

25

## Example (cont.)

- Test statistic = $\sum_i \sum_j \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} = \boxed{22.476}$
- Given α = 0.01, find rejection region
  - Degree of freedom = (I-1) (J-1) = (3-1)(3-1) = 4
  - Thus, $\chi^2_{0.01,4}$=13.277
- Null hypothesis is rejected
- Gasoline station condition and aggressiveness in pricing are dependent

Supaporn Erjongmanee
fengspe@ku.ac.th
Analysis of Categorical Data
Slide 26
Department of Computer Engineering
Kasetsart University

26

## Example 2

- Is there a relationship between marital status and educational level?

| Education | Married once | Married more than once |
|---|---|---|
| College degree | 550 | 61 |
| No college degree | 681 | 144 |

- Test independency





## References

1. J.L. Devore and K.N.Berk, Modern Mathematical Statistics with Applications, Springer, 2012.
2. J.A. Rice, Mathematical Statistics and Data Analysis, Duxbury Press, 1995.