Analysis of Categorical Data

Dr. Supaporn Erjongmanee

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

Outline

- P-Value
- Analysis of Categorical Data
  - Introduction
  - Homogeneity test
  - Independence test
  - Examples

P-value

- **Smallest** significance level at which null hypothesis is rejected
- Also call observed significance level (OSL)
- Think of P-value as area under the curve

Example

- Nicotine level in cigarette is normally distributed
  - Average nicotine level = μ = 1.5, σ = 0.2
- Customer wants to check nicotine level
  - H_0: $\mu = 1.5$
  - H_a: $\mu > 1.5$
- If test statistic z = 2.10, then
  - α = 0.1, z_α = 1.2816: $z > z_\alpha$ => reject H_0
  - α = 0.05, z_α = 1.6449: $z > z_\alpha$ => reject H_0
  - α = 0.01, z_α = 2.3263: $z < z_\alpha$ => do not reject H_0

What's smallest α to reject H_0?

Goal is to minimize rejection region α

Example: P-value for z-test

α decreases as z_α increases

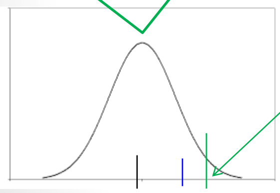
- **Upper-tailed test case**

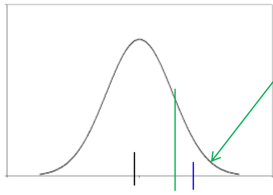
Goal is to minimize rejection region α

Equivalently, find largest z_α that results to reject H_0

Let z = test statistic

When $z < z_\alpha$, then H_0 is not rejected





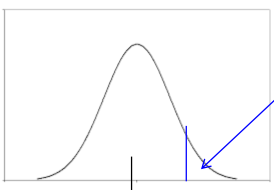
If $z_\alpha < z$, then H_0 is rejected but α is not minimum still.

α is minimum when $z_\alpha = z$.

Example: P-value for z-test

P-value = Smallest significance level at which H_0 is rejected

- **Upper-tailed test case (cont.)**



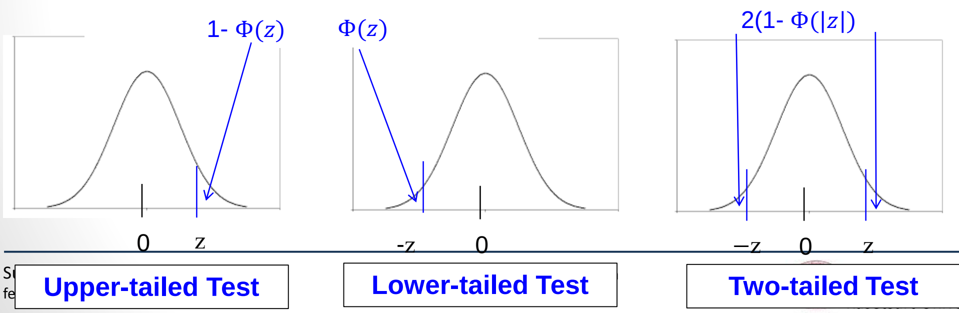
- Our goal is to minimize α
- Minimum α occurs at critical value z_α = test statistic z
- Thus, **P-value = 1 - $\Phi(z)$**

P-value for z-test (cont.)

- Null hypothesis: $\mu = \mu_0$

Alternative Hypothesis	P-value	Test
$H_a : \mu > \mu_0$	$1 - \Phi(z)$	**Upper-tailed test**
$H_a : \mu < \mu_0$	$\Phi(z)$	**Lower-tailed test**
$H_a : \mu \neq \mu_0$	$2(1-\Phi(\|z\|))$ or $2(\Phi(-\|z\|))$	**Two-tailed test**

Z = test statistic



Example

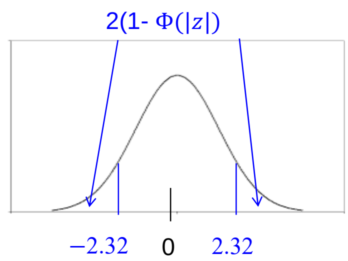
- Target thickness of silicon wafer = 245 µm
- 50 wafers are sampled and collected for thickness
  - Sample mean = $\bar{X} = 246.18$ µm
  - Sample standard deviation = S = 3.60 µm
- Question: What is p-value to reject H_0?
- Our goal is to check wafer thickness level
  - μ = average wafer thickness
  - $\mu_0 = 245$
  - H_0: $\mu = 245$
  - H_a: $\mu \neq 245$

Example (cont.)

H_0: μ = 245
H_a: μ ≠ 245

- Test statistic:

$$Z = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{246.18-245}{3.60/\sqrt{50}} = 2.32$$



- This is two-tailed test
  - **P-value** = 2(1- Φ(|z|)

$$= 2\,(1 - \Phi(|2.32|)$$

$$= 2\,(1 - 0.9898) = 0.0204$$

Question:
Given α = 0.01 and p-value = 0.0204, do we reject H_0 ?

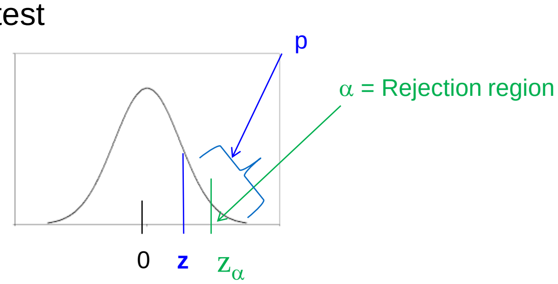
Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example (cont.)

Question:
Given α = 0.01 and p-value = 0.0204, do we reject H_0 ?

Consider upper-tailed test



If p > α, then test statistic z does not fall in rejection region.

Do not reject H_0

H_0 is rejected when p < α

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

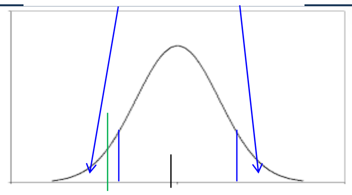
Example (cont.)

H_0: μ = 245
H_a: μ ≠ 245

- Test statistic:

$$Z = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{246.18-245}{3.60/\sqrt{50}} = 2.32$$

- This is two-tailed test
  - P-value = 2(1- Φ(|z|) = 2 (1 – Φ(|2.32|) = 0.0204

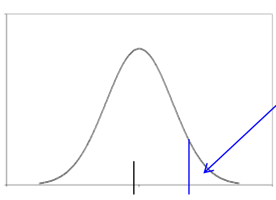


Given α = 0.01 and p-value = 0.0204, do we reject H_0 ?

  - Given α = 0.01 < p-value = 0.0204
    - Test statistic falls outside rejection region for α /2
    - Null hypothesis is not rejected
    - At significance level = 0.01, wafer thickness is not different from the target value

Example: P-value for t-test

- Similar to z-test
- **Upper-tailed test case**:

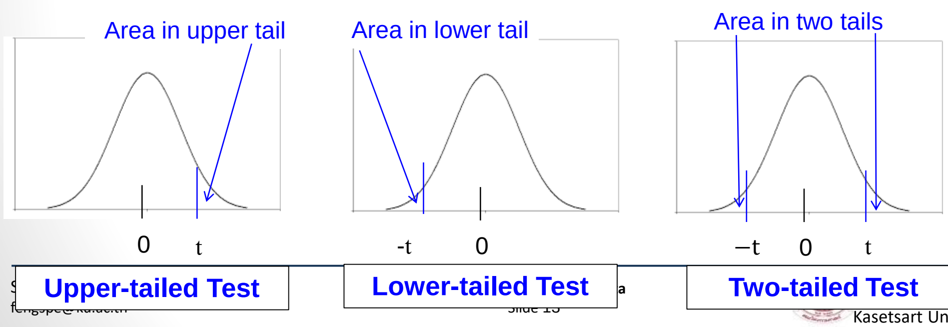


- Our goal is to minimize α
- Minimum α occurs at critical value $t_{\alpha,\ df}$ = test statistic **t**
- Thus, **P-value =** area in upper tail of test statistic **t**

P-value for t-test (cont.)

- Null hypothesis: $\mu = \mu_0$

Alternative Hypothesis	P-value	Test
$H_a : \mu > \mu_0$	Area in upper tail of test statistic t	**Upper-tailed test**
$H_a : \mu < \mu_0$	Area in lower tail of test statistic t	**Lower-tailed test**
$H_a : \mu \neq \mu_0$	Area in two tails of test statistic t	**Two-tailed test**



Example

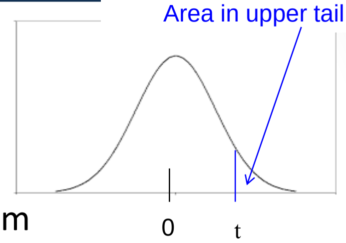
- Our goal is to check fuel efficiency whether it is better than average = 20 mpg
- Collect fuel efficiency (miles per gallon (mpg)) of 4 cars
  - x_1 = 20.830, x_2 = 22.232, x_3 = 20.276, x_4 = 17.718
  - Sample mean = $\bar{X}$ = 20.264 mpg
  - Sample standard deviation = s = 1.8864 mpg
- Question: What is p-value to reject claim ?
- Set up hypothesis
  - μ = average fuel efficiency
  - $\mu_0 = 20$
  - H_0: $\mu = 20$
  - H_a: $\mu > 20$

Example (cont.)

Upper-tailed Test

- Test statistic:

$$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{20.264-20}{1.8864/\sqrt{4}} = 0.2799$$



- This is upper-tailed test with 3 degree of freedom
  - P-value = area on the right of t = 0.2799

    = 1- 0.6011 = 0.3989

Given α = 0.05 and p-value = 0.3989, do we reject H_0 ?

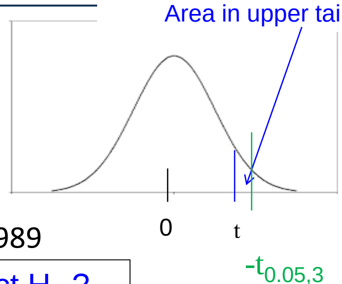
* *Tool: http://stattrek.com/online-calculator/t-distribution.aspx*

Example (cont.)

Upper-tailed Test

- Test statistic:

$$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{20.264-20}{1.8864/\sqrt{4}} = 0.2799$$



- This is upper-tailed test with 3 degree of freedom
  - P-value = area on the right of 0.2799 = 1- 0.6011 = 0.3989

Given α = 0.05 and p-value = 0.3989, do we reject H_0 ?

- Given α = 0.05 < p-value = 0.3989,
  - Test statistic falls outside rejection region for α
  - H_0 is not rejected
  - At significance level = 0.05, fuel efficiency is 20 mpg

Outline

- P-Value
- Analysis of Categorical Data
  - Introduction
  - Homogeneity test
  - Independence test
  - Examples

Introduction

- A study of data in categories
- 2 cases:
  1. Population I of interest; Each population is separated into J categories
     - Example: 3 department stores vs. 5 payment methods (case, check, store credit card, Visa, Mastercard)
  2. Single population with two factors; One factor with I categories, and the other factor with J categories
     - Example: One department store, 6 department vs. 5 payment methods (case, check, store credit card, Visa, Mastercard)

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 <u>two-way contingency table</u>

	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}

Introduction (cont.)

- 2 cases:
  1. <u>Population *I*</u> of interest; Each population is separated into <u>*J* categories</u>
  2. Single population with two factors; One factor with <u>*I* categories</u>, and the other factor with <u>*J* categories</u>
- Hypothesis test
  1. Proportion of all categories in each population are the same
     - Homogeneity test
  2. Two factors occur independently
     - Independence test

Outline

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

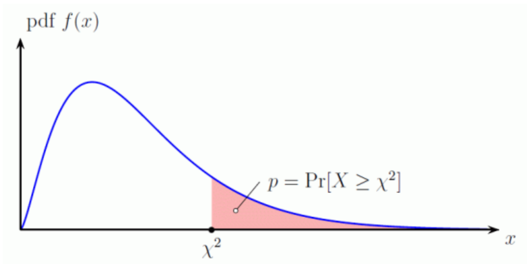
Homogeneity Test

1. 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 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): $p_{1j} = p_{2j} = \ldots = p_{Ij}$
    - Proportion of samples in j category for each population is the same
  - Alternative hypothesis (H_a): H_0 is not true

Homogeneity Test (cont.)

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

Example

- A can food company have three product sizes; each size is produced at different production lines
- Test in nonconformity of cans
  - 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

Example (cont.)

- Hypothesis
  - H_0: All production lines are homogeneous in term of nonconformity categories
    - I = number of production lines = 3, J = types of nonconformity = 5
    - That is we test whether $p_{1j} = p_{2j} = p_{3j}$ for j = 1, 2, …, 5
  - H_a: Production lines are not homogeneous

Example (cont.)

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

		$\hat{e}_{ij}$					
		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

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

Example (cont.)

- Test statistic = $\sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 14.159$
- Degree of freedom = (I-1) (J-1) = (3-1)(5-1) = (2)(4) = 8
- P-Value = 0.077
- Thus, we reject hypothesis at α = 0.1, but not α = 0.05 or 0.01
- At significance level = 0.05, all production lines are homogeneous in term of nonconformity categories

Outline

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

Independence Test

2. Single population with two factors; 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

Independence Test (cont.)

- Let $\hat{e}_{ij}$ = expected number of samples = $np_i p_j = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n}$
- Test statistic
  - $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$
- Rejection region
  - $\chi^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

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

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

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

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

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

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

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

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

Department of Computer Engineering
Kasetsart University

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.