Feature Selection Day 2

12 April 2022 07:03

Gender	med A	med B	Total
M	2 2.45	7 6.54	9
F	4 3.54	9 9.45	13
	6	16	22

H_0 = there is no association between Gender & med

H_a = there is association between Gender & med.

Exp freq = (row total x column total) / grand total

E_1 = $(9 \times 6)/22 = 2.45$

E_2 = $(13 \times 6)/22 = 3.54$

E_3 = $(9 \times 16)/22 = 6.54$

E_4 = $(13 \times 16)/22 = 9.45$

21.98 ≃ 22

pvalue = 1.0 > 0.05 (α)

H_0 = Accept

Chi-2

$$\chi_c^2 = \frac{(O-E)^2}{E}$$

c = degree of freedom

O = observed value

E = Exp freq

Use it as numbers. not as percentage and proportion

Discrete variable.
ordinal var

Gender	meal A	meal B	meal C	total
m	74 91	22 20.1	42 26.9	138
f	102 85	17 18.8	10 25.1	129
	176	39	52	267

$Df = (\text{no. of rows} - 1)(\text{no. col.} - 1)$
$= (2-1)(3-1)$
$= 1 \times 2$
$= 2$

$E_1 = (138 \times 176)/267 = 91$
$E_2 = (129 \times 176)/267 = 85$
$E_3 = (138 \times 39)/267 = 20.1$
$E_4 = (129 \times 39)/267 = 18.8$
$E_5 = (138 \times 52)/267 = 26.9$
$E_6 = (129 \times 52)/267 = 25.1$

$\chi_1 = (74-91)^2/91 = 3.1$
$\chi_2 = (102-85)^2/85 = 3.4$
$\chi_3 = (22-20.1)^2/20.1 = 0.2$
$\chi_4 = (17-18.8)^2/18.8 = 0.17$
$\chi_5 = (42-26.9)^2/26.9 = 8.4$
$\chi_6 = (10-25.1)^2/25.1 = 9.08$

24.31

df : 2
$\alpha = 0.05$

24.31 > 5.991

H_0 = reject p_value < 0.05

H_a = accept

chi is high :-
1) feature is important
2) p_value will be low

chi is low :-
1) feature is not important
2) p_value value will be high.

Missing value Ratio.

Ratio # missing value = $\frac{\text{no. of missing}}{\text{total}} \times 100$

Variance Threshold. Method

ANOVA Test

Gender	wt	ward
M	40	A
F	45	B
M	54	A
F	36	A

M = 94
F = 81

One sample :- percentage distribution

M = 50 %
F = 50 %

H_0 = mean of male 47
H_a = mean is not 47

M = 60 %
F = 50 %

Ha = mean is not 47

Two sample :- Cat VS Num

when we have 2 class in cat var (Gender)
(M, F)

& one continuous ⇒ 2 Sample t-test

&

when we have more than 2 class in cat. var

& cont. cont ⇒ ANOVA test

Z-test : # of sample > 30 | std. d. is known
n > 30

ttest :- n < 30 | std. d. is unknown.

	Cat	Cont
Cat	chi2 test	t-test, z-test, ANOVA.
Cont		corr, Var inf.