Chapter 10 : Hypothesis testing for categorical data

* categorical data means qualitative data
* testing 2 sample proportions

Method (1)

اخطوات :

① $H_0 : P_1 = P_2$ vs $H_1 : P_1 \neq P_2$

② Test Stat

→ Z

→ Z_{conv}

$$Z = \frac{(\hat{P}_1 - \hat{P}_1) - (P_1 - P_2)}{\sqrt{P^* \cdot q^* \cdot \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$Z_{conv} = \frac{|P_1 - \hat{P}_2| - \left(\frac{1}{2n} + \frac{1}{2m}\right)}{\sqrt{P^* q^* \left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$(P_1 - P_2)$ = zero

P^* (Pooled proportion) = $\frac{x+y}{m+n}$

$q^* = 1 - p^*$

n : Sample (1) size

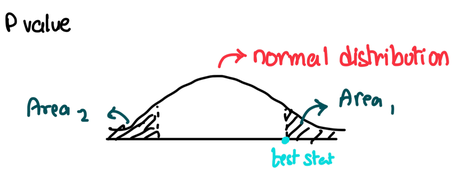
m : Sample (2) size

$x = nP_1$

$y = nP_2$

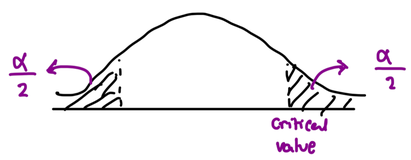
③ → P value

→ Rejection Region



Pvalue = 2 x Area

Pvalue > α ⟹ accept H_0



* test stat > critical value ⟹ reject H_0

Method (2) : contingency table

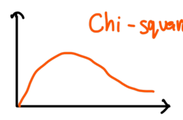
① H_0 vs H_1

$H_0 : P_1 = P_2$ ⟹ variables are independent

$H_1 : P_1 \neq P_2$ ⟹ variables are dependent

② the main aim of this table is to summarize a large set of data

③ test stat : χ^2 or χ^2_{conv}



→ * all values are positive
* It is right skewed

Observed table

			Total
	O_{11}	O_{12}	row margin
	O_{21}	O_{22}	row margin
Total	column margi	column margi	Grand Total

$E = \frac{R * C}{\text{Grand Total}}$

Expected table

			Total
	E_{11}	E_{12}	
	E_{21}	E_{22}	
Total			Grand Total

R x C table

* R : number of rows
  C : number of columns
* test stat → $\chi^2 = \sum \frac{(O-E)^2}{E}$
* d.f = (R-1)(C-1)
* Conditions of expected table
  1) All expected values > 1
  2) Only $\frac{1}{5}$ of All cells have expected value < 5

2*2 table

Special case from R*C table

* R = 2
  C = 2
* test stat → $\chi^2 = \sum \frac{(O-E)^2}{E}$

→ $\chi^2_{conv} = \sum \frac{(|O-E| - \frac{1}{2})^2}{E}$ → Yates corrected chi squared

* Conditions:
  Always expected values > 5

Goodness of fit test

① Approximation of discrete random variable to continuous random variable

A
- $P(X < a) \rightarrow P(X < a-1)$
- $P(X > a) \rightarrow P(X > a+1)$

B
- $P(X \geq b) \rightarrow P(X \geq b-0.5)$
- $P(X \leq b) \rightarrow P(X \leq b+0.5)$
- $P(a \leq X \leq b) \rightarrow P(a-0.5 \leq X \leq b+0.5)$

② Expected = $\frac{\text{Grand Total}}{\text{Proportion}}$

③ $\chi^2 = \sum \frac{(O-E)^2}{E}$

④ $d.f = g - k - 1$

g → number of groups

k → number of estimators

* if we reject H_0 → normal method doesn't provide an adequate fit to the data