Chi-squared Test for Independence

χ^2 test for independence is a hypothesis test to test for independence between variables.

If A is independent of B

$P(A|B) = P(A)$ $\quad$ $P(A \text{ and } B) = P(A)P(B)$

$P(A \text{ and } B) = P(\cancel{A|B})P(B)$

$P(A)$

Consider the following:

A 2020 study of employees at two academic hospitals in Philadelphia.

The study was interested in vaccine hesitancy.

One of the other variables measured was risk of Covid-19, as either high, medium, or low.

Question: Is vaccine hesitancy independent of workplace risk?

$n = 5608$

Work Place Risk	Not Hesitant	Hesitant	
High	1230	1019	45.3% 2249
Medium	1684	1464	46.5% 3148
Low	106	105	49.8% 211
	3020	2588	

$VH \perp\!\!\!\perp WR$

H_0: Vaccine hesitancy is independent of workplace risk.

H_a: Vaccine hesitancy is not independent of workplace risk.

$VH \not\perp\!\!\!\perp WR$

If H_0 is true, we expect the # of individuals in (Eg.) High risk and hesitant group to be proportional to

$P(HR \text{ and Hesitant}) = P(HR) \times P(H)$

and hesitant group to be proportional to

Expected counts $P(\text{HR and Hesitant}) = P(\text{HR}) \times P(H)$

Risk	Not Hesitant	Hesitant	totals
H	$\frac{2249 \times 3020}{5608} = 1211.1$	$\frac{2249 \times 2588}{5608} = 1037.9$	2249
M	$\frac{3148 \cdot 3020}{5608} = 1695.3$	$\frac{3148 \cdot 2588}{5608} = 1452.8$	3148
L	$\frac{211 \cdot 3020}{5608} = 113.6$	$\frac{211 \cdot 2588}{5608} = 97.4$	211
	3020	2588	5608

Calculate Expected Count

Expected prop who are High risk + not hesitant

Sample size

$= P(\text{high risk}) \times P(\text{not hesitant}) \times N$

$N = 5608$ $\quad P(\text{High risk}) = \frac{\#\text{ high risk}}{N}$

$P(\text{not hesitant}) = \frac{\#\text{ not hesitant}}{N}$

Expected count High Risk + not hesitant

$$= \frac{\#\text{ high risk} \times \#\text{ not hesitant}}{N} = \hat{p}_{\text{high risk}} \times \hat{p}_{\text{not hesitant}} \times N$$

In general expected count for a cell

$$= \frac{\#\text{Row total} \times \#\text{Column total}}{\#\text{Grand Total}}$$

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

$$= \sum_{\text{every cell}} \frac{(O-E)^2}{E}$$

high risk + not hesitant

high risk + hesitant

$$\chi^2 = \frac{(1230 - 1211.1)^2}{1211.1} + \frac{(1019 - 1037.9)^2}{1037.9} + \dots + \dots$$

$$+ \dots + \frac{(105 - 97.4)^2}{97.4}$$

low risk + hesitant

$\chi^2 = 1.9086$

test statistic

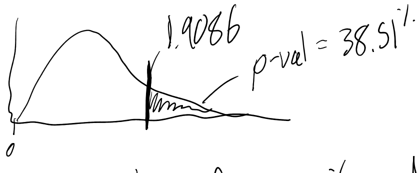
χ^2 distribution is defined by the degrees of freedom

For a 2-way table

$df = (\#\text{ rows} - 1)(\#\text{ of columns} - 1)$

$= (3-1)(2-1) = 2$

$$P(\chi^2 > 1.9086) = 0.3851$$



Our p-value of 38.5% indicates that there is not significant evidence against the independence of workplace risk and vaccine hesitancy. We fail to reject H_0.