2-Sample Categorical Data

BIOSTAT 201A Fall 2025

Discussion 7 – November 14, 2025

Cindy J. Pang

Contingency Tables	Two Sample Test for Difference in Population Proportions	Pearson's Chi-Square Test
Hypotheses	$H_0: p_1 = p_2 \iff p_1 - p_2 = 0$ $H_1: p_1 \neq p_2 \iff p_1 - p_2 \neq 0$	H_0: Group 1 and Group 2 are independent H_1: Group 1 and 2 are associated.
Assumptions	(1) p_1 and p_2 are independent. (2) Rule of Five satisfied for both proportions s.t. $n_1 p_1 (1-p_1) \geq 5$ and $n_2 p_2 (1-p_2) \geq 5$	All expected counts are ≥ 5 ($E_{ij} \geq 5$)
Data	Binary covariate $\begin{array}{l\|c\|c\|l} \text{Status} & \text{"success"} & \text{"failure"} & \text{Total} \\ \hline \text{Exposed} & x_1 & n_1 - x_1 & n_1 \\ \text{Not Exposed} & x_2 & n_2 - x_2 & n_2 \\ \hline & x_1 + x_2 & n_1 + n_2 - (x_1 + x_2) & n_1 + n_2 \end{array}$	Group 2 (j) $\begin{array}{cc\|c\|c\|l} & & 1 & 0 & \\ \hline \text{Group 1 (i)} & 1 & a & b & a+b \\ & 0 & c & d & c+d \\ \hline & & a+c & b+d & N = a+b+c+d \end{array}$ $E_{11} = \frac{(a+b)(a+c)}{N}$ $E_{10} = \frac{(a+b)(b+d)}{N}$ $E_{01} = \frac{(c+d)(a+c)}{N}$ $E_{00} = \frac{(c+d)(b+d)}{N}$
Test Statistic	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p (1 - \hat{p}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $\hat{p}_p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$	$\chi^2_{TS} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ where $E_{ij} = \frac{(\text{row total in } i)(\text{col total in } j)}{N}$ $= \frac{(a - E_{11})^2}{E_{11}} + \frac{(b - E_{10})^2}{E_{10}} + \frac{(c - E_{01})^2}{E_{01}} + \frac{(d - E_{00})^2}{E_{00}}$
Decision Rule	$\|z\| > z_{1-\alpha/2} \Rightarrow$ reject H_0	$\chi^2_{TS} > \chi^2_{df=(\#rows-1)(\#cols-1),\, 1-\alpha/2} \Rightarrow$ reject H_0.

A hypothesis suggests that breast cancer risk increases with a longer interval between the onset of menstruation and a woman's first childbirth, making age at first birth a potential risk factor. To test this, an international study examined women from hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Among women who had given birth, 21.2% of those with breast cancer (683 of 3,220) and 14.6% of those without breast cancer (1,498 of 10,245) had their first child at age 30 or older. The question is whether this difference reflects a true association or simply occurred by chance.

(a) Test this hypothesis using a 2-Sample Test for Difference in Population Proportions

$H_0: P_{BC,\ Age\ 30+} = P_{BC,\ Age<30}$ vs. $H_1: P_{BC,\ Age\ 30+} \neq P_{BC,\ Age<30}$

Data:

	Age 30+	Age < 30	
BC	$x_1 =$ 683	2,537	$n_1 =$ 3220
BC	$x_2 =$ 1,498	8,747	$n_2 =$ 10,245
	$x_1 + x_2 =$ 2,181	11,284	$n_1 + n_2 =$ 13,465

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{683}{3220} = 21.2\% = .212$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{1,498}{10,245} = 14.6\% = .146$

$\hat{p}_P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{2,181}{13,465} = 0.1619755$

Test Stat: $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_P(1-\hat{p}_P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{.212 - .146}{\sqrt{0.1619755(1-.1619755)\left(\frac{1}{3220}+\frac{1}{10245}\right)}} = 8.85 > z_{.975} = 1.96$$

$\boxed{\text{Conclusion:}}$ Since our z-statistic is greater than the 1.96 critical value, we reject the H_0 and conclude that this difference reflects a true association and there is a difference in breast cancer risk after the age of 30.

A hypothesis suggests that breast cancer risk increases with a longer interval between the onset of menstruation and a woman's first childbirth, making age at first birth a potential risk factor. To test this, an international study examined women from hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. Among women who had given birth, 21.2% of those with breast cancer (683 of 3,220) and 14.6% of those without breast cancer (1,498 of 10,245) had their first child at age 30 or older. The question is whether this difference reflects a true association or simply occurred by chance.

(b) Test this hypothesis using a Pearson's Chi-square Test. *Are your conclusions different from part a?*

H_0: Risk of Breast Cancer is independent of age vs. H_1: There is an association between Breast Cancer Risk and age.

Data:

	Age 30+	Age < 30	row totals
BC	a= 683	b= 2,537	3,220
$\overline{BC}$	c= 1,498	d= 8,747	10,245
col totals:	2,181	11,284	13,465

$$E_{BC, Age\ 30+} = \frac{(3220)(2181)}{13465} = 521.561$$

$$E_{BC, Age<30} = \frac{(3220)(11284)}{13465} = 2698.439$$

$$E_{\overline{BC}, Age\ 30+} = \frac{(10245)(2181)}{13465} = 1659.439$$

$$E_{\overline{BC}, Age<30} = \frac{(10245)(11284)}{13465} = 8585.561$$

$$\chi^2_{TS} = \frac{(683-521.561)^2}{521.561} + \frac{(2{,}537-2698.439)^2}{2698.439} + \frac{(1498-1659.439)^2}{1659.439} + \frac{(8747-8585.61)^2}{8585.61}$$

$$= 78.36984 > \chi^2_{1, .975} = 5.024$$

$\boxed{\text{Conclusion}}$: Since our $\chi^2_{TS} > \chi^2_{df=1, 0.975}$, we reject the Ho and conclude that there is an association between breast cancer risk and age, which matches our result from part (b).