Week 6

Comparison of Population Proportions

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Present basic inferential methods for categorical data; in particular, the analysis of binary data classified in two-way contingency tables.
- Compare the two population proportions:
  - Case-control pair-matched data with a single binary exposure:
    - We will use either z score test statistic or McNemar's chi-square test statistic.
  - Large independent samples, where the z score test statistic is used to test whether there are differences in the two population proportions or not.
- Introduce Mantel–Haenszel method to investigate the relationship between two binary variables, where the confounders are controlled.
- Present the chi-square test for independency for the general case of two-way contingency tables (data not necessarily binary).

Analysis of Pair-Matched Data (Binary Data Classified in Two-Way Contingency Tables)

- Applies to cases where each subject or member of a group is observed twice for the presence or absence of a certain characteristic (e.g., at admission to and discharge from a hospital).
- Also applies to matched pairs observed for the presence or absence of the same characteristic.
- A popular application is an epidemiological design called a pair-matched case–control study:
  - cases of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions,
  - controls are sampled either as disease-free persons from the population at risk or as hospitalized patients having a diagnosis other than the one under investigation.
- **Technique to control confounding factors**
  - Individual cases are matched, often one to one, to controls chosen to have similar values for confounding variables such as age, gender, and race.

Pair-Matched Data with a Single Binary Exposure

- For pair-matched data with a single binary exposure (e.g., smoking versus non-smoking), data can be represented by a 2×2 table where $(+, -)$ denotes the (exposed, nonexposed) outcome.
  - a denotes the number of pairs with two exposed members,
  - b denotes the number of pairs where the case is exposed but the matched control is unexposed,
  - c denotes the number of pairs where the case is unexposed but the matched control is exposed,
  - d denotes the number of pairs with two unexposed members.

TABLE 6.1

	Control	
Case	+	−
+	a	b
−	c	d

One-sided test (right tailed)
H_0: The exposure has nothing to do with the disease (there is no effect)
H_a: The exposure has a positive effect on the disease (there is a $+$ effect)
One-sided test (left tailed)
H_0: The exposure has nothing to do with the disease (no effect)
H_a: The exposure has a negative effect on the disease (there is a $-$ effect)
Two-sided test (two tailed)
H_0: The exposure has nothing to do with the disease (there is no effect)
H_a: The exposure has an effect on the disease (there is an effect)

McNemar's chi-square Test for Pair-Matched Data with a Single Binary Exposure

Compare the incidence of exposure among the cases versus the controls

- Under the null hypothesis that the exposure has nothing to do with the disease, the test statistic is either the standardized z score

$$Z = \frac{b - c}{\sqrt{b + c}} \sim N(0, 1)$$

or the McNemar's chi-square

$$X^2 = \frac{(b - c)^2}{b + c} \sim \chi^2(df = 1)$$

  - If the test is one-sided, z score is used and the null hypothesis is rejected at the 0.05 level when $|z| \geq 1.65$.
  - If the test is two-sided, X^2 is used and the null hypothesis is rejected at the 0.05 level when $\mathrm{X}^2 \geq 3.84$.

Example 6.2

- It has been noted that metal workers have an increased risk for cancer of the internal nose and paranasal sinuses, perhaps as a result of exposure to cutting oils.
- Therefore, a study was conducted to see whether this particular exposure also increases the risk for squamous cell carcinoma of the scrotum.
- Cases included all 45 squamous cell carcinomas of the scrotum diagnosed in Connecticut residents from 1955 to 1973, as obtained from the Connecticut Tumor Registry.
- Matched controls were selected for each case based on age at death (within eight years), year of death (within three years), and number of jobs as obtained from combined death certificate and directory sources.

TABLE 6.2

	Controls	
Cases	Yes	No
Yes	2	26
No	5	12

- An occupational indicator of metal worker (yes/no) was evaluated as the possible risk factor in this study; results are shown in Table 6.2.

We have, for a one-tailed test, $Z = \frac{26-5}{\sqrt{26+5}} = 3.77$ indicating a very highly significant increase of risk associated with the exposure ($p < 0.001$).

Example 6.3

- A study in Maryland identified 4032 white persons, enumerated in a unofficial 1963 census, who became widowed between 1963 and 1974.
- These people were matched, one to one, to married persons on the basis of race, gender, year of birth, and geography of residence.
- The matched pairs were followed to a second census in 1975. The overall male mortality is shown in Table 6.3.

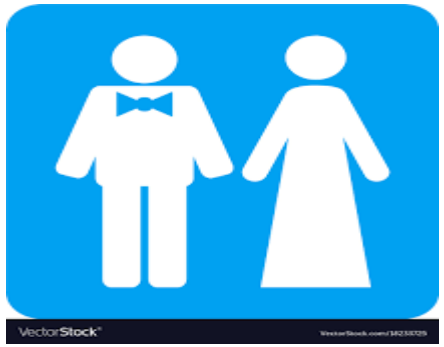


TABLE 6.3

	Married Men	
Widowed Men	Dead	Alive
Dead	2	292
Alive	210	700

An application of McNemar's chi-square test (two-sided) yields

$$X^2 = \frac{(292 - 210)^2}{292 + 210} = 13.39$$

It can be seen that the null hypothesis of equal mortality should be rejected at the 0.05 level $(13.39 > 3.84)$.

Comparison of Two Proportions (When Large Independent Samples Will be Used)

- One-sided (right tailed) test

$$H_0: \pi_1 = \pi_2 \text{versus } H_a: \pi_1 > \pi_2$$

Reject H_0 at the significance of level α if $Z.test \geq Z_\alpha$

- One-sided (left tailed) test

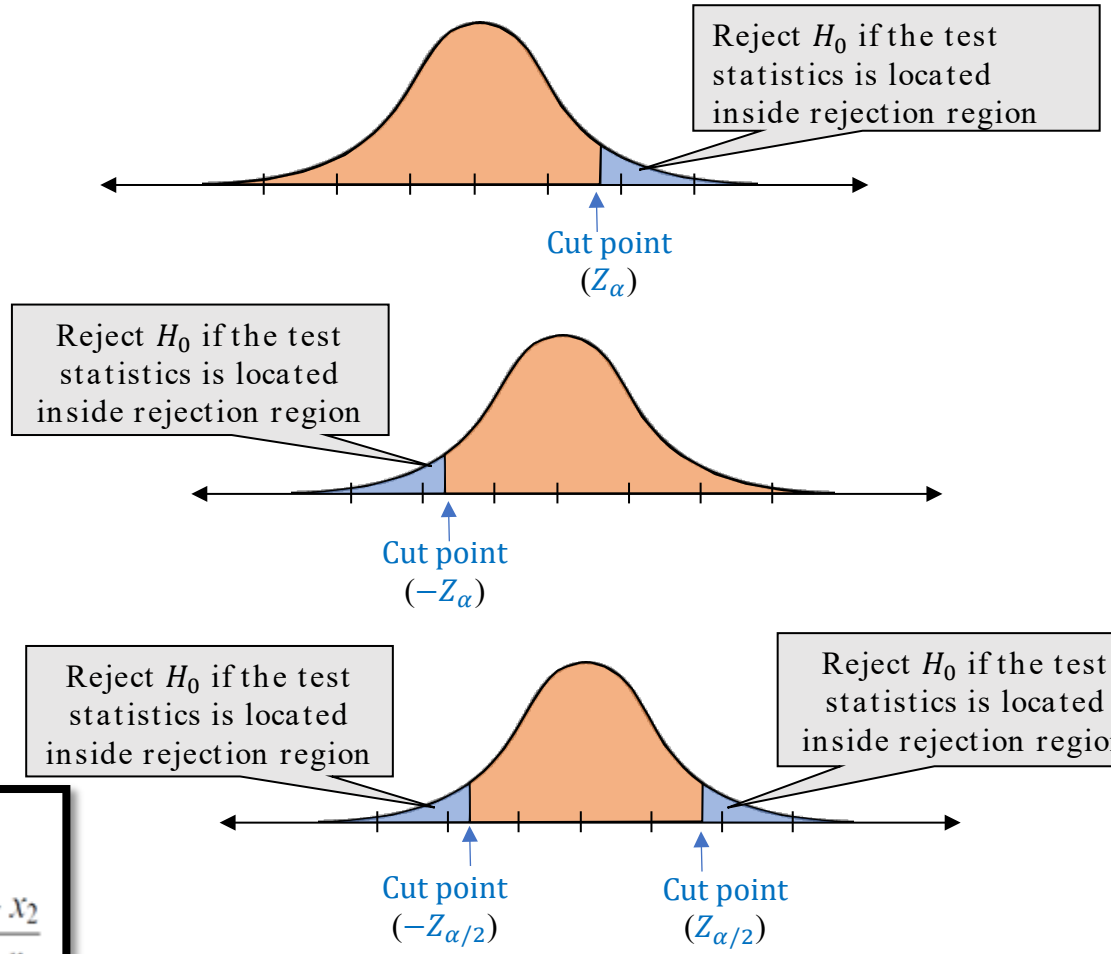
$$H_0: \pi_1 = \pi_2 \text{versus } H_a: \pi_1 < \pi_2$$

Reject H_0 at the significance of level α if $Z.\text{test} \leq -Z_\alpha$

- Two-sided (two tailed) test

$$H_0: \pi_1 = \pi_2 \text{versus } H_a: \pi_1 \neq \pi_2$$

Reject H_0 at the significance of level α if $Z.\text{test} \geq Z_{\alpha/2}$ *or if* $Z.\text{test} \leq -Z_{\alpha/2}$



$$z = \frac{p_2 - p_1}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$$

where $p_1 = \frac{x_1}{n_1}$, $p_2 = \frac{x_2}{n_2}$ and p is the pooled proportion, defined by $p = \frac{x_1 + x_2}{n_1 + n_2}$

Comparison of Two Proportions (chi-square test)

- For two-sided test $H_0: \pi_1 = \pi_2$ vs $H_a: \pi_1 \neq \pi_2$, the square of the z score, denoted X^2, is more often used.
- The test is referred to as the chi-square test:

$$X^2 = \frac{(n_1 + n_2)[x_1(n_2 - x_2) - x_2(n_1 - x_1)]^2}{n_1 n_2 (x_1 + x_2)(n_1 + n_2 - x_1 - x_2)} = \frac{(a + b + c + d)(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

and the null hypothesis is rejected at the 0.05 level when $X^2 \geq 3.84$.

TABLE 6.4

Factor	Sample 1	Sample 2	Total
Present	a	c	$a + c$
Absent	b	d	$b + d$
Sample size	$n_1 = a + b$	$n_2 = c + d$	$N = a + b + c + d$

$$P_1 = \frac{a}{n_1}$$
$$P_2 = \frac{c}{n_2}$$
$$P = \frac{a + c}{n_1 + n_2}$$

Example 6.4

Example 6.4 A study was conducted to see whether an important public health intervention would significantly reduce the smoking rate among men. Of $n_1 = 100$ males sampled in 1965 at the time of the release of the Surgeon General's report on the health consequences of smoking, $x_1 = 51$ were found to be smokers. In 1980 a second random sample of $n_2 = 100$ males, similarly gathered, indicated that $x_2 = 43$ were smokers.

An application of the method above yields

$$p = \frac{51 + 43}{100 + 100}$$

$$= 0.47$$

$$z = \frac{0.51 - 0.43}{\sqrt{(0.47)(0.53)\left(\frac{1}{100} + \frac{1}{100}\right)}}$$

$$= 1.13$$

It can be seen that the rate observed was reduced from 51% to 43%, but the reduction is not statistically significant at the 0.05 level ($z = 1.13 < 1.65$).

Example 6.5

- An investigation was made into fatal poisonings of children by two drugs which were among the leading causes of such deaths. In each case, an inquiry was made as to how the child had received the fatal overdose and responsibility for the accident was assessed. Results are shown in Table 6.5.

TABLE 6.5

	Drug A	Drug B
Child responsible	8	12
Child not responsible	31	19

We have the proportions of cases for which the child is responsible:

$$p_A = \frac{8}{8+31} = 0.205 \quad \text{or} \quad 20.5\%$$

$$p_B = \frac{12}{12+19} = 0.387 \quad \text{or} \quad 38.7\%$$

suggesting that they are not the same and that a child seems more prone to taking drug B than drug A. However, the chi-square statistic

$$X^2 = \frac{(39+31)[(8)(19)-(31)(12)]^2}{(39)(31)(20)(50)}$$

$$= 2.80 \quad (<3.84;\ \alpha = 0.05)$$

shows that the difference is not statistically significant at the 0.05 level.

Example 6.6

- In Example 1.2, a case–control study was conducted to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia.
- The primary risk factor under investigation was employment in shipyards during World War II,
- Table 6.6 provides data for nonsmokers.

TABLE 6.6

Shipbuilding	Cases	Controls
Yes	11	35
No	50	203

For the cases, $P_2 = \frac{11}{61} = 0.18$

For the controls, $P_1 = \frac{35}{238} = 0.147$

The pooled proportion $P = \frac{11+35}{61+238} = 0.154$

leading to $$z = \frac{0.180 - 0.147}{\sqrt{(0.154)(0.846)\left(\frac{1}{61} + \frac{1}{238}\right)}} = 0.64$$

It can be seen that the rate of employment for the cases (18.0%) was higher than that for the controls (14.7%), but the difference is not statistically significant at the 0.05 level ($z = 0.64 < 1.65$).

Example 6.7

- The role of smoking in the etiology of pancreatitis has been recognized for many years.
- To provide estimates of the quantitative significance of these factors, a hospital-based study was carried out in eastern Massachusetts and Rhode Island between 1975 and 1979.
- Ninety-eight patients who had a hospital discharge diagnosis of pancreatitis were included in this unmatched case–control study.
- The control group consisted of 451 patients admitted for diseases other than those of the pancreas and biliary tract.
- Risk factor information was obtained from a standardized interview with each subject, conducted by a trained interviewer.
- Some data for the males are shown in Table 6.7

TABLE 6.7

Use of Cigarettes	Cases	Controls
Current smokers	38	81
Never or ex-smokers	15	136
Total	53	217

With currently smoking being the exposure,

for the cases, $P_2 = \frac{38}{53} = 0.717$ and

for the controls, $P_1 = \frac{81}{217} = 0.373$.

Thus the pooled proportion $P = \frac{38+81}{53+217} = 0.441$ leading to

$$z = \frac{0.717 - 0.373}{\sqrt{(0.441)(0.559)\left(\frac{1}{53} + \frac{1}{217}\right)}} = 4.52$$

It can be seen that the proportion of smokers among the cases (71.7%) was higher than that for the controls (37.7%) and the difference is highly statistically significant ($p < 0.001$).

Mantel-Haenszel Method

- **Mantel–Haenszel method** is used to investigate the relationship between two binary variables (e.g., a disease and an exposure); however, we have to control for confounders.
- Confounding variable: may be associated with either the disease or exposure or both.

The process can be summarized as follows:

1. We form 2×2 tables, one at each level of the confounder.
2. At a level of the confounder, we have the frequencies (Table 6.8.)

TABLE 6.8

	Disease Classification		
Exposure	+	−	Total
+	a	b	r_1
−	c	d	r_2
Total	c_1	c_2	n

- Under the null hypothesis the Mantel–Haenszel test is based on the standardized z statistic:

$$z = \frac{\sum a - \sum \frac{r_1 c_1}{n}}{\sqrt{\sum [r_1 r_2 c_1 c_2 / (n^2 (n-1))]}}$$

where the summation Σ is across levels of the confounder.

- One can use the square of the z score, a chi-square test at one degree of freedom, for two-sided alternatives.

- When the test above is statistically significant, the association between the disease and the exposure is *real*.
- Since we assume that the confounder is not an effect modifier, the odds ratio is constant across its levels.
- The odds ratio at each level is estimated by ad/bc; the Mantel–Haenszel procedure pools data across levels of the confounder to obtain a combined estimate:

$$\mathrm{OR}_{\mathrm{MH}} = \frac{\sum (ad/n)}{\sum (bc/n)}$$

Example 6.8

- A case–control study was conducted to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia.
- The primary risk factor under investigation was employment in shipyards during World War II,
- Data are tabulated separately in Table 6.9 for three levels of smoking.
- There are three 2×2 tables, one for each level of smoking;
  - In Example 1.1, the last two tables were combined and presented together for simplicity.
- We begin with the 2×2 table for **nonsmokers** (Table 6.10):

TABLE 6.9

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Moderate	Yes	70	42
	No	217	220
Heavy	Yes	14	3
	No	96	50

TABLE 6.10

Shipbuilding	Cases	Controls	Total
Yes	11 (a)	35 (b)	46 (r_1)
No	50 (c)	203 (d)	253 (r_2)
Total	61 (c_1)	238 (c_2)	299 (n)

Example 6.8 (Cont.)

- **For nonsmokers:**

$$
\begin{aligned}
a &= 11 \\
\frac{r_1 c_1}{n} &= \frac{(46)(61)}{299} \\
&= 9.38 \\
\frac{r_1 r_2 c_1 c_2}{n^2(n-1)} &= \frac{(46)(253)(61)(238)}{(299)^2(298)} \\
&= 6.34 \\
\frac{ad}{n} &= \frac{(11)(203)}{299} \\
&= 7.47 \\
\frac{bc}{n} &= \frac{(35)(50)}{299} \\
&= 5.85
\end{aligned}
$$

- **For moderate smokers:**

$$
\begin{aligned}
a &= 70 \\
\frac{r_1 c_1}{n} &= \frac{(112)(287)}{549} \\
&= 58.55 \\
\frac{r_1 r_2 c_1 c_2}{n^2(n-1)} &= \frac{(112)(437)(287)(262)}{(549)^2(548)} \\
&= 22.28 \\
\frac{ad}{n} &= \frac{(70)(220)}{549} \\
&= 28.05 \\
\frac{bc}{n} &= \frac{(42)(217)}{549} \\
&= 16.60
\end{aligned}
$$

- **For heavy smokers:**

$$
\begin{aligned}
a &= 14 \\
\frac{r_1 c_1}{n} &= \frac{(17)(110)}{163} \\
&= 11.47 \\
\frac{r_1 r_2 c_1 c_2}{n^2(n-1)} &= \frac{(17)(146)(110)(53)}{(163)^2(162)} \\
&= 3.36 \\
\frac{ad}{n} &= \frac{(14)(50)}{163} \\
&= 4.29 \\
\frac{bc}{n} &= \frac{(3)(96)}{163} \\
&= 1.77
\end{aligned}
$$

These results are combined to obtain the z score:

$$z = \frac{(11 - 9.38) + (70 - 58.55) + (14 - 11.47)}{\sqrt{6.34 + 22.28 + 3.36}} = 2.76$$

and a z score of 2.76 yields a one-tailed p value of 0.0029, which is beyond the 1% level. This result is stronger than those for tests at each level because it is based on more information, where all data at all three smoking levels are used. The combined odds ratio estimate is

$$\mathrm{OR}_{\mathrm{MH}} = \frac{7.47 + 28.05 + 4.29}{5.85 + 16.60 + 1.77} = 1.64$$

representing an approximate increase of 64% in lung cancer risk for those employed in the shipbuilding industry.

Inferences for General Two-Way Tables (Data Not Necessarily Binary)

- Let X_1 and X_2 denote two categorical variables,
- X_1 has I levels and X_2 has J levels;
  - There are IJ combinations of classifications (IJ cells).
  - The probabilities of the combinations are $\{\pi_{ij}\}$, where π_{ij} denotes the probability that the outcome (X_1, X_2) falls in the cell in row i and column j.
  - Under the assumption of independence, we have in cell (i, j), π_{ij} can be estimated by the *estimated expected frequencies*

$$e_{ij} = \frac{(\text{row total})(\text{column total})}{\text{sample size}}$$

		Factor X_1				Total
		Level 1	**Level 2**	· · ·	**Level I**	
Factor X_2	**Level 1**	Observed= x_{11} Expected=$e_{11} = \frac{(x_{1+})(x_{+1})}{n}$	Observed= x_{12} Expected=$e_{12} = \frac{(x_{1+})(x_{+2})}{n}$	· · ·	Observed= x_{1I} Expected=$e_{1I} = \frac{(x_{1+})(x_{+I})}{n}$	x_{1+}
	Level 2	Observed= x_{21} Expected=$e_{21} = \frac{(x_{2+})(x_{+1})}{n}$	Observed= x_{22} Expected=$e_{22} = \frac{(x_{2+})(x_{+2})}{n}$	· · ·	Observed= x_{2I} Expected=$e_{2I} = \frac{(x_{2+})(x_{+I})}{n}$	x_{2+}
	·	·	·	·	·	·
	·	·	·	·	·	·
	·	·	·	·	·	·
	Level J	Observed= x_{J1} Expected=$e_{J1} = \frac{(x_{J+})(x_{+1})}{n}$	Observed= x_{J2} Expected=$e_{J2} = \frac{(x_{J+})(x_{+2})}{n}$	· · ·	Observed= x_{JI} Expected=$e_{JI} = \frac{(x_{J+})(x_{+I})}{n}$	x_{J+}
Total		x_{+1}	x_{+2}	· · ·	x_{+I}	n

Pearson's chi-Square Statistic for Testing Independency

- Suppose we wanted to do the following test:

H_0 : There is no relationship between the exposure and disease.

H_a : There is a relationship between the exposure and disease.

- Note: The alternative hypothesis is no longer one -sided or two-sided. It is many-sided.
- We want to see is if the two factors or variables X_1 and X_2 are independence or not?
- We achieve that by comparing the observed frequencies (the x's) versus those expected (the e's) under the null hypothesis of independence.
- The Pearson's chi-square statistic:

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i,j} \frac{(observed - expected)^2}{expected} \sim \chi^2(df = (I-1)(J-1))$$

- We will reject H_0 when χ^2 test statistic is greater than the corresponding critical value found from the χ^2 table with degrees of freedom, $df = (I-1)(J-1)$

Example 6.10

- In 1979 the U.S. Veterans Administration conducted a health survey of 11,230 veterans.
- The advantages of this survey are that it includes a large random sample with a high interview response rate,
- It was done before the public controversy surrounding the issue of the health effects of possible exposure to Agent Orange.
- The data shown in Table 6.13 relate Vietnam service to having sleep problems among the 1787 veterans who entered the military service between 1965 and 1975.

TABLE 6.13

	Service in Vietnam		
Sleep Problems	Yes	No	Total
Yes	173	160	333
No	599	851	1450
Total	772	1011	1783

We have

$$e_{11} = \frac{(333)(772)}{1783} = 144.18$$
$$e_{12} = 333 - 144.18 = 188.82$$
$$e_{21} = 772 - 144.18 = 627.82$$
$$e_{22} = 1011 - 188.82 = 822.18$$

leading to

$$X^2 = \frac{(173-144.18)^2}{144.18} + \frac{(160-188.82)^2}{188.82} + \frac{(599-627.82)^2}{627.82} + \frac{(851-822.18)^2}{822.18}$$
$$= 12.49$$

This statistic, at 1 df, indicate a significant correlation ($p < 0.001$) relating Vietnam service to having sleep problems among the veterans.

Example 6.11

Example 6.11 Table 6.14 shows the results of a survey in which each subject of a sample of 300 adults was asked to indicate which of three policies they favored with respect to smoking in public places. The numbers in parentheses are expected frequencies. An application of Pearson's chi-quare test, at 6 degrees of freedom, yields

$$X^2 = \frac{(5-8.75)^2}{8.75} + \frac{(44-46)^2}{46} + \cdots + \frac{(10-4.5)^2}{4.5}$$
$$= 22.57$$

TABLE 6.14

Highest Education Level	Policy Favored: No Restrictions on Smoking	Smoking Allowed in Designated Areas Only	No Smoking at All	No Opinion	Total
College graduate	5 (8.75)	44 (46)	23 (15.75)	3 (4.5)	75
High school	15 (17.5)	100 (92)	30 (31.50)	5 (9)	150
Grade school	15 (8.75)	40 (46)	10 (15.75)	10 (4.5)	75
Total	35	184	63	18	300

The result indicates a high correlation between education levels and preferences about smoking in public places ($p = 0.001$).