Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

Ex 10.21

- Comparing two different chemotherapy treatments for breast cancer, A and B.
  - The two groups should be as comparable as possible on other prognostic factors.
- A matched study
  - The patients are assigned to pairs matched on age and clinical conditions
  - A random member of each matched pair gets treatment A and the other gets treatment B.
  - The patients are followed for 5 years, with survival as the outcome variable.

A 2 × 2 contingency table comparing treatments A and B for breast cancer based on 1242 patients

Treatment	Outcome: Survive for 5 years	Outcome: Die within 5 years	Total
A	526	95	621
B	515	106	621
Total	1041	201	1242

- Yates-corrected chi-square statistic is 0.59, which is not significant.
- Using this test assumes that the samples are independent.

A 2 × 2 contingency table with the matched pair as the sampling unit based on 621 matched pairs

Outcome of treatment A patient	Outcome of treatment B patient: Survive for 5 years	Die within 5 years	Total
Survive for 5 years	510	16	526
Die within 5 years	5	90	95
Total	515	106	621

- Probability that the treatment B member of the pair survived given that the treatment A member of the pair survived = 510/526 = .970
- Probability that the treatment B member of the pair survived given that the treatment A member of the pair died = 5/95 = .053
- Concordant pair
  - A matched pair in which the outcome is the same for each member of the pair.
- Discordant pair
  - A matched pair in which the outcomes differ for the members of the pair.
- Type A discordant pair
  - Treatment A member of the pair has the event and B does not.
- Type B discordant pair
  - Treatment B member of the pair has the event and A does not.

- Let p = probability that a discordant pair is of type A.
- H_0: $p = 1/2$ versus H_1: $p \neq 1/2$.

McNemar's Test for Correlated Proportions—Normal-Theory Test

(1) Form a 2 × 2 table of matched pairs, where the outcomes for the treatment A members of the matched pairs are listed along the rows and the outcomes for the treatment B members are listed along the columns.

(2) Count the total number of discordant pairs (n_D) and the number of type A discordant pairs (n_A).

(3) Compute the test statistic

$$X^2 = \left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2 \Big/ \left(\frac{n_D}{4}\right)$$

An equivalent version of the test statistic is also given by

$$X^2 = \left(|n_A - n_B| - 1\right)^2 / (n_A + n_B)$$

(4) For a two-sided level α test,

if $X^2 > \chi^2_{1,1-\alpha}$

then reject H_0;

if $X^2 \le \chi^2_{1,1-\alpha}$

then accept H_0.

(5) The exact p-value is given by p-value = $Pr(\chi^2_1 \ge X^2)$.

(6) Use this test only if $n_D \ge 20$.

Ex. 10.24

- Assess the statistical significance of the data in previous example.
- $n_D = 21 \geq 20$, $n_D(1/2)(1/2) = 5.25 > 5$
  - normal approximation is valid.

$$X^2 = \frac{\left(|5-10.5| - \frac{1}{2}\right)^2}{21/4} = \frac{\left(5.5 - \frac{1}{2}\right)^2}{5.25} = \frac{5^2}{5.25} = \frac{25}{5.25} = 4.76$$

Equivalently, we could also compute the test statistic from

$$X^2 = \frac{(|5-16| - 1)^2}{5+16} = \frac{10^2}{21} = 4.76$$

From Table 6 in the Appendix, note that

$$\chi^2_{1,.95} = 3.84$$
$$\chi^2_{1,.975} = 5.02$$

Thus, because $3.84 < 4.76 < 5.02$, it follows that $.025 < p < .05$,

$R \times C$ Contingency Tables

- $R \times C$ contingency table: a table with R rows and C columns.

Ex. 10.33

- Suppose we want to study further the relationship between age at first birth and development of breast cancer. In particular, we would like to know if the effect of age at first birth follows a consistent trend
  - more protection for women whose age at first birth is < 20 than for women age at first birth is 25-29.
  - higher risk for women age at first birth is ≥ 35 than for women whose age at first birth is 30-34.

Data from the international study in Example 10.4 investigating the possible association between age at first birth and case–control status

Case–control status	Age at first birth <20	20–24	25–29	30–34	≥35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465
% cases	.184	.214	.259	.298	.351	.239

Source: Reprinted with permission of *WHO Bulletin, 43*, 209–221, 1970.

Computation of the Expected Table for an *R* × *C* Contingency Table The expected number of units in the (i, j) cell $= E_{ij}$ = the product of the number of units in the ith row multiplied by the number of units in the jth column, divided by the total number of units in the table.

$$\text{Expected value of the (1,1) cell} = \frac{\text{first row total} \times \text{first column total}}{\text{grand total}} = \frac{3220(1742)}{13{,}465} = 416.6$$

$$\text{Expected value of the (1,2) cell} = \frac{\text{first row total} \times \text{second column total}}{\text{grand total}} = \frac{3220(5638)}{13{,}465} = 1348.3$$

$$\vdots$$

$$\text{Expected value of the (2,5) cell} = \frac{\text{second row total} \times \text{fifth column total}}{\text{grand total}} = \frac{10{,}245(626)}{13{,}465} = 476.3$$

Expected table for the international study data in Table 10.18

Case–control status	Age at first birth					
	<20	20–24	25–29	30–34	≥35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10,245
Total	1742	5638	3904	1555	626	13,465

- We want to compare the observed table with the expected table.
- Again, $(O\text{-}E)^2/E$ is used.
- Under null hypothesis, for an R×C contingency table, the sum of $(O\text{-}E)^2/E$ over the RC cells approximately follow a chi-square distribution with $(R\text{-}1)\times(C\text{-}1)$ df.

Chi-Square Test for an $R \times C$ Contingency Table To test for the relationship between two discrete variables, where one variable has R categories and the other has C categories, use the following procedure:

(1) Analyze the data in the form of an $R \times C$ contingency table, where O_{ij} represents the observed number of units in the (i, j) cell.

(2) Compute the expected table as shown in Equation 10.22, where E_{ij} represents the expected number of units in the (i, j) cell.

(3) Compute the test statistic

$$X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$$

which under H_0 approximately follows a chi-square distribution with $(R-1) \times (C-1)$ df.

(4) For a level α test,

if $X^2 > \chi^2_{(R-1)\times(C-1),1-\alpha}$, then reject H_0.

If $X^2 \le \chi^2_{(R-1)\times(C-1),1-\alpha}$, then accept H_0.

(5) The approximate p-value is given by the area to the right of X^2 under a $\chi^2_{(R-1)\times(C-1)}$ distribution.

(6) Use this test only if both of the following two conditions are satisfied:

(a) No more than 1/5 of the cells have expected values < 5.

(b) No cell has expected value < 1.

Ex. 10.35 Assess the statistical significance of the data in the previous example (Ex. 10.33).

All expected values are ≥ 5

$$X^2 = \frac{(320-416.6)^2}{416.6} + \frac{(1206-1348.3)^2}{1348.3} + \cdots + \frac{(406-476.3)^2}{476.3} = 130.3$$

Under H_0, X^2 follows a chi-square distribution with $(2\text{-}1)\times(5\text{-}1)$ df.

Because $\chi^2_{4,.999} = 18.47 < 130.3 = X^2$

$p < 1 - .999 = .001$

There is a significant relationship between age at first birth and development of breast cancer.

Chi-Square Test for Trend in Binomial Proportions

- The result of the previous example shows some relationship between breast cancer and age at first birth.
- It does not tell us specifically about the nature of the relationship.
- There is an increasing trend in the proportion of women with breast cancer in each succeeding column.
- Can we use a specific test to detect such trends?
- Score variable S_i
  - represent some particular numeric attribute of the group
- We can assign scores of 1, 2, 3, 4, and 5 to the five groups in the international study of breast cancer example.

Chi-Square Test for Trend in Binomial Proportions (Two-Sided Test) Suppose there are k groups and we want to test if there is an increasing (or decreasing) trend in the proportion of "successes" p_i (the proportion of units in the first row of the ith group) as i increases.

(1) Set up the data in the form of a $2 \times k$ contingency table, where success or failure is listed along the rows and the k groups are listed along the columns.

(2) Denote the number of successes in the ith group by x_i, the total number of units in the ith group by n_i, and the proportion of successes in the ith group by $\hat{p}_i = x_i/n_i$. Denote the total number of successes over all groups by x, the total number of units over all groups by n, the overall proportion of successes by $\overline{p} = x/n$, and the overall proportion of failures by $\overline{q} = 1 - \overline{p}$.

(3) Construct a score variable S_i to correspond to the ith group. This variable will usually either be 1, 2, . . . , k for the k groups or be defined to correspond to some other numeric attribute of the group.

(4) More specifically, we wish to test the hypothesis H_0: There is no trend among the p_i's versus H_1: The p_i are an increasing or decreasing function of the S_i, expressed in the form $p_i = \alpha + \beta S_i$ for some constants α, β. To relate p_i and S_i, compute the test statistic $X_1^2 = A^2/B$, where

$$A = \sum_{i=1}^{k} n_i(\hat{p}_i - \bar{p})(S_i - \bar{S})$$

$$= \left(\sum_{i=1}^{k} x_i S_i\right) - x\bar{S} = \left(\sum_{i=1}^{k} x_i S_i\right) - x\left(\sum_{i=1}^{k} n_i S_i\right) \Big/ n$$

$$B = \bar{p}\,\bar{q}\left[\left(\sum_{i=1}^{k} n_i S_i^2\right) - \left(\sum_{i=1}^{k} n_i S_i\right)^2 \Big/ n\right]$$

which under H_0 approximately follows a chi-square distribution with 1 *df*.

(5) For a two-sided level α test,

if $X_1^2 > \chi^2_{1,1-\alpha}$, then reject H_0.

If $X_1^2 \le \chi^2_{1,1-\alpha}$, then accept H_0.

(6) The approximate *p*-value is given by the area to the right of X_1^2 under a χ_1^2 distribution.

(7) The direction of the trend in proportions is indicated by the sign of A. If $A > 0$, then the proportions increase with increasing score, if $A < 0$, then the proportions decrease with increasing score.

(8) Use this test only if $n\bar{p}\,\bar{q} \ge 5.0$.

Ex 10.37

- Using the internationals study data, assess whether or not there is an increasing trend in the proportion of breast cancer cases as age at first birth increases.
- S_i = 1, 2, 3, 4, 5. x_i = 320, 1206, 1011, 463, 220. n_i = 1742, 5638, 3904, 1555, 626. x = 3220, n = 13,465,

  $\overline{p} = x/n = .239,\ \overline{q} = 1 - \overline{p} = .761.$

- A = 320(1) + 1206(2) +…+ 220(5) – (3220)[1742(1) +…+ 626(5)]/13,465 = 567.16
- B = (.239)(.761){1742(1^2) +… +626(5^2)-[1742(1) +…+ 626(5)]^2/13,465} = 2493.33

  $\chi_1^2 = A^2 / B = 129.01$

  $\chi_{1,.999}^2 = 10.83 < 129.01 = X_1^2$, H_0 can be rejected with $p < .001$.

Chi-Square Goodness-of-Fit Test

Ex. 10.39 Diastolic blood-pressure measurement were collected at home in a community-wide screening program of 14,736 adults ages 30-69 in East Boston, MA, as part of a nationwide study to detect and treat hypertensive people. Two measurements taken at one visit

- A frequency distribution of the mean diastolic blood pressure is given in the following table.
- We want to assume these measurements came from an underlying normal distribution.
- How can the validity of this assumption be tested?

Frequency distribution of mean diastolic blood pressure for adults 30–69 years old in a community-wide screening program in East Boston, Massachusetts

Group (mm Hg)	Observed frequency	Expected frequency	Group	Observed frequency	Expected frequency
<50	57	77.9	$\geq 80, <90$	4604	4478.5
$\geq 50, <60$	330	547.1	$\geq 90, <100$	2119	2431.1
$\geq 60, <70$	2132	2126.7	$\geq 100, <110$	659	684.1
$\geq 70, <80$	4584	4283.3	≥ 110	251	107.2
			Total	14,736	14,736

- Compute the expected table and compare with the observed table.
- We can use $(O\text{-}E)^2/E$ for the test.
  - The agreement between observed and expected frequencies can be summarized over the whole table by summing $(O\text{-}E)^2/E$ over all groups.
- If we have the correct underlying model, this sum will approximately follow a chi-square distribution with g-1-k *df*
  - g = number of groups and k = the number of parameters estimated from the data to compute the expected frequencies.

(4) If O_i and E_i are, respectively, the observed and expected number of units within the ith group, then compute

$$X^2 = \left(O_1 - E_1\right)^2 / E_1 + \left(O_2 - E_2\right)^2 / E_2 + \cdots + \left(O_g - E_g\right)^2 / E_g$$

where g = the number of groups.

(5) For a test with significance level α, if

$$X^2 > \chi^2_{g-k-1,1-\alpha}$$

then reject H_0; if

$$X^2 \le \chi^2_{g-k-1,1-\alpha}$$

then accept H_0.

(6) The approximate p-value for this test is given by

$$Pr\left(\chi^2_{g-k-1} > X^2\right)$$

(7) Use this test only if

(a) No more than 1/5 of the expected values are < 5.

(b) No expected value is < 1.

Ex. 10.41

- Test for goodness of fit of the normal-probability model for the previous example (Table 10.22).

Two parameters have been estimated from the data (μ, σ^2), and there are 8 groups. Therefore, $k = 2$, $g = 8$. Under H_0, X^2 follows a chi-square distribution with $8 - 2 - 1 = 5$ *df*.

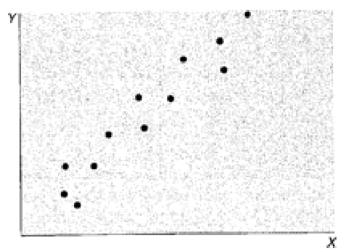
$$X^2 = (O_1 - E_1)^2 / E_1 + \cdots + (O_8 - E_8)^2 / E_8$$
$$= (57 - 77.9)^2 / 77.9 + \cdots + (251 - 107.2)^2 / 107.2 = 350.2 \sim \chi_5^2 \text{ under } H_0$$

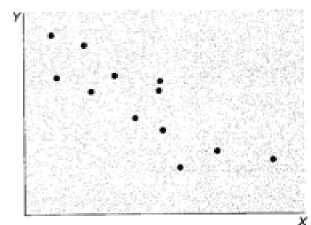
Because $\chi^2_{5,.999} = 20.52 < 350.2 = X^2$, the p-value $< 1 - .999 = .001$ and the results are very highly significant.

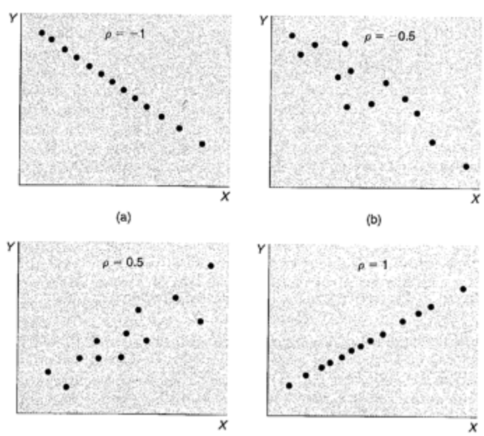
- The normal model does not provide an adequate fit to the data.

Correlation methods

1. Understand the nature and strength of the association between two measurement variables X and Y.
2. Population correlation coefficient, ρ, quantifies the linear relationship between X and Y.







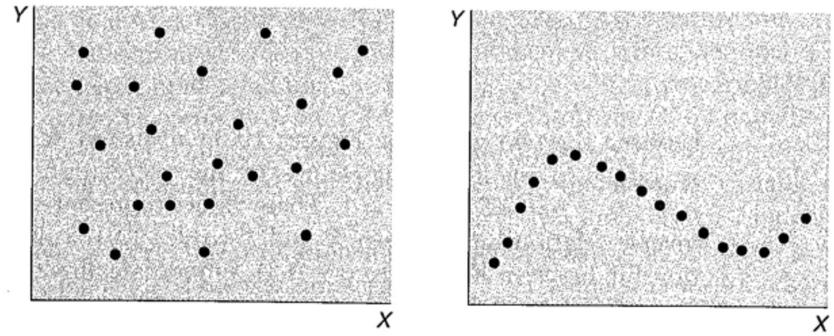
$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$-1 \leq \rho(X,Y) \leq 1$$

$$\text{Cov}(X,Y) = E((X-\mu)(Y-\nu))$$

3

3. Correlation coefficient is a measure of linear association between X and Y. So corr=0 implies that either there isn't any relationship between X and Y or a possibly non-linear relationship between X and Y.



(a) No relationship between X and Y (b) Nonlinear relationship between X and Y

4. Correlation coefficient can be affected by truncation, e.g. Relationship between SAT scores measured during the senior year of high school and GPA measured at the completion of the freshman year in college.

5. Correlation coefficient may be affected by confounding variables Relationship between size of a home and its selling price. Association between age of first job and starting salary. Sample correlation coefficient r

6. $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ Compute the sample correlation coefficient r. Compute the test statistic $t = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}$ which under H_0 follows a t distribution with $n-2$ df. For a two-sided level α test if $t > t_{n-2,1-\alpha/2}$ or $t < -t_{n-2,1-\alpha/2}$ reject H_0, otherwise accept H_0. The p-value is given by

$$p = 2 \times (\text{area to the left of}\, t \;\text{ under a}\, t_{n-2}\, \text{distribution})\, if\, t < 0$$
$$p = 2 \times (\text{area to the right of}\, t \;\text{ under a}\, t_{n-2}\, \text{distribution})\, if\, t < 0$$

7. Suppose serum-cholesterol levels in spouse pairs are measured to determine whether or not there is a correlation between cholesterol levels in spouses. Specifically, we wish to test the hypothesis $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. Suppose that r = .25 based on 100 spouse pairs. Is this evidence enough to warrant rejecting H_0? We have $n = 100, r = .25t = .25(98)^{1/2}/(1 - .252)^{1/2} = 2.56, qt(.975, 98) = 1.98$ H_0 is rejected.