EE3211 Modelling Techniques

Lecture 5

Hypothesis Testing: Categorical Data

Overview

- Methods of hypothesis testing for comparing categorical data

Types of Data

- **Categorical variables**: two or more categories that do not have any ordering

  e.g. race and ethnicity

- **Dichotomous variables**: two possible values

  e.g. gender, death, disease status

- **Ordinal variables**: more than two ranked or ordered values

  e.g., amount of current smoking: none, <10/day, 10-20/day, 21-30/day, >30/day

A Scenario

- Breast cancer in women: caused partially by happenings that occur between age at menarche and age at first childbirth
  - Risk of breast cancer ↑ as time between age at menarche and age at first childbirth ↑

- Breast cancer cases: selected hospitals in the US, Greece, Yugoslavia, Brazil, Taiwan and Japan

- Controls: women of comparable age in the hospital at the same time as the cases but who did not have breast cancer
  asked all women about their age at first birth

- Divide women with at least one birth into two categories:
  - Women whose age at first birth was ≤ 29 years
  - Women whose age at first birth was ≥ 30 years

- Results: 683 of 3220 (21.2%) women with breast cancer (case women) and 1498 of 10,245 (14.6%) women without breast cancer (control women) had an age at first birth>=30

***How can we evaluate whether this difference is significant or simply due to chance?**

Contingency-Table Method

- 2 × 2 contingency table: a table with two rows and two columns

- Display data that can be classified by two different variables
  -each has only two possible outcomes
  -one variable is arbitrarily assigned to rows and another variable to columns

- Each cell: represents number of units for each variable
  (1,1) cell: cell in the first row and first column
  (1,2) cell: cell in the first row and second column …..
  (2,2) cell: cell in the second row and second column

- The observed number of units: O_{11}, O_{12}, O_{21}, and O_{22}

1. **Row marginal totals** or **row margins:** the number of units in each row and show them in the right margins.
2. **Column marginal totals** or **column margins :** the number of units in each column and show them in the bottom margins.
3. **Grand total**: the total number of units in the four cells, which is displayed in the lower right-hand corner of the table

Example: Suppose all women with at least one birth in the breast cancer study described previously are classified as either cases or controls and with age at first birth as either <=29 or >=30. The our possible combinations are shown in the first table here:

Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls

	Age at first birth		
Status	≥30	≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

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

- **Test for homogeneity of binomial proportions**: tests whether the proportions are the same in two independent samples
  - one set of margins is fixed (e.g. rows)
  - no. of successes in each row: random variable

- Test of independence or a test of association: tests whether there is some association between two reported measures of a characteristic

Example:

- Association between two reported measures of dietary cholesterol for same person? (reproducibility)
- Food-frequency questionnaire (FFQ): measure dietary intake
  - person specifies number of servings consumed per day of different food items
- total nutrient composition is calculated from specific dietary components of food item

A comparison of dietary cholesterol assessed by a food-frequency questionnaire at two different times

	Second food-frequency questionnaire		
First food-frequency questionnaire	High	Normal	Total
High	15	5	20
Normal	9	21	30
Total	24	26	50

Significance Testing: Using Contingency-Table Approach

H_0: p1=p2

H_1: p1≠p2

Table 10.4 General contingency table for the international-study data in Example 10.4 if (1) of n_1 women in the case group, x_1 are exposed and (2) of n_2 women in the control group, x_2 are exposed (that is, having an age at first birth ≥30)

	Age at first birth		
Case–control status	≥30	≤29	Total
Case	x_1	$n_1 - x_1$	n_1
Control	x_2	$n_2 - x_2$	n_2
Total	$x_1 + x_2$	$n_1 + n_2 - (x_1 + x_2)$	$n_1 + n_2$

$$\hat{p} = \frac{n_1\widehat{p_1} + n_2\widehat{p_2}}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$n_1\hat{p} = \frac{n_1(x_1+x_2)}{n_1+n_2}$$

x_1 = no. of exposed women in group 1 (age at first birth ⩾30)

x_2 = no. of exposed women in group 2 (age at first birth ⩽29)

Computation of expected values for 2 × 2 contingency tables

Expected number of units in the (*i,j*) cell (***E_{ij}***)**:**

ith row margin X ***jth*** column margin

grand total

Observed Data

Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls

Status	Age at first birth ≥30	Age at first birth ≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

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

Expected table: contingency table that would be expected if there were no relationship between parameters, i.e.

H_0: $p_1 = p_2 = p$ were true

Expected Data

Table 10.5 Expected table for the breast-cancer data in Example 10.4

Case–control status	Age at first birth ≥30	Age at first birth ≤29	Total
Case	521.6	2698.4	3220
Control	1659.4	8585.6	10,245
Total	2181	11,284	13,465

E(1,1)=3220(2181)/13465
=521.6
E(1,2)=3220(11284)/13465
=2698.4
E(2,1)=10245(2181)/13465
=1659.4
E(2,2)=10245(11284)/13465
=8585.6

Observed table vs. Expected table

- Corresponding cell values in two tables are close → accept H_0
- Comparing cells in two tables: $\frac{(O-E)^2}{E}$

O: observed number
E: expected number

- Under H_0: Sum of $\frac{(O-E)^2}{E}$ over 4 cells ~ χ^2 distribution (df=1)

Pearson χ^2 statistics

-Sum is large → reject H_0 (poor agreement between O and E tables)

-Sum is small → accept H_0 (good agreement between O and E tables)

- Condition: normal approximation to binomial distribution valid

- No expected value < 5 → normal approximation true

rule of five

2×2 Contingency Table:
Yates-Corrected Chi-Square Test (more accurate p-values)

H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$

O_{ij} : observed number of units in the (i,j) cell

E_{ij} : expected number of units in the (i,j) cell

1. Compute test statistic which under $H_0 \sim \chi_1^2$ distribution

$$X^2 = \left(|O_{11} - E_{11}| - .5\right)^2 / E_{11} + \left(|O_{12} - E_{12}| - .5\right)^2 / E_{12} + \left(|O_{21} - E_{21}| - .5\right)^2 / E_{21} + \left(|O_{22} - E_{22}| - .5\right)^2 / E_{22}$$

2. Level α test:

$X^2 > \chi^2_{1,1-\alpha}$ → reject H_0

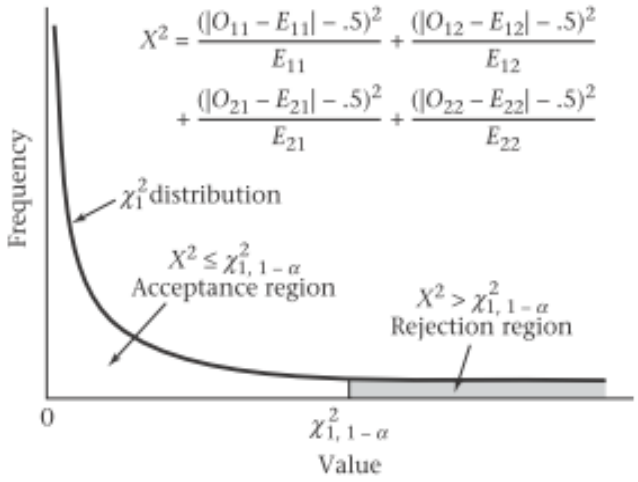
$X^2 \leq \chi^2_{1,1-\alpha}$ → accept H_0

3. P-value = area to the right of X^2 under a χ_1^2 distribution

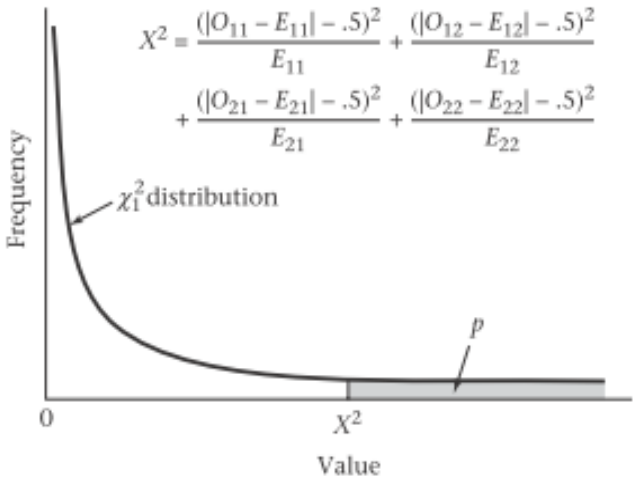
* Use this test only if none of the **four expected values < 5**

- The Yates-corrected chi-square test: *two-sided* test
- Critical region (chi-square distribution): one-sided
- Large values of $|O_{ij} - E_{ij}|$, test statistic X^2 : obtained under H_1
  - $p_1 < p_2$ or $p_1 > p_2$
- Small values of X^2: in favor of H_0

Acceptance and rejection regions for the Yates-corrected chi-square test for a 2 × 2 contingency table



Computation of the *p*-value for the Yates-corrected chi-square test for a 2 × 2 contingency table



Example on 2x2 Contingency Table: breast cancer

Q: Assess the breast-cancer data for statistical significance, using a contingency-table approach.

TABLE 10.1 Data for the international study in Example 10.4 comparing age at first birth in breast-cancer cases with comparable controls

	Age at first birth		
Status	≥30	≤29	Total
Case	683	2537	3220
Control	1498	8747	10,245
Total	2181	11,284	13,465

Solution:

- First compute observed and expected tables
- Check all expected values in expected table are at least 5

Example on 2x2 Contingency Table: breast cancer

TABLE 10.5 Expected table for the breast-cancer data in Example 10.4 (p. 373)

	Age at first birth		
Case–control status	≥30	≤29	Total
Case	521.6	2698.4	3220
Control	1659.4	8585.6	10,245
Total	2181	11,284	13,465

$$X^2 = \frac{(|683 - 521.6| - 0.5)^2}{521.6} + \frac{(|2537 - 2698.4| - 0.5)^2}{2698.4} + \frac{(|1498 - 1659.4| - 0.5)^2}{1659.4} + \frac{(|8747 - 8585.6| - 0.5)^2}{8585.6}$$

$$= \frac{(160.9)^2}{521.6} + \frac{(160.9)^2}{2698.4} + \frac{(160.9)^2}{1659.4} + \frac{(160.9)^2}{8585.6} = 49.661 + 9.599 + 15.608 + 3.017 = 77.89 \sim \chi_1^2 \text{ under } H_0$$

- $\chi^2_{1,.999} = 10.83 < 77.89 = X^2$ → p < 1 - 0.999=0.001 → results are extremely significant

Conclusion: breast cancer incidence is significantly associated with having a first child after age 30

TABLE 6 Percentage points of the chi-square distribution ($\chi^2_{d,u}$)[a]

	u													
d	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
1	0.0^4393[b]	0.0^3157[c]	0.0^3982[d]	0.00393	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	[illegible]
3	0.0717	0.115	0.216	0.352	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

- Use of a continuity correction for the contingency table is a debated subject
- Generally, p-values obtained using the continuity correction are slightly larger and are slightly less significant than comparable results (other methods)

R commands to perform the chi-square test for 2x2 tables

#*x and y are vectors pertaining to 2 variables

>chisq.test(x, y)

#use the matrix command to for the contingency table

>table=matrix(c(a, b, c, d), nrow=2)

#to obtain Yates-corrected chi-square statistic:

>chisq.test(table)

Fisher's Exact Test

- Offers exact levels of significance for any 2×2 table but it is only necessary for tables with small expected values
- H_0: $p_1 = p_2 = p$ vs. H_1: $p_1 \neq p_2$

Example: relation between high salt intake and death from CVD (retrospective or longitudinal study)

- approximately same number of 50-54 men who died from CVD (the cases) and men who died from other causes (the controls) over a 1-month period
- identify high- and low-salt users over a period of time
- compare relative frequency of death from CVD in two groups
- data are presented in the left table (2x2 contingency table):

Data concerning the possible association between cause of death and high salt intake

	Type of diet		
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

→

Fisher's Exact Test:

- same level of significance for any 2x2 table with small expected values
- similar results as χ^2 test (if applicable)

E(1,1)=7(25)/60=2.92
E(1,2)=7(35)/60=4.08

General layout of data for Fisher's exact test example

Cause of death	Type of diet: High salt	Type of diet: Low salt	Total
Non-CVD	a	b	a + b
CVD	c	d	c + d
Total	a + c	b + d	n

p_1= probability that a man was on high-salt diet with a non-CVD death

p_2= probability that a man was on high-salt diet with a CVD death

H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$

- assume that margins of this table are fixed
- exact probability of observing the table with cells a, b, c, d is:

Exact Probability of Observing A Table with Cells a, b, c, d

$$Pr(a,b,c,d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

TABLE 10.10 Hypothetical 2 × 2 contingency table in Example 10.19

2	5	7
3	1	4
5	6	11

$$\Pr(2, 5, 3, 1) = \frac{7!4!5!6!}{11!2!5!3!1!} = \frac{5040(24)(120)(720)}{(39916800)(2)(120)(6)} = 0.182$$

Enumeration of All Possible Tables with Same Margins as the Observed Table

1. Rearrangement of the rows and columns of the observed table

- Smaller row total in first row
- Smaller column total in first column

2. Begin the table with 0 in the (1, 1) cell. Other cells determined from the row and column margins.

3. Increase the (1, 1) cell by 1 in the next table, decrease the (1, 2) and (2, 1) cells by 1, increase the (2, 2) cell by 1.

4. Continue step 3 until one of the cells is 0

Data concerning the possible association between cause of death and high salt intake

	Type of diet		
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

TABLE 10.11 **Enumeration of all possible tables with fixed margins and their associated probabilities, based on the hypergeometric distribution for Example 10.19**

0	25	1	24	2	23	3	22
7	28	6	29	5	30	4	31
.017		.105		.252		.312	
4	21	5	20	6	19	7	18
3	32	2	33	1	34	0	35
.214		.082		.016		.001	

Fisher's Exact test: General Procedure and Computation of *p*-Value

H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$

*the expected value of at least one cell is <5

1. Enumerate all possible tables with the same row and column margins as the observed table

2. Compute exact probability of each table from step 13. 3. Suppose observed table is "a table" and the last table enumerated is "k table"

   i. Hypothesis H_0: $p_1 = p_2$ vs. $\boldsymbol{H_1: p_1 \neq p_2}$

   p-value = $2 \times \min[Pr(0)+Pr(1)+\ldots+Pr(a), Pr(a)+Pr(a+1)+\ldots+Pr(k),.5]$

   ii. Hypothesis H_0: $p_1 = p_2$ vs. $\boldsymbol{H_1: p_1 < p_2}$

   p-value = $Pr(0)+Pr(1)+\ldots+Pr(a)$

   Iii Hypothesis H_0: $p_1 = p_2$ vs. $\boldsymbol{H_1: p_1 > p_2}$

   p-value = $Pr(a)+Pr(a+1)+\ldots+Pr(k)$

Example on Fisher's Exact Test: Cardiovascular Disease, Nutrition

Q: Evaluate the statistical significance of the data using a two-sided alternative.

Data concerning the possible association between cause of death and high salt intake

	Type of diet		
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

$H0: p1 = p2$ vs. $H1: p1 \neq p2$

- "2" table's probability: 0.252
- compute the *p*-value: smaller of the tail probabilities corresponding to the "2" table is computed and doubled

Example on Fisher's Exact Test: Cardiovascular Disease, Nutrition

First compute left-hand tail area:

$Pr(0) + Pr(1) + Pr(2) = .017 + .105 + .252 = .375$

Right-hand tail area:

$Pr(2) + Pr(3) + \ldots + Pr(7) = .252 + .312 + .214 + .082 + .016 + .001 = .878$

$p = 2 \times \min(.375, .878, .5) = 2(.375) = .749$

If a one-sided alternative of the form $H0$: $p1 = p2$ vs. $H1$: $p1 < p2$ is used
- p-value equals:

$Pr(0) + Pr(1) + Pr(2) = .017 + .105 + .252 = .375$

Conclusion: Two proportions in this example are *not* significantly different with either a one-sided or two-sided test

We *cannot* say (based on limited amount of data) that there is a significant association between salt intake and cause of death

R command to perform Fisher's Exact test for 2x2 tables

#use matrix command to form the 2x2 table and assign it the name table

>table=matrix(c(a, b, c, d), rnow=2)

#compute the p-value when H_1: $p_1<p_2$

>p.value.lower=fisher.test(table, alternative="l")

#compute the p-value when H_1: $p_1>p_2$

>p.value.upper=fisher.test(table, alternative="g")

#copute the two-sided pvalue

>p.value.two.sided=2*min(p.value.lower, p.value.upper, 0.5)

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

Example: compare two different chemotherapy regimens for breast cancer after mastectomy. The two treatment groups should be as comparable as possible on other prognostic factors.

TABLE 10.12 A 2 × 2 contingency table comparing treatments A and B for breast cancer based on 1242 patients

A: 526/621=0.847
B: 515/621=0.829

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

Matched study:

- random member

of each matched pair receives treatment A (chemotherapy) perioperatively (within 1 week after mastectomy)

- other member gets

treatment B (chemotherapy only perioperatively)

- in pairs matched on age (+/- 5)
- Follow for 5 years
- Outcome: survival

Expected table:

Treatment	Survival for 5 years	Die within 5 years	TOTAL
A	E(1,1) 520.5	E(1,2) 100.5	621
B	E(2,1) 570.5	E(2,2) 100.5	621
TOTAL	1041	201	1242

$$E(1,1) = \frac{1041\,(621)}{1242} = 520.5$$

$$E(1,2) = \frac{201\,(621)}{1242} = 100.5$$

$$E(2,1) = \frac{1041\,(621)}{1242} = 520.5$$

$$E(2,2) = \frac{201\,(621)}{1242} = 100.5$$

Observed table:

Treatment	Survival for 5 years	Die with 5 years
A	526	95
B	515	106

~~Independent samples~~

$$\chi^2 = \frac{(|526-520.5|-0.5)^2}{520.5} + \frac{(|95-100.5|-0.5)^2}{100.5} + \frac{(|515-520.5|-0.5)^2}{520.5} + \frac{(|106-100.5|-0.5)^2}{100.5}$$

$$= \frac{5^2}{520.5} + \frac{5^2}{100.5} + \frac{5^2}{520.5} + \frac{5^2}{100.5} = 0.0480 + 0.2488 + 0.0480 + 0.2488$$

$$= 0.5936 \sim \chi^2_1 \text{ under } H_0$$

$\chi^2_{1,0.05} = 3.84 > 0.5936 \Rightarrow$ Not significant.

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

A: 526/621=0.847
B: 515/621=0.829

TABLE 10.12 A 2 × 2 contingency table comparing treatments A and B for breast cancer based on 1242 patients

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

Prob that treatment B member survived given treatment A member survived =510/526=0.97

Prob that treatment B member survived given treatment A member died =5/95=0.053

Dependent data

TABLE 10.13 A 2 × 2 contingency table with the matched pair as the sampling unit based on 621 matched pairs

	Outcome of treatment B patient		
Outcome of treatment A 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

- **Concordant pair:** matched pair, same outcome for each member of the pair
- **Discordant pair:** matched pair, different outcomes for members of the pair
- **Type A discordant pair**: discordant pair, treatment A member of the pair has the event and treatment B member does not
- **Type B discordant pair:** discordant pair, treatment B member of the pair has the event and the treatment A member does not

p = probability that a discordant pair is of type A

i. If treatment is equally effective: type A and type B discordant pairs would be approx. equal, and p = ½
ii. If treatment A is more effective: type A would be fewer than type B discordant pairs, and p < ½
iii. If treatment B is more effective: type B would be fewer than type A discordant pairs, and p > ½

McNemar's Test for Correlated Proportions

Normal-Theory Test

Form a 2×2 table of matched pairs:

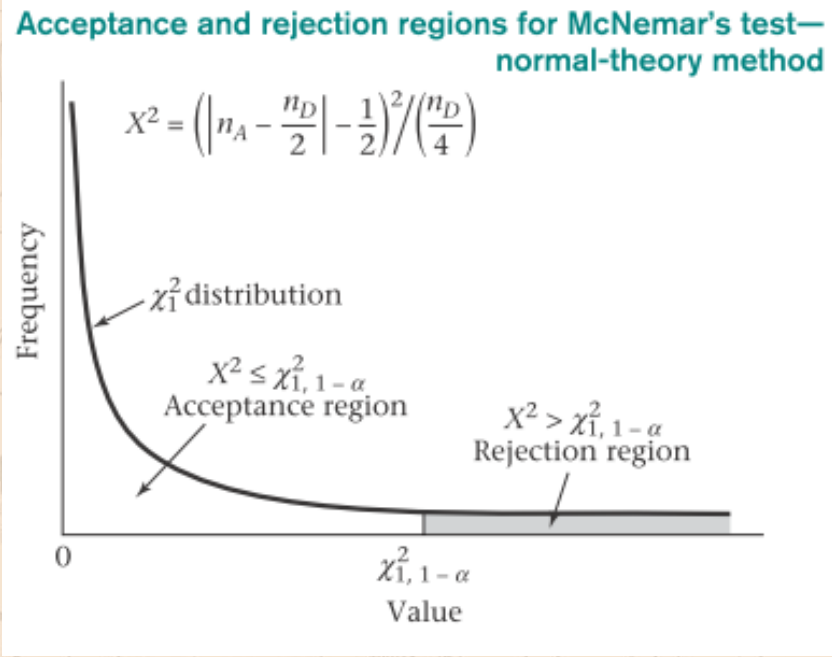
- Rows: outcomes for the treatment A members of the matched pairs
- Columns: outcomes for the treatment B members

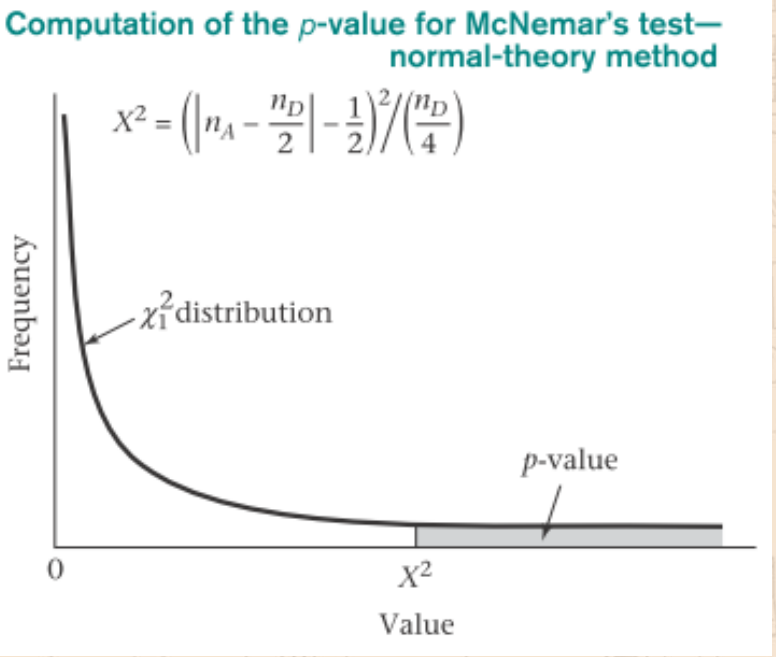
Count total number of discordant pairs (n_D) and number of type A discordant pairs (n_A)

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

2. An equivalent version: $X^2 = (|n_A - n_B| - 1)^2/(n_A + n_B)$
   where n_B= number of type B discordant pairs
3. For a two-sided level α test:
   if $X^2 > \chi_{1,1-\alpha}^2$ → reject H_0
   If $X^2 \leq \chi_{1,1-\alpha}^2$ → accept H_0
4. The exact p-value $= Pr(\chi_1^2 \geq X^2)$

*Use this test only if $\boldsymbol{n_D \geq 20}$





- A two-sided test despite the one-sided nature of the critical region

  p<1/2 or p>1/2 → | n_A - n_D /2| is large → X^2 is large

McNemar's Test for Correlated Proportions

Exact Test

- $n_D/4 < 5$ ($n_D < 20$): normal approximation to the binomial distribution cannot be used

  → test based on exact binomial probabilities is required

- The exact *p*-value:

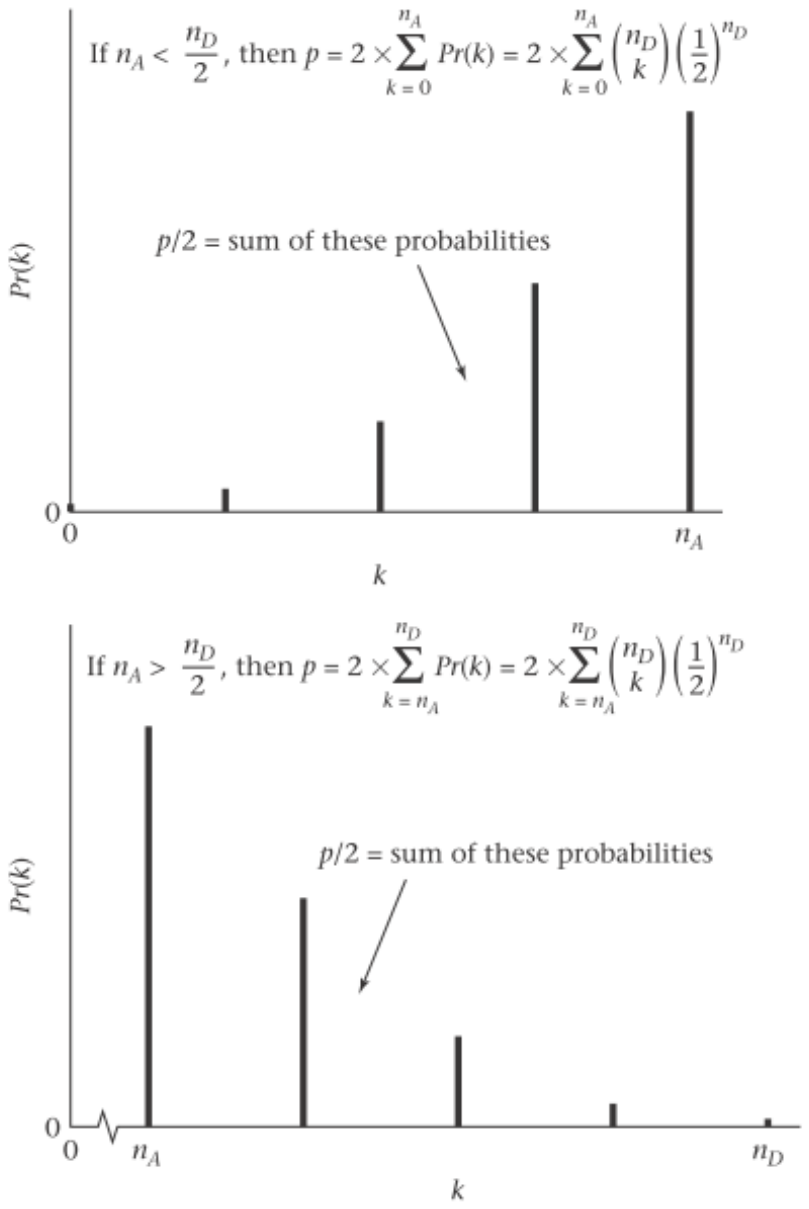
$$\text{(a)}\quad p = 2 \times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} \text{ if } n_A < n_D/2$$

$$\text{(b)}\quad p = 2 \times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} \text{ if } n_A > n_D/2$$

$$\text{(c)}\quad p = 1 \text{ if } n_A = n_D/2$$

- valid for any number of discordant pairs (n_D) but is particularly useful for $n_D < 20$ (normal-theory test cannot be used)

Figure 10.7 Computation of the *p*-value for McNemar's test—exact method



Example on McNemar's Test: Hypertension

- Blood pressure recording via automated blood-pressure machine in a computer device (with small fee)
- Study: compare the computer device with standard methods of blood pressure measurement
- 20 patients are recruited with hypertensive status assessed by both computer device and a trained observer
- Hypertensive status: either hypertensive (+)
  - systolic blood pressure is >=160 mm Hg or higher or
  - diastolic blood pressure is >=95 mm Hg or higher
- Otherwise: patient is normotensive (-)

Q: assess the statistical significance of these findings

- Each person used himself/herself as control → dependent samples → can't use Yates-corrected χ^2 test

Hypertensive status of 20 patients as judged by a computer device and a trained observer

	Hypertensive status			Hypertensive status	
Person	Computer device	Trained observer	Person	Computer device	Trained observer
1	–	–	11	+	–
2	–	–	12	+	–
3	+	–	13	–	–
4	+	+	14	+	–
5	–	–	15	–	+
6	+	–	16	+	–
7	–	–	17	+	–
8	+	+	18	–	–
9	+	+	19	–	–
10	–	–	20	–	–

Comparison of hypertensive status as judged by a computer device and a trained observer

	Trained observer	
Computer device	+	–
+	3	7
–	1	9

Concordant pairs: 9+3=12

Discordant pairs: 7+1=8

$n_A = 7, n_D = 8$

$n_D < 20$, $n_A > \frac{n_D}{2} = 4$

(a) $p = 2 \times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$ if $n_A < n_D/2$

(b) $p = 2 \times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$ if $n_A > n_D/2$

(c) $p = 1$ if $n_A = n_D/2$

$$P = 2 * \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} = 2 * \sum_{k=7}^{8} \binom{8}{k} \left(\frac{1}{2}\right)^{8}$$

Table for Binomial probabilities:

N=8, p=0.5, Pr(X≥7|p=0.5)=0.313+0.0039
=0.0352

Two tailed p-value: 2(0.0352)=0.07

- results are not statistically significant (p>0.05)
- cannot conclude that there is a significant difference between the 2 methods
- a trend can be detected toward the computer device identifying more hypertensives than the trained observer

TABLE 1 Exact binomial probabilities $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$ *(continued)*

n	k	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0313
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571

R command to perform McNemar's Test for correlated proportions

#use matrix command to form the 2x2 table :
>table=matrix(c(a, c, b, d), nrow=2)

#x and y variables (vectors) consisting of outcomes for matched pairs
>mcnemar.test(x, y)

Scenario: Age at first birth and Development of breast Cancer

- Study link between age at first birth and development of breast cancer
- Investigate whether effect of age at first birth follows a consistent trend:
  - More protection for women whose age at first birth is <20 than for women whose age at first birth is 25-29
  - Higher risk for women whose age at first birth is >=35 than for women whose age at first birth is 30-34

R × C Contingency Tables

- **R × C contingency table:** R rows and C columns; variable in the rows has R categories, variable in the columns has C categories

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 by *WHO Bulletin*, *43*, 209–221, 1970.

- Expected number of units in the (*i*,*j*) cell

$= E_{ij} =$ no. of units in the *i*th row X no. of units in the *i*th column / total no. of units

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

$(1,1)=\frac{3220(1742)}{13465}=416.6$

$(1,2)=\frac{3220(5638)}{13465}=1348.3$

…

$(2,5)=\frac{10245(626)}{13465}=476.3$

Checking: sum of the expected values across any row or column must equal corresponding row of column total

- To test relationship between two discrete variables (one variable has R categories and the other has C categories), use the following procedure:

1.R × C contingency table: O_{ij} represents the observed number of units in the (*i,j*) cell

2. Expected table: E_{ij} represents the expected number of units in the (*i,j*) cell

3. Test statistic: $X^2 = (O_{11} - E_{11})^2 / E_{11} + (O_{12} - E_{12})^2 / E_{12} + \cdots + (O_{RC} - E_{RC})^2 / E_{RC}$

$H_0 \sim \chi^2$ distribution with $(R - 1) \times (C - 1)$ *df*

do not need correction for contingency tables > 2x2

4. For a level α test:

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

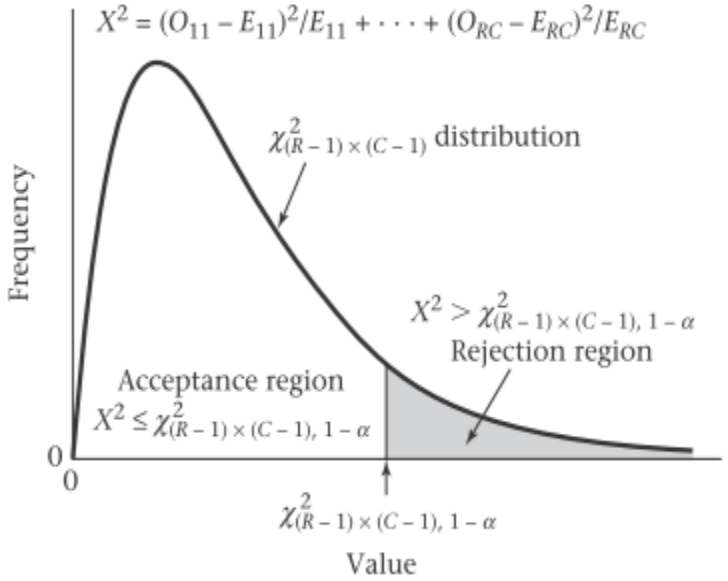
   if $X^2 \leq \chi^2_{(R-1)\times(C-1),1-\alpha}$ → accept H_0

5. P-value=area to the right of X^2 under a $\chi^2(R-1)\times(C-1)$ distribution

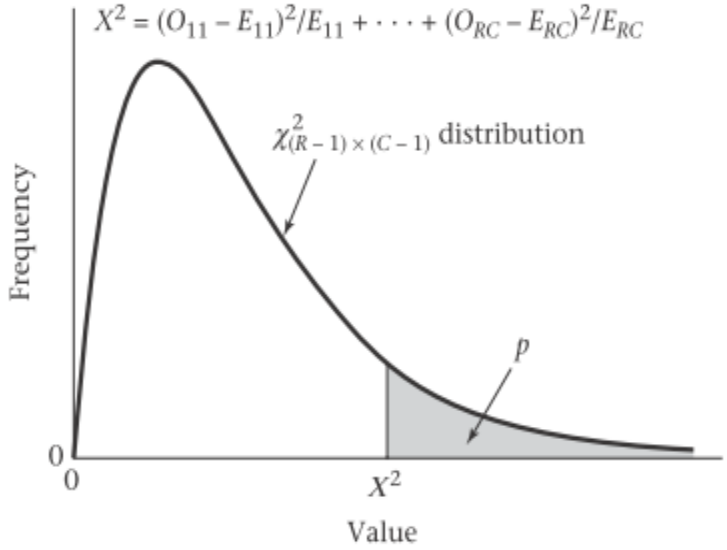
*Use this test only if both of the following conditions are satisfied:

- **No more than 1/5 of the cells have expected values <5**
- **No cell has an expected value <1**

Acceptance and rejection regions for the chi-square test for an $R \times C$ contingency table



Computation of the *p*-value for the chi-square test for an $R \times C$ contingency table



R command to perform the Chi-square test for R x C Tables

#we first create the matrix form of R x C table

>table=matrix(c(….))

#then, using the chisq.test to analyze the ata

>chisq.test(table)

Example on R x C Table: Cancer

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 whether the effect of age at first birth follows a consistent trend, that is, (1) more protection for women whose age at first birth is <20 than for women whose age at first birth is 25−29 and (2) higher risk for women whose age at first birth is ≥35 than for women whose age at first birth is 30−34. The data are presented in Table 10.16, where case−control status is indicated along the rows and age at first birth categories are indicated along the columns. The data are arranged in the form of a 2 × 5 contingency table because case−control status has two categories and age at first birth has five categories. We want to test for a relationship between age at first birth and case−control status.

TABLE 10.16 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

Q: Assess the statistical significance of the data

Example on R x C Table: Cancer

Solution:

$$Expected\ Value\ of\ the\ (1,1)\ cell = \frac{first\ row\ total \times first\ column\ total}{grand\ total} = \frac{3220(1742)}{13465} = 416.6$$

$$Expected\ Value\ of\ the\ (1,2)\ cell = \frac{first\ row\ total \times second\ column\ total}{grand\ total} = \frac{3220(5638)}{13465} = 1348.3$$

$$Expected\ Value\ of\ the\ (2,5)\ cell = \frac{second\ row\ total \times fifth\ column\ total}{grand\ total} = \frac{10245(626)}{13465} = 476.3$$

Here shows all 10 expected values:

TABLE 10.17 Expected table for the international study data in Table 10.18

	Age at first birth					
Case–control status	<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

Example on R x C Table: Cancer

- all expected values are ≥5

$$X^2 = \frac{(320 - 416.6)^2}{416} + \frac{(1206 - 1348.3)^2}{1348.3} + \cdots + \frac{(406 - 476.3)^2}{476.3} = 130.3$$

Under *H*0, *X*2 follows a chi-square distribution with (2 − 1) × (5 − 1), or 4, *df* Because

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

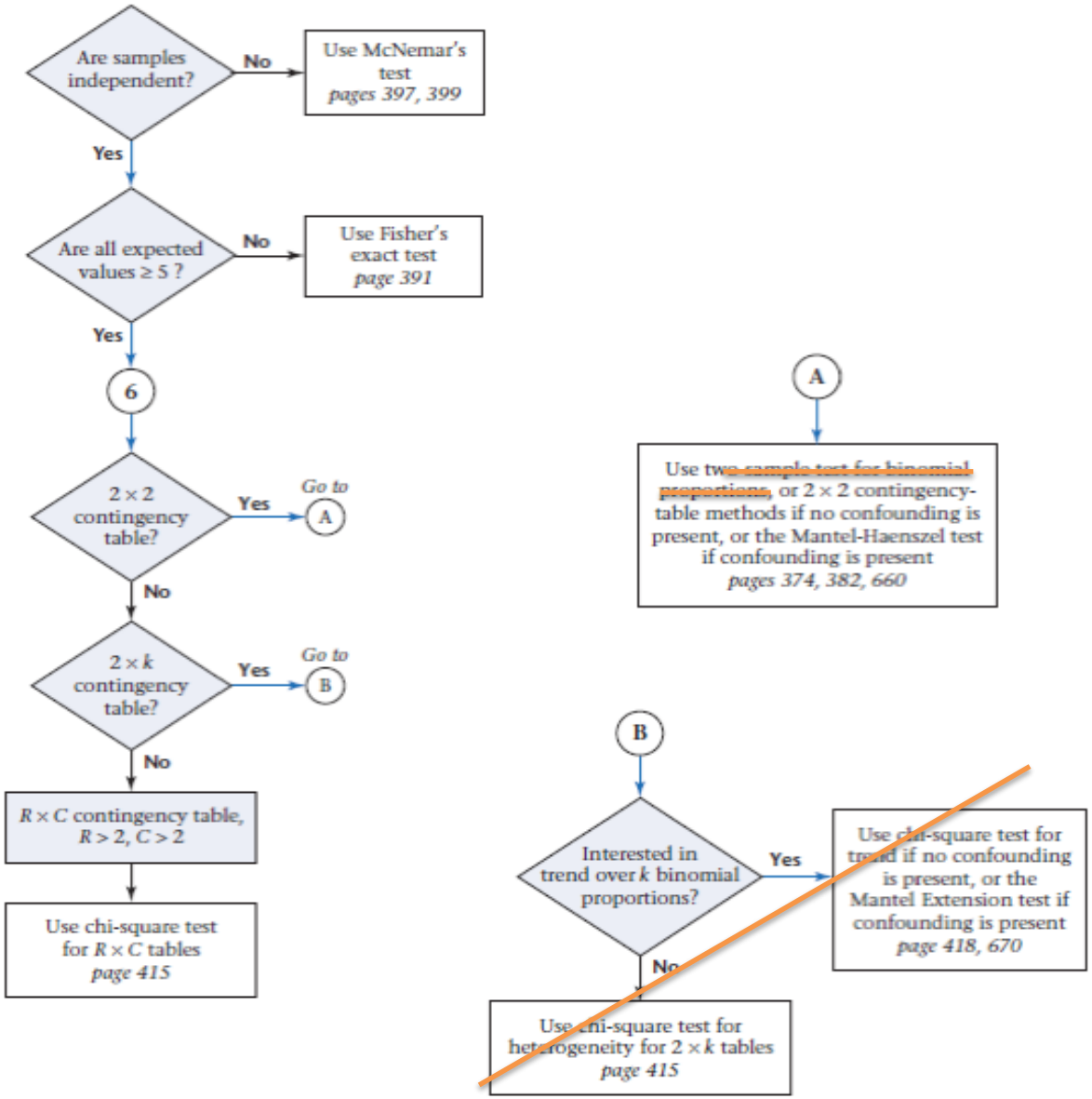
It follows that $p < 1 - .999 = .001$

- Results are very highly significant
- Conclusion: there is a significant relationship between age at first birth and prevalence of breast cancer

TABLE 6 Percentage points of the chi-square distribution ($\chi^2_{d,u}$)[a]

	u													
d	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
1	0.0^4393[b]	0.0^3157[c]	0.0^3982[d]	0.00393	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.0717	0.115	0.216	0.352	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	[illegible]
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

FIGURE 10.16 Flowchart for appropriate methods of statistical inference for categorical data



Summary

- Techniques for analyzing qualitative or categorical data
- Comparison of proportions from two independent samples using
  i. chi-square test
  ii. Fisher's exact test
  iii. McNemar's test for correlated proportions