Association between categorical variables (proportions) Chi square (χ^2) test

Professor Dr Abubakir M. Saleh

Biostatistics NUR304

Fall semester

8th week

7/11/2023

Outline

1) Construct 2-way table to examine association between two categorical variables.

2) Conduct Chi Square ($\chi 2$) test to assess evidence for association between two or more categorical variables.

Objectives

At the end of this lecture, students should be able to :

- Know how to use chi-square test for categorical variables.
- Obtain P-value and interpret it.

Constructing a two-way table

- Shows distribution of (relationship between) 2 categorical variables.

- Example: Relationship between physical exercise and the sex of individuals?

- If rows are independent variable, use row %’s.
- **2x2 table**

Sex	Exercise		No exercise		Total	
	No.	%	No.	%	No.	%
Male	31	**75.6**	10	**24.4**	41	100
Female	101	**83.5**	20	**16.5**	121	100
Total	132	**81.5**	30	**18.5**	162	100

Another example

- Drug A: out of 93 patients, 49 had response
- Drug B: Out of 91 patients, 18 had response
- Construct a two-way table
- **2x2 table**

Drug	Tumor response		Total
	Yes	No	
Drug A	49 **(53%)**	44 (47%)	93 (100%)
Drug B	18 **(20%)**	73 (80%)	91 (100%)
Total	67 (36%)	117 (64%)	184 (100%)

Larger tables

- **3x3 table**

Age group	Fever after operation						Total	
	Mild		Moderate		Severe			
	No.	%	No.	%	No.	%	No.	%
<30 Y	37	**59**	14	**22**	**12**	**19**	63	100
30-45 Y	18	**33**	17	**31**	**19**	**35**	54	100
>45 Y	24	**50**	14	**29**	**10**	**21**	48	100
Total	79		45		**41**		165	

Association between two variables

- What do we mean by association between two variables?
- Two variables are associated if distribution of one varies according to value of other
- Knowing value of one variable tells us something about value of other
- In example,

  Knowing sex of student will tell us something about physical exercise (association).

- Usually examine distribution of dependent variable according to levels of independent variable

- Distribution of physical exercise (dependent) across sex (independent)

Sex	Exercise		No exercise		Total	
	No.	%	No.	%	No.	%
Male	31	**75.6**	10	**24.4**	41	100
Female	101	**83.5**	20	**16.5**	121	100
Total	132	**81.5**	30	**18.5**	162	100

- Distribution of physical exercise differs according to sex but.....by more than we expect by chance??....

Example: Gender and Exercise among students

Sex	Exercise		No exercise		Total	
	No.	%	No.	%	No.	%
Male	31	**75.6**	10	24.4	41	100
Female	101	**83.5**	20	16.5	121	100
Total	132	**81.5**	30	18.5	162	100

75.6% of male students exercise regularly
83.5% of female students exercise regularly

Is there a real difference or it is due to chance?

Significance test for association

- Examining percentages indicates whether association may exist between exposure and disease
- But is association likely to be real or due to sampling variability?
- Need a .....

Significance test for association

- Examining percentages indicates whether association may exist between exposure and disease
- But is association likely to be real or due to sampling variability?
- Need a **…. significance test.**
- **Null hypothesis (H_0): "no association between the two variables"**
- **H_0: distribution of physical exercise is same in each group (male and female).**

Significance test for comparing proportions

- The test is called Chi Square ($\chi 2$) test
- **Step 1 – Calculate expected table**
  For H_0, as there is not real association
- **Step 2 – Calculate** $\chi 2$
- **Step 3 – Obtain p-value and interpret it**

Note: Steps 1 & 2 can be done in one quick step only for 2x2 tables

Step 1-Calculate expected table

- Only numbers, without percentages

Sex	Exercise	No exercise	Total
Male	33.4	7.6	41
Female	98.6	22.4	121
Total	132	30	162

Expected number = $\frac{\text{Row total x Column total}}{\text{Overall total}}$

Observed

41x132/162=33.4

Sex	Exercise	No exercise	Total
Male	**31**	10	**41**
Female	101	20	121
Total	**132**	30	**162**

Sex	Exercise	No exercise	Total
Male	**33.4**		41
Female			121
Total	132	30	162

Quick way

$$\text{Expected number} = \frac{\text{Row total x Column total}}{\text{Overall total}}$$

Observed

Sex	Exercise	No exercise	Total
Male	31	**10**	**41**
Female	101	20	121
Total	132	**30**	**162**

41x132/162=33.4

41x30/162=7.6

Expected

Sex	Exercise	No exercise	Total
Male	33.4	**7.6**	41
Female			121
Total	132	30	162

Quick way

$$\text{Expected number} = \frac{\text{Row total x Column total}}{\text{Overall total}}$$

Observed

Sex	Exercise	No exercise	Total
Male	31	10	41
Female	**101**	20	**121**
Total	**132**	30	**162**

41x132/162=33.4

41x30/162=7.6

121x132/162=98.6

Expected

Sex	Exercise	No exercise	Total
Male	33.4	7.6	41
Female	**98.6**		121
Total	132	30	162

Quick way

$$\text{Expected number} = \frac{\text{Row total x Column total}}{\text{Overall total}}$$

Observed

Sex	Exercise	No exercise	Total
Male	31	10	41
Female	101	20	**121**
Total	132	**30**	**162**

41x132/162=33.4

41x30/162=7.6

121x132/162=98.6

121x30/162=22.4

Expected

Sex	Exercise	No exercise	Total
Male	33.4	7.6	41
Female	98.6	**22.4**	121
Total	132	30	162

Step 2 – calculate χ2

Compare each observed value with each expected value

Observed

Sex	Exercise	No exercise	Total
Male	**31**	**10**	41
Female	**101**	**20**	121
Total	132	30	162

Expected

Sex	Exercise	No exercise	Total
Male	**33.4**	**7.6**	41
Female	**98.6**	**22.4**	121
Total	132	30	162

and obtain χ2 test statistic. $\chi 2 = \Sigma \{(O-E)^2/E\}$

- Compare each observed value with each expected value and obtain $\chi 2$ test statistic.
- $\chi 2 = \Sigma \{(O-E)^2/E\}$
- Calculate (O-E)2/E for each cell and sum over all cells
- $\chi 2 = (31 - 33.4)^2/33.4 + (10 - 7.6)^2/7.6 + (101 - 98.6)^2/98.6 + (20 - 22.4)^2/22.4 =$ **1.25**
- If $\chi 2$ value is large then (O-E) is, in general, large and data do not support H_0, i.e. real association
- If $\chi 2$ value is small then (O-E) is, in general, small and data do support H_0, i.e. no association

Step 3 - Obtain p-value

- Refer χ2 value to tables of chi-squared distribution
- Need "degrees of freedom", *ν , to take into account* number of "cells" in table
- *ν = (r - 1) x (c - 1) r = no. of rows, c = no. of columns.*
- In example, r = c = 2, so ν = (2-1) x (2-1) = 1
- Refer to table, χ2 = 1.25, d.f. =1

Percentage points of the χ^2 distribution.

	P value							
d.f.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32

- In example, $r = c = 2$, so $\nu = (2-1) \times (2-1) =$
- From table, $\chi 2$ value of 3.84, $P > 0.05$

Step 4 - Interpret p-value

- No evidence of association

Quick method for χ2

- **There is a quick formula to test for association in 2x2 table**

- **If we label cells of 2x2 table as follows:**

a b |e
c d |f
----
g h | N

Sex	Exercise	No exercise	Total
Male	31 (a)	10 (b)	41(e)
Female	101 (c)	20 (d)	121 (f)
Total	132 (g)	30 (h)	162 (N)

- **Then easiest way to calculate χ2 is using:**

$$\chi 2 = \frac{(|ad - bc|)^2 \times N}{efgh} = \frac{(31x20-101x10)^2 x162}{41x121x132x30} = 1.25$$

Another example – Tumor response

Observed

Drug	Tumor response		Total
	Yes	No	
Drug A	**49 (53%)**	44	93
Drug B	18 (20%)	73	91
Total	67 (36%)	117	184

Expected

Drug	Tumor response		Total
	Yes	No	
Drug A	**33.86**	59.4	93
Drug B	33.14	57.86	91
Total	67 (36%)	117	184

$\chi 2 = (49 - 33.86)^2/33.86 + (18 - 33.14)^2/33.14 + (44 - 59.14)^2/59.14 + (73 - 57.86)^2/57.86 = 21.52.$

Percentage points of the χ^2 distribution.

	P value							
d.f.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32

- $\chi 2$ of 21.52
- r = c = 2, so (2-1) x (2-1) = 1 d.f. and $p<0.001$

- Quick formula

Drug	Tumor response		Total
	Yes	No	
Drug A	**49 (53%)**	44	93
Drug B	18 (20%)	73	91
Total	67 (36%)	117	184

$$\chi 2 = \frac{(|ad - bc|)^2 \times N}{efgh}$$

$$= \frac{(49 \times 73 - 44 \times 18)^2 \times 184}{93 \times 91 \times 67 \times 117}$$

$$= 21.51$$

Summary

What to do when confronted with categorical data?

- **6 Step Guide....**

Step 1: Construct 2-way table to display data

Step 2: Calculate row (independent) %'s

Step 3: Carry out (O-E) $\chi 2$ test of association (or quick formula for 2x2 tables only)

Step 4: Calculate degrees of freedom for $\chi 2$ test

Step 5: Refer to tables to obtain P-value

Step 6: Interpret p-value

References

- Essential Medical Statistics, by Betty Kirkwood & Jonathan Sterne (Published by Blackwell)

  Statistics Without Tears, a Primer for Non-mathematicians, by Derek Rowntree (Published by Penguin)

Questions?