Analysis of Categorical Data

Goodness of Fit Test

(Examine Distribution)

1

Example:

The color distribution of A&A candies is supposed to be 30% red, 20% green, and 50% yellow. In a random sample of 200 A&A candies taken from the production line, 56 red, 52 green and 92 yellow. Does the sample suggest that the distribution of different color of candies still follows the 30/20/50 distribution?

2

Chi-square Goodness of Fit Test (For Distribution of Univariate Data)

The Chi-square Goodness-of-fit Test (Pearson, 1900)

$F(x)$: the true but unknown distribution function.

$F^*(x)$: a completely specified distribution function.

H_0: $F(x) = F^*(x)$ for all x

H_a: $F(x) \neq F^*(x)$ for at least one x

	Classes				
	1	2	...	k	***Total***
Observed frequencies	O_1	O_2	...	O_k	***N***
Probability distribution from $F^*(x)$	π_1^*	π_2^*		π_k^*	**1**
Expected frequencies	E_1	E_2	...	E_k	***N***

(Expected Cell Count: $E_j = \pi_j^* N$ $j = 1, 2, ..., k$)

3

Test Statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2\{k - 1\}$$

Under H_0, χ^2 has a chi-square distribution with $(k-1)$ degrees of freedom

Cochran's guidelines: (Large sample conditions)

- None of the expected cell counts less than 1.
- No more than 20% of the expected cell frequencies are less than 5.

Decision Rule:

Reject H_0 If $\chi^2 > \chi^2_\alpha$ or ***p*-value** $< \alpha$.

4

Example: (continue)

	Red	Green	Yellow	Total
Observed Frequencies	56	52	92	200
Expected Frequencies	60	40	100	200

$$\chi^2 = \frac{(56-60)^2}{60} + \frac{(52-40)^2}{40} + \frac{(92-100)^2}{100}$$

$= 4.507 < 5.991 = \chi^2_{.05}(2)$

=> fail to reject H_0

Conclusion: There is no sufficient evidence to support that the color distribution is significantly different from 30/20/50.

5

Example:

It is believe that the outcomes is uniformly distributed from a game of drawing a ball from 4 balls numbered from 1, 2, 3, and 4 in a box. The outcomes from the past 100 games were recorded, and the frequency distribution (count) is the following: 1 occurred 20 times, 2 occurred 25 times, 3 occurred 41 times and 4 occurred 14 times. Does the sample suggest that the outcome distribution of this game is not significantly different from uniform distribution?

6

Test of Homogeneity and Independence in a Two-way Table

(Examine Correlation between Two Categorical Variables)

7

Is there a relationship between Treatment and Heart Disease?

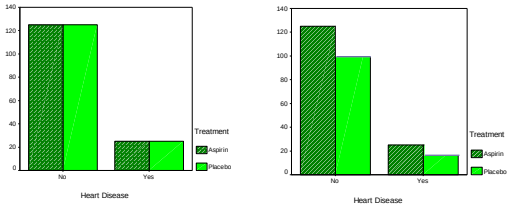
Heart Disease Variable:
"Have the disease" or "Do not have the disease."
Treatment Variable:
"Placebo" or "Aspirin".

	Heart Disease		
Treatment	Yes +	No –	Total
Placebo	36	114	150
Aspirin	14	136	150
Total	50	250	300

8

Evidence of No Correlation

Cluster bar chart

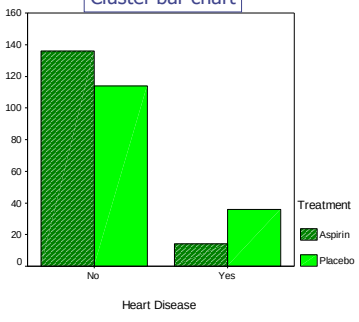


9

Evidence of Correlation

(as in the example)

Cluster bar chart



10

Questions about the sample

- Did this happen by chance?
- How likely would it happen if there is no significant treatment effect?

Testing statistical hypotheses!

11

Null hypothesis: There is **no relationship** between treatment variable and outcome variables.

Alternative hypothesis: There **is relationship** between treatment variable and outcome variables

	Heart Disease		
Treatment	Yes +	No –	Total
Placebo	36 (25)	114 (125)	150
Aspirin	14 (25)	136 (125)	150
Total	50	250	300

$\frac{150 \times 50}{300} = 25$

Observed frequency

Expected frequency

12

Expected Frequency:

Numbers in (..) :
(i,j)th cell expected freq. $= \dfrac{M_i \times N_j}{T}$

M_i: i-th **row** total
N_j: j-th **column** total
T : grand total

13

Chi-square Test for Independence

Test statistic:

$\chi^2 = (36 - 25)^2/25 + (14 - 25)^2/25$
$+ (114 - 125)^2/125 + (136 - 125)^2/125 =$ **11.616**

	Heart Disease		
Treatment	Yes +	No –	Total
Placebo	36 (25)	114 (125)	150
Aspirin	14 (25)	136 (125)	150
Total	50	250	300

14

Test Statistic:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2\{(r-1)(c-1)\}$$

Under H_0, χ^2 has a chi-square distribution with $(r-1)(c-1)$ degrees of freedom

Cochran's guidelines: (Large sample conditions)
- None of the expected cell counts less than 1.
- No more than 20% of the expected cell frequencies are less than 5.

Decision Rule:

Reject H_0 If $\chi^2 > \chi^2_\alpha$ or ***p*-value** $< \alpha$.

15

Decision Rule

- *p*-value approach:
  Reject H_0 if *p*-value $< .05$
- Critical Value approach:
  Reject H_0 if the test statistic $> \chi^2_\alpha = 3.84$

TABLE A.8
Percentiles of the chi-square distribution

	Area in Upper Tail				
df	0.100	0.050	0.025	0.010	0.001
1	2.71	3.84	5.02	6.63	10.83
2	4.61	5.99	7.38	9.21	13.82
3	6.25	7.81	9.35	11.34	16.27

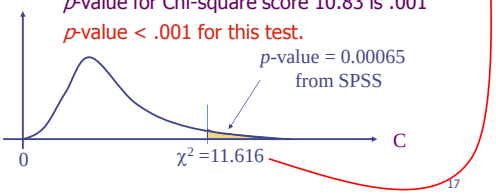
16

p-value

	Area in Upper Tail				
df	0.100	0.050	0.025	0.010	0.001
1	2.71	3.84	5.02	6.63	10.83
2	4.61	5.99	7.38	9.21	13.82
3	6.25	7.81	9.35	11.34	16.27

The probability of observing an evidence (statistic) that is at least as extreme as the observed sample evidence, if the null hypothesis is true.

p-value for Chi-square score 10.83 is .001
p-value < .001 for this test.



17

Decision (Conclusion)

Reject H_0 since *p*-value $< .001 < 0.05$ and $\chi^2 =$ **11.616** $> \chi^2_\alpha =$ **3.84**.

There is sufficient evidence to support the alternative hypothesis that there is statistically significant association between treatment and heart disease.

18

Relative Risk

Relative risk of getting heart disease for taking Placebo versus Aspirin is (36/150) / (14/150) = 2.57

	Heart Disease		
Treatment	Yes +	No –	Total
Placebo	36	114	150
Aspirin	14	136	150
Total	50	250	300

19

Step 1: Test the hypothesis at 5% level of sig.

Hypothesis:

H_0: There is **NO** relation between variable 1 (treatment) and variable 2 (outcome variables).

H_A: There is relation between two variables.

	Relapse		
	No	Yes	Total
Desipramine	14 (8)	10 (16)	24
Lithium	6 (8)	18 (16)	24
Placebo	4 (8)	20 (16)	24
Total	24	48	72

20

Step 2:

Test Statistics: $\chi^2 = \frac{(14-8)^2}{8} + \frac{(10-16)^2}{16} + \frac{(6-8)^2}{8} + \frac{(18-16)^2}{16} + \frac{(4-8)^2}{8} + \frac{(20-16)^2}{16} = \mathbf{10.5}$

Step 3:

d.f.= (3–1)(2–1) = 2

C.V. approach: Reject null hypothesis if $\chi^2 > \chi^2_{.05} = 5.99$

p-value approach: Reject null hypothesis if the ***p*-value** < α.

Step 4:

Conclusion: Since $\chi^2 = 10.5 > \chi^2_{.05} = 5.99$ or $\chi^2 = 10.5 > 9.21$, the ***p*-value** < 0.01 < 0.05. The relation between treatment and outcome variables is statistically significant.

21

TABLE A.8

Percentiles of the chi-square distribution

	Area in Upper Tail				
df	0.100	0.050	0.025	0.010	0.001
1	2.71	3.84	5.02	6.63	10.83
2	4.61	5.99	7.38	9.21 **10.5**	13.82
3	6.25	7.81	9.35	11.34	16.27

0.001 < *p*-value < .01

22

Example: Two independent random samples from two countries were collected.
Country A: 551 of 1500 adults were smokers.
Country B: 652 of 2000 adults were smokers.
At α = 0.05, test whether there is a significant difference between the percentages of smokers in country A and country B?

	Smoke	Not Smoke	Total
A	551	949	1500
B	652	1348	2000
Total	1203	2297	3500

23

Cautions

- Confounding factors: race, age, …
- Randomization, Design of experiment.
- Use of regression or multiple contingency tables.

24