Contingency Tables

1

Overview

1) **Types of Variables**
2) **Comparing (2) Categorical Variables**
   - Contingency (two-way) tables
   - χ^2 Tests
3) **2 x 2 Tables**
   - Sampling designs
   - Testing for association
   - Estimation of effects
   - Paired binary data
4) **Stratified Tables**
   - Confounding
   - Effect Modification

2

Factors and Contingency Tables

Definition: A **factor** is a categorical (discrete) variable taking a small number of values that represent the levels of the factor.

Examples

Gender with two levels: 1 = Male and 2 = Female

Disease status with three levels: 1 = Progression, 2 = Stable, 3 = Improved

AgeFactor with 4 levels: 1 = 20-29 yrs, 2 = 30-39, 3 = 40-49, 4 = 50-59

3

Factors and Contingency Tables

- We use one-way tables of the frequencies of factor levels to summarize the distribution of factors
- To assess whether two factors are related, we often construct an R x C table that cross-classifies the observations according to the 2 factors. We can test whether the factors are related using a χ^2 test.
- Examining two-way tables of Factor A vs Factor B at each level of a third Factor C shows how the A/B association may be explained or modified by C (later).

4

Categorical Data – R x C table

Example: From Doll and Hill (1952) - retrospective assessment of smoking frequency. The table displays the daily average number of cigarettes for lung cancer patients and control patients. Note there are equal numbers of cancer patients and controls.

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	7 0.5%	55 4.1%	489 36.0%	475 35.0%	293 21.6%	38 2.8%	1357
Control	61 4.5%	129 9.5%	570 42.0%	431 31.8%	154 11.3%	12 0.9%	1357
Total	68	184	1059	906	447	50	2714

5

Categorical Data - χ^2 Test

We want to test whether the smoking frequency is the same for each of the populations sampled.

H_0: distribution of smoking same in both groups

H_A: distribution of smoking not the same

What does H_0 predict we would observe if all we knew were the marginal totals?

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer							1357
Control							1357
Total	68	184	1059	906	447	50	2714

6

Categorical Data - χ^2 Test

H_0 predicts the following **expectations**:

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	34	92	529.5	453	223.5	25	1357
Control	34	92	529.5	453	223.5	25	1357
Total	68	184	1059	906	447	50	2714

- Each group has the same proportion in each cell as the overall **marginal proportion.** The "equal" expected number for each group is the result of the equal sample size in each group.
- We can test Ho by summarizing the difference between the observed and expected cell counts

7

Categorical data - χ^2 Test

Exercise 1: What does H_0 predict we would observe in the first 3 columns in this case?

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer							700
Control							2100
Total	68	184	1059	etc			2800

8

Categorical Data - χ^2 Test

Summing the differences between the observed and expected counts provides an overall assessment of H_0.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)\times(c-1))$$

- X^2 is known as the **Pearson's Chi-square Statistic.**
  - Large values of X^2 suggests the data are not consistent with H_0
  - Small values of X^2 suggests the data are consistent with H_0
- The χ^2 distribution approximates the distribution of X^2 when Ho true
  - Computer intensive "exact" tests also possible

χ^2 Test

In the Doll and Hill data the contributions to the X^2 statistic are:

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	$\frac{(7-34)^2}{34}$	$\frac{(55-92)^2}{92}$	etc.				
Control	$\frac{(61-34)^2}{34}$						
Total							

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	21.44	14.88	3.10	1.07	21.61	6.76	
Control	21.44	14.88	3.10	1.07	21.61	6.76	
Total							

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 137.7$$

$p = P(X^2 > \chi^2(5) \mid H_0 \text{ true}) < 0.0001$

Conclusion?

Categorical Data - χ^2 Test

	Factor Levels				
	1	2	…	C	Total
1	O_{11}	O_{12}	…	O_{1C}	N_1
Group 2	O_{21}				N_2
3	O_{31}				N_3
⋮	⋮				
R	O_{R1}			O_{RC}	N_R
Total	M_1	M_2		M_C	T

1. Compute the expected cell counts under null hypothesis (no association):
   $E_{ij} = N_i M_j / T$
2. Compute the chi-square statistic:
   $$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
3. Compare X^2 to $\chi^2(df)$ where
   df = (R-1) x (C-1)
4. Interpret acceptance/rejection or p-value.

Applications In Epidemiology – 2x2 table

In the specific (very common) case of 2 x 2 tables, we write

	D	not D	Total
E	a	b	$(a + b) = n_1$
not E	c	d	$(c + d) = n_2$
Total	$(a + c) = m_1$	$(b + d) = m_2$	N

For 2x2 tables we can write X^2 as

$$X^2 = \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

Compare X^2 to $\chi^2(1)$.

Applications In Epidemiology – 2x2 table

Example 1: Pauling (1971)

Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

Q: Is treatment with Vitamin C associated with a reduced probability of getting a cold?

Q: If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

13

Applications In Epidemiology – 2x2 table

	Cold - Y	Cold - N	Total
Vitamin C	17 (12%)	122 (88%)	139
Placebo	31 (22%)	109 (78%)	140
Total	48	231	279

H_0 : probability of disease does not depend on treatment

H_A : probability of disease does depend on treatment

$$X^2 = \frac{279(17\times109 - 31\times122)^2}{139\times140\times48\times231} = 4.81$$

For the p-value we compute $P(\chi^2(1) > 4.81) = 0.028$. Therefore, we reject the homogeneity of disease probability in the two treatment groups.

14

Applications In Epidemiology – 2x2 table

Example 1 fixed the number of E and not E, then evaluated the disease status after a fixed period of time (same for everyone). This is a **prospective cohort study**. Given this design we can estimate the **relative risk**:

$$RR = \frac{P(D|E)}{P(D|\bar{E})} = \frac{p_1}{p_2}$$

The range of RR is [0, ∞). By taking the logarithm, we have (- ∞, +∞) as the range for ln(RR) and a better approximation to normality

$$\ln(\widehat{RR}) = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) = \ln\left(\frac{a/n_1}{c/n_2}\right)$$

$$\ln(\widehat{RR}) \sim N\left(\ln(p_1/p_2), \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right)$$

A 95% CI can be calculated by $\ln(\widehat{RR}) \pm 1.96 * \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}$ and exponentiating.

15

Applications In Epidemiology – 2x2 table

Exercise 2: Compute the estimated RR and a 95% CI for the Pauling dataset

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

16

Applications In Epidemiology – 2x2 table

Example 2: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

Q: Is oral cancer associated with smoking?

Q: If smoking is associated with oral cancer, then what is the magnitude of the risk?

17

Applications In Epidemiology – 2x2 table

In **Example 2** we fixed the number of **cases** and **controls** then ascertained exposure status. Such a design is known as **case- control study**. Based on this we are able to directly estimate:

$$P(E|D) \quad \text{and} \quad P(E|\overline{D})$$

However, we generally are interested in the relative risk of disease given exposure, which is not estimable from these data alone - we've fixed the number of diseased and diseased free subjects. Further,

$$P(D \mid E) \neq P(E \mid D)$$

$$\frac{P(D|E)}{P(D|\overline{E})} \neq \frac{P(E|D)}{P(E|\overline{D})}$$

$$\frac{P(E|D)/\left(1-P(E|D)\right)}{P(E|\overline{D})/\left(1-P(E|\overline{D})\right)} = \frac{P(D|E)/\left(1-P(D|E)\right)}{P(D|\overline{E})/\left(1-P(D|\overline{E})\right)}$$

18

Applications In Epidemiology – 2x2 table

Instead of the relative risk we can estimate the **exposure odds ratio** which (surprisingly) is equivalent to the **disease odds ratio**:

$$\frac{P(E|D)/\left(1-P(E|D)\right)}{P(E|\overline{D})/\left(1-P(E|\overline{D})\right)} = \frac{P(D|E)/\left(1-P(D|E)\right)}{P(D|\overline{E})/\left(1-P(D|\overline{E})\right)}$$

Furthermore, for rare diseases, 1 - P(D | E) ≈ 1 so the disease odds ratio approximates the relative risk:

$$\frac{P(D|E)/\left(1-P(D|E)\right)}{P(D|\overline{E})/\left(1-P(D|\overline{E})\right)} \approx \frac{P(D|E)}{P(D|\overline{E})}$$

For rare diseases (e.g., prevalence <5%), the (sample) odds ratio estimates the (population) relative risk.

19

Applications In Epidemiology – 2x2 table

Like the relative risk, the odds ratio has [0, ∞) as its range. The **log odds ratio** has (- ∞, +∞) as its range and the normal approximation is better as an approximation to the dist of the estimated log odds ratio.

$$OR = \frac{p_1/1-p_1}{p_2/1-p_2}$$

$$\hat{O}R = \frac{ad}{bc}$$

Confidence intervals are based upon:

$$\ln(\hat{O}R) \sim N\left(\ln(OR), \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right)$$

Therefore, a 95% confidence interval for the log odds ratio is given by:

$$\ln\left(\frac{ad}{bc}\right) \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

20

Applications In Epidemiology – 2x2 table

Exercise 3: Compute χ^2, the estimated OR and a 95% CI for the Keller dataset

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

21

Applications In Epidemiology – 2x2 table

Example 3: Sex-linked traits

Suppose we collect a random sample of Drosophila and cross classify eye color and sex.

	male	female	Total
red	165	300	465
white	176	81	257
Total	341	381	722

Q: Is eye color associated with sex?

Q: If eye color is associated with sex, then what is the magnitude of the effect?

22

Applications In Epidemiology – 2x2 table

Example 3 is an example of a **cross-sectional** study since only the total for the entire table is fixed in advance. The row totals or column totals are not fixed in advance.

	male	female	Total
red	165	300	465
white	176	81	257
Total	341	381	722

Cross-sectional studies

- Sample from the entire population, not by disease status or exposure status
- Use chi-square test to test for association
- Use RR or OR to summarize association
- Cases of disease are **prevalent** cases (compared to incident cases in a prospective or cohort study)

23

Applications In Epidemiology – 2x2 table

Exercise 4: Compute χ^2 and the estimated OR for the Drosophila dataset

	male	female	Total
red	165	300	465
white	176	81	257
Total	341	381	722

24