Contingency Tables

Session 7

Module 1 Probability & Statistical Inference

The Summer Institutes
DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY of WASHINGTON

Overview

1. Defining Categorical Variables

- Contingency (two-way) tables
- χ^2 Tests

2. Comparing Two Categorical Variables

3. 2 x 2 Tables

- Sampling designs
- Testing for association
- Estimation of effects

Factor

A **factor** is a type of variable that can take one of a small number of possible values. The possible values are called the **levels** of the factor. *Also known as a categorical variable or discrete variable.*

Examples

Gender with three levels:
1 = Male, 2 = Female, 3 = Non-binary

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

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

Factors and Contingency Tables

One-way tables summarize the proportion of observations within each level of <u>one</u> factor.

Contingency tables, aka **two-way tables**, summarize the proportion of observations within each combination of levels from <u>two</u> factors.

- Also called an **R x C** table
- Often used to assess whether two factors are related
- 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 (Session 8).

EXAMPLE

Categorical Data: R x C table

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 the equal numbers of cases and controls.

Observed data

	None	**<5** cigarettes	**5-14** cigarettes	**15-24** cigarettes	**25-49** cigarettes	**50+** cigarettes	
Cases (Cancer)	7 0.5%	55 4.1%	489 36.0%	475 35.0%	293 21.6%	38 2.8%	1357
Controls (No Cancer)	61 4.5%	129 9.5%	570 42.0%	431 31.8%	154 11.3%	12 0.9%	1357
	68	184	1059	906	447	50	**2714**

EXAMPLE

Categorical Data: χ^2 test

Doll and Hill (1952)

Scientific Question

Is the distribution of smoking frequencies for those with cancer different from the distribution for those without cancer?

Restate scientific question as statistical hypotheses:

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?

Expected values?

	None	<5 cigarettes	5-14 cigarettes	15-24 cigarettes	25-49 cigarettes	50+ cigarettes	
Cases (Cancer)	34	92	529.5	453	223.5	25	1357
Controls (No Cancer)	34	92	529.5	453	223.5	25	1357
	68	184	1059	906	447	50	**2714**

EXAMPLE

Categorical Data: χ^2 test

Doll and Hill (1952)

Scientific Question
Is the distribution of smoking frequencies for those with cancer different from the distribution for those without cancer?

Each group has the same proportion in each cell as the overall **marginal proportion.** The “equalness” is just because the sample sizes were the same.

We can test H_0 by summarizing the difference between the observed and expected cell counts

Expected values

	None	**<5** cigarettes	**5-14** cigarettes	**15-24** cigarettes	**25-49** cigarettes	**50+** cigarettes	
Cases (Cancer)	34	92	529.5	453	223.5	25	1357
Controls (No Cancer)	34	92	529.5	453	223.5	25	1357
	68	184	1059	906	447	50	**2714**

Categorical Data

χ^2 Test Statistic

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

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((R-1)(C-1))$$

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 H_0 true
- Computer intensive "exact" tests also possible

EXAMPLE

Categorical Data: χ^2 test

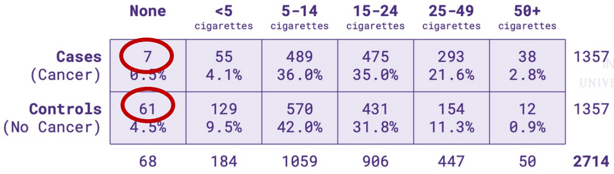
Doll and Hill (1952)

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

21.4+21.4+14.9+14.9+3.1+3.1+1.1+1.1+21.6+21.6+6.8+6.8 = 137.8

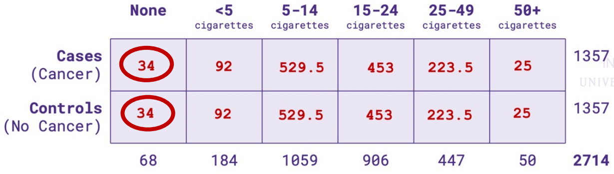
$$\frac{(7\text{-}34)^2}{34} = 21.4$$

$$\frac{(61\text{-}34)^2}{34} = 21.4$$



	None	<5 cigarettes	5-14 cigarettes	15-24 cigarettes	25-49 cigarettes	50+ cigarettes	
Cases (Cancer)	7 0.5%	55 4.1%	489 36.0%	475 35.0%	293 21.6%	38 2.8%	1357
Controls (No Cancer)	61 4.5%	129 9.5%	570 42.0%	431 31.8%	154 11.3%	12 0.9%	1357
	68	184	1059	906	447	50	**2714**

Observed



	None	<5 cigarettes	5-14 cigarettes	15-24 cigarettes	25-49 cigarettes	50+ cigarettes	
Cases (Cancer)	34	92	529.5	453	223.5	25	1357
Controls (No Cancer)	34	92	529.5	453	223.5	25	1357
	68	184	1059	906	447	50	**2714**

Expected

The contributions to the X^2 statistic are…

	None	**<5** cigarettes	**5-14** cigarettes	**15-24** cigarettes	**25-49** cigarettes	**50+** cigarettes
Cases (Cancer)	$\frac{(7\text{-}34)^2}{34} = 21.4$	$\frac{(55\text{-}92)^2}{92} = 14.9$	3.1	1.1	21.6	6.8
Controls (No Cancer)	$\frac{(61\text{-}34)^2}{34} = 21.4$	14.9	3.1	1.1	21.6	6.8

EXAMPLE

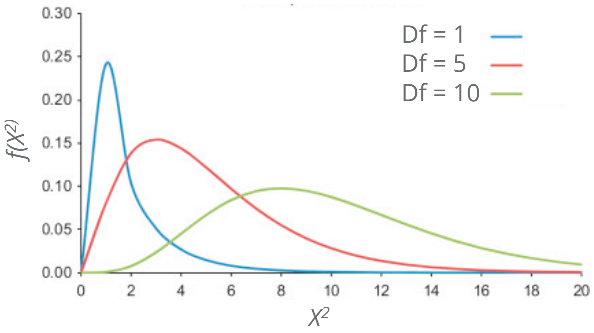
Categorical Data: χ^2 test

Doll and Hill (1952)

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 137.8$$

Degrees of freedom = (r-1)(c-1)

X^2 distributions vary by their degrees of freedom

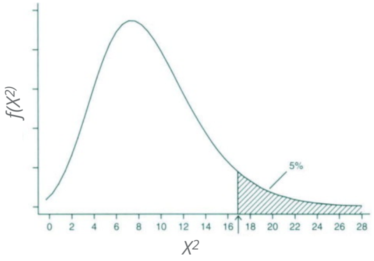


Is this significant?

H_0: distribution of smoking same in both groups

H_A: distribution of smoking not the same

Sufficiently large values of X^2 will reject H_0.



Df = (r-1)(c-1)

Df = (2-1)(6-1)

Df = 5

```
pchisq(137.8, df = 5, lower.tail=FALSE)
```

p-value = $P(X^2 > 137.8 \mid H_0) < 0.0001$

Conclusion Reject H_0 at $\alpha = 0.05$

Categorical Data: χ^2 Test

Summary: conducting χ^2 a test

1. Compute the expected cell counts under null hypothesis of no association:

$$E_{ij} = N_i M_j / T$$

2. Compute the chi-square statistic:

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3. Compare X^2 to χ^2_{df}, where $df = (R-1)\ x\ (C-1)$.
4. Interpret p-value.

Factor Levels

Groups

	One	Two	Three	. . .	C	TOTAL
One	O_{11}	O_{12}	O_{13}	…	O_{1C}	N_1
Two	O_{21}					N_2
Three	O_{31}					N_3
. . .	…					. . .
R	O_{R1}				O_{RC}	N_R
TOTAL	M_1	M_2	M_3	. . .	M_C	T

2 x 2 Tables

Epidemiological Applications

We can write the chi-square statistic for a 2 x 2 table as:

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

Then, we can compare our observed X^2 statistic to a χ^2 distribution with Df=1.

Disease Status

		D	Not D	TOTAL
Exposure Status	E	a	b	$(a+b)=n_1$
	Not E	c	d	$(c+d)=n_2$
	TOTAL	$(a+c)=m_1$	$(b+d)=m_2$	T

EXAMPLE

2 x 2 Tables

Epidemiological Applications: Pauling (1971)

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

Question 1 Is treatment with Vitamin C associated with a reduced probability of getting a cold?

Question 2 If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

Disease Status

Exposure Status	Cold	no Cold	TOTAL
Vit C	17	122	139
Placebo	31	109	140
TOTAL	48	231	279

EXAMPLE

2 x 2 Tables

Epidemiological Applications: Pauling (1971)

Scientific Q1

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

Restate scientific question as statistical hypotheses:

H_0 : probability of disease does not depend on treatment

H_A : probability of disease does depend on treatment

$$X^2 = \frac{279(17 \cdot 109 - 31 \cdot 122)^2}{139 \cdot 140 \cdot 48 \cdot 231}$$
$$= 4.81$$

```
pchisq(4.81, df = 1, lower.tail=FALSE)
```

p-value = $P(X^2 > 4.81 \mid H_0) = 0.028$

🧐 **Conclusion** Reject H_0 at $\alpha = 0.05$

Disease Status

Exposure Status		Cold	no Cold	TOTAL
	Vit C	17 (12%)	122 (88%)	139
	Placebo	31 (22%)	109 (78%)	140
	TOTAL	48	231	279

DEFINITION

2 x 2 Tables

Epidemiological Applications: Risk Ratio

Scientific Q2
If Vitamin C is associated with reducing colds, what is the magnitude of the effect?

In the Pauling (1971) example, they 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 **risk ratio (RR)** as: $RR = \dfrac{P(D|E)}{P(D|\bar{E})} = \dfrac{p_1}{p_2}$

$\bar{E} = E^C$

The range of RR is [0, ∞). The range of ln(RR) is (- ∞, +∞).

Using the natural log of RR, we can use a Normal approximation to calculate a confidence interval!

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

$$\ln\left(\widehat{RR}\right) \sim N\left[\ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right), \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right]$$

To calculate **95% CI**, determine:

$$\ln\left(\widehat{RR}\right) \pm 1.96\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}$$

Then exponentiate the endpoints.

Paws

Work through questions 1-2

EXAMPLE

2 x 2 Tables

Epidemiological Applications: 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.*

		Disease Status		
		Case	**Control**	**TOTAL**
Exposure Status	**Smoker**	484	385	**869**
	Non-Smoker	27	109	**117**
	TOTAL	**511**	**475**	**986**

Question 1 Is oral cancer associated with smoking?

Question 2 If smoking is associated with oral cancer, then what is the magnitude of the risk?

EXAMPLE

Keller (AJPH, 1965)

Since we fixed the number of **cases** and **controls** then ascertained exposure status, this study design is a **case-control study**.

Based on this we are able to directly estimate:

$$P(E \mid D) \text{ and } P(E \mid \overline{D}) \longleftarrow$$

Probability of exposure, given disease status

But we want…

Probability of disease, given exposure status

…to get Risk Ratio

However, the **risk ratio** of disease given exposure is **not estimable from these data alone**, since the number of diseased and disease-free subjects is fixed.

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

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

Odds of exposure (conditional on having the disease)

$$\frac{P(E \mid D)/(1 - P(E \mid D))}{P(E \mid \overline{D})/(1 - P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))}$$

Disease odds ratio

Odds Ratio

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

Odds of exposure (conditional on having the disease)

$$\frac{P(E \mid D)/(1 - P(E \mid D))}{P(E \mid \overline{D})/(1 - P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))}$$

😒 **exposure odds ratio** 🙂 **disease odds ratio**

And for rare diseases $1 - P(D \mid E) \approx 1$, $1 - P(D \mid \overline{E}) \approx 1$ so the disease odds ratio approximates the risk ratio:

$$\frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))} \approx \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

🙂 **disease odds ratio** 😄 **risk ratio**

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

Odds Ratio

		Disease Status		
		D	Not D	TOTAL
Exposure Status	E	a	b	(a+b)=n_1
	Not E	c	d	(c+d)=n_2
	TOTAL	(a+c)=m_1	(b+d)=m_2	T

Like the risk ratio, the odds ratio ranges from [0, ∞].

$$OR = \frac{p_1(1-p_1)}{p_2(1-p_2)}$$

Population odds ratio

$$\widehat{OR} = \frac{a \cdot d}{b \cdot c}$$

Sample odds ratio

The **log odds ratio** has (-∞, +∞) as its range and the Normal distribution approximates its sampling distribution. Confidence intervals are based upon:

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

The **95% CI** for the log odds ratio is given by:

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

Exponentiate the endpoints to get the CI for the odds ratio on its original scale.

EXAMPLE

2 x 2 Tables

Epidemiological Applications: Sex-Linked Traits

Suppose we collect a random sample of Drosophila fruit flies and cross-classify by eye color and sex.

Question 1 Is eye color associated with sex?

Question 2 If eye color is associated with sex, then what is the magnitude of the effect?

Sex

Eye Color	Male	Female	TOTAL
Red	165	300	465
White	176	81	257
TOTAL	341	381	722

EXAMPLE

2 x 2 Tables

Epidemiological Applications: Sex-Linked Traits

This is 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.

- Sample from the entire population, not by disease status or exposure status
- Use chi-squared 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 study).

		Sex		
		Male	**Female**	**TOTAL**
Eye Color	**Red**	165	300	**465**
	White	176	81	**257**
	TOTAL	**341**	**381**	**722**