

## Lecture 15

### Chapters 12&13 Relationships between Two Categorical Variables

- Tabulating and Summarizing
- Table of Expected Counts
- Statistical Significance for Two-Way Tables

## Constructing & Assessing a Two-Way Table

- Decide variables' roles, explanatory & response
- Put explanatory in rows, response in columns
- Compare conditional rates in response of interest for two (or more) explanatory groups

### Example: Constructing a Two-Way Table

- **Background:** A study recorded heavy drinking or not for bipolar alcoholics taking Valproate or placebo.
- **Question:** What are the explanatory and response variables; what should go in the rows and columns of a two-way table for the data?

□ **Response:** Explanatory is \_\_\_\_\_  
Response is \_\_\_\_\_

### Example: What to Report in a Two-Way Table

- **Background:** A study recorded incidence of heavy drinking for bipolar alcoholics taking Valproate or placebo.

	Drinking	No drinking	Total
Valproate	14	18	32
Placebo	15	7	22
Total	29	25	54

- **Question:** The numbers who drank are 14 for Valproate, 15 for placebo. Should we say the incidence of drinking was about the same for both groups?
- **Response:**

### Example: Significance in a Two-Way Table

□ **Background:** The conditional rate of heavy drinking was  $14/32=0.44$  for Valproate-takers,  $15/22=0.68$  for placebo.

	Drinking	No drinking	Total
Valproate	14	18	32
Placebo	15	7	22
Total	29	25	54

- **Question:** Does the difference seem “significant”?
- **Response:** If the difference were 0.55 vs. 0.57, we’d say \_\_\_\_\_. If it were 0.36 vs. 0.76 (more than twice as much) we’d say \_\_\_\_\_. For a difference of 0.44 vs. 0.68 from a small sample, it’s \_\_\_\_\_.

### Definition (Review)

- **Statistically significant relationship:** one that cannot easily be attributed to chance. (If there were actually no relationship in the population, the chance of seeing such a relationship in a random sample would be less than 5%.)

(We’ll learn to assess statistical significance in Chapters 13, 22, 23.)

### Another Comparison in Considering Categorical Relationships

Instead of considering how different are the *proportions* in a two-way table, we may consider how different the **counts** are from what we’d expect if the “explanatory” and “response” variables were in fact unrelated. This gives us a way to assess significance.

### Example: Expected Counts in a Two-Way Table

- **Background:** A two-way table shows heavy drinking or not observed for bipolar alcoholics taking Valproate or placebo.

Observed	Drinking	No drinking	Total
Valproate	14	18	32
Placebo	15	7	22
Total	29	25	54

- **Question:** What counts would we expect to see, if there were no relationship whatsoever between the two variables?
- **Response:** We’d expect to see counts for which the **rate** of drinking is the same (overall \_\_\_\_\_) for both groups.

### Example: Expected Counts (continued)

- Response (continued): If exactly 29/54 in each group drank,

Expected	Drinking	No drinking	Total
Valproate	$(29/54) \times 32 = 17.2$	$(25/54) \times 32 = 14.8$	32
Placebo	$(29/54) \times 22 = 11.8$	$(25/54) \times 22 = 10.2$	22
Total	29	25	54

Note: Each expected count is  $\frac{\text{Column total} \times \text{Row total}}{\text{Table total}}$

### Example: Comparing Counts

- Background:** Tables of observed and expected counts in Valproate/drinking experiment:

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

  

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- Question:** How do the counts compare?
- Response:**

### Example: Comparing Counts

- Background:** Observed and expected counts differ.

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

  

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- Question:** Is the difference significant?
- Response:** We need a way of putting the four differences in perspective...

### Components and Chi-Square Statistic

- Components to compare observed and expected counts, one table cell at a time:

$$\text{component} = \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Components are **individual** standardized squared differences.

- Chi-square** statistic combines all components by summing them up:

$$\text{chi-square} = \text{sum of } \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Chi-square is **sum** of standardized squared differences.

## Example: Chi-Square Components

- Background: Observed and Expected Tables:

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- Question: Find each component =  $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$
- Response:

## Example: Chi-Square Statistic

- Background: Observed and Expected Tables:

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- Question: Find chi-square = sum of  $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$
- Response:

## Example: Assessing Significance

- Background: Chi-square =  $0.6 + 0.7 + 0.9 + 1.0 = 3.2$ .

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- Question: Is the relationship significant?
- Response: Need to assess the relative size of 3.2.

## Statistical Significance in a 2x2 Table

It can be shown that for a 2x2 table, a chi-square statistic larger than 3.84 indicates a large enough difference between observed and expected values that there's almost certainly a relationship.

**Note:** 1.96 is the "magic" z value for which the chance of being at least that extreme is 0.05. In fact, chi-square for a 2x2 table corresponds to the square of z:  $1.96^2 = 3.84$ .

### Example: Assessing Chi-Square Statistic

□ **Background:** Chi-square =  $0.6 + 0.7 + 0.9 + 1.0 = 3.2$ .

Obs	D	ND	T
V	14	18	32
P	15	7	22
T	29	25	54

Exp	D	ND	T
V	17.2	14.8	32
P	11.8	10.2	22
T	29	25	54

- **Question:** Is the difference between observed and expected counts significant?
- **Response:** Since 3.2 is not as large as 3.84, the difference is \_\_\_\_\_  
(A larger sample would help, but not easy to get here...)

### Are Variables in a 2x2 Table Related?

1. Compute each expected count =  $\frac{\text{Column total} \times \text{Row total}}{\text{Table total}}$

2. Calculate each component =  $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

3. Find chi-square = sum of  $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

4. If chi-square > 3.84, there is a statistically significant relationship. Otherwise, we don't have evidence of a relationship.

### Example: Smoking and Alcohol Related?

□ **Background:** Overall proportion alcoholic is  $\frac{40}{1000} = 0.04$

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Non-smoker	10	760	770
Total	40	960	1000

- **Questions:** If proportions were same for smokers and non-smokers, what counts do we expect?
- **Response:** Expect...
  - \_\_\_\_\_ smokers to be alcoholic
  - \_\_\_\_\_ non-smokers to be alcoholic; also
  - \_\_\_\_\_ smokers not alcoholic
  - \_\_\_\_\_ non-smokers not alcoholic

**EXTRA CREDIT** (Max. 5 pts.) Choose two categorical variables included in the survey data 800surveyf06.txt at [www.pitt.edu/~nancyp/stat-0800/index.html](http://www.pitt.edu/~nancyp/stat-0800/index.html) (see instructions to highlight, copy, and paste into MINITAB). Follow steps 1 through 4 outlined above to determine if there is a statistically significant relationship between them.