# MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

# Review

- Last time, we finished Module 9 the inference methods for populations proportions, and started on the $\chi^2$-test.

- Today we cover Chapter 15, and use the $\chi^2$-test on the contingency tables.

# $\chi^2$-test as a general goodness-of-fit test

- The $\chi^2$-test for data in finite many K categories (cells):

(1) Under null hypothesis $\mathbf{H_0}$: find the best estimated frequencies $E_i$ ;

(2) $\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$

(3) df= number of cells - number of estimated parameters -1

Reject $\mathbf{H_0}$ if $\chi^2_{Obs} > \chi^2_{\alpha,df}$

- The $\chi^2$-test is an approximate test. As a rule of thumb, we may use it when each cell has ≥5 observations.

# $\chi^2$-test as a general goodness-of-fit test

- Pigeons Example: Do pigeons know their way to home when released? **$H_0$**: $p_1 = p_2 = p_3 = p_4$ =1/4.

  $\chi^2_{Obs}$ =7.20,

  **d.f. = 4-0-1 =3**

Use R: 1-pchisq(7.2, df=3) to get p-value=0.06578905

Fail to reject **$H_0$** at $\alpha$=0.05 level.

<u>Conclusion</u>: Pigeons do not know their direction when released.

# $\chi^2$-test as a general goodness-of-fit test

- Example: Flying bomb hits in London. (R.D. Clarke)

Divide London into 576 districts with ¼ square kilometers area each. Count the number of bombs falling into each district.

| k=# of hits | 0 | 1 | 2 | 3 | 4 | ≥5 | Total |
|---|---|---|---|---|---|---|---|
| # of districts with k hits | 229 | 211 | 93 | 35 | 7 | 1 | 576 |

- Model (**$H_0$**): X = # of hits in a district ~ Poisson($\lambda$)

# $\chi^2$-test as a general goodness-of-fit test

- Bomb hits Example:

Model ($\mathbf{H_0}$): X = # of hits in a district ~ Poisson($\lambda$)

- Under what is the best fit?

estimate $\hat{\lambda}=\frac{0(229)+1(211)+2(93)+3(95)+4(7)+5(1)}{576}=\frac{537}{576}=0.9323$

| k=# of hits | 0 | 1 | 2 | 3 | 4 | ≥5 | Total |
|---|---|---|---|---|---|---|---|
| # of districts with k hits | 229 | 211 | 93 | 35 | 7 | 1 | 576 |
| Expected under $\mathbf{H_0}$ is n*P(X=k) for Poisson(0.9323) | 226.7 | 211.4 | 98.5 | 30.6 | 7.1 | 1.6 | |

# $\chi^2$-test as a general goodness-of-fit test

- Bomb hits Example: Merge last two cells since too few counts (<5).

| k=# of hits | 0 | 1 | 2 | 3 | ≥4 | Total |
|---|---|---|---|---|---|---|
| # of districts with k hits | 229 | 211 | 93 | 35 | 8 | 576 |
| Expected n*P(X=k) | 226.7 | 211.4 | 98.5 | 30.6 | 8.7 | |

$$\chi^2_{Obs} = \frac{(226.7-229)^2}{226.7} + \frac{(211.4-211)^2}{211.4} + \frac{(98.5-93)^2}{98.5} + \frac{(30.6-35)^2}{30.6} + \frac{(8.7-8)^2}{8.7} = 1.02$$

d.f. = # of cells - # of est. para -1 = 5-1-1 =3. p-value>0.10 (Table A.8).

Use R: 1-pchisq(1.02, df=3) to get p-value=0.7964

- Fail to reject $H_0$ at $\alpha$=0.05 level.
- Conclusion: The bomb hits are random in space.

# Module 10 Contingency Tables (Chapter 15)

- Now we apply the $\chi^2$-test on contingency tables. Particularly, applying the $\chi^2$-test on 2 by 2 table reproduces the two proportions comparison tests.
- Recall: for two proportions comparison, we have tests for paired samples and two independence samples. We will do the two independence samples first.

# Module 10 Contingency Tables (Chapter 15)

- (1) Two independence population proportions.
- Example: Bicycle helmet safety effectiveness.

Data (p342)

| Head Injury | Wearing Helmet Yes | Wearing Helmet No | Total |
|---|---|---|---|
| Yes | 17 | 218 | 235 |
| No | 130 | 428 | 558 |
| Total | 147 | 646 | 793 |

Apply the $\chi^2$-test to this table, what do we get?

# Module 10 Contingency Tables (Chapter 15)

- (1) Two independence population proportions.

**Observed**

| | | |
|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{1\cdot}$ |
| $n_{21}$ | $n_{22}$ | $n_{2\cdot}$ |
| $n_{\cdot 1}$ | $n_{\cdot 2}$ | $n$ |

**Frequency**

| | | |
|---|---|---|
| $p_{11}$ | $p_{12}$ | $p_{1\cdot}$ |
| $p_{21}$ | $p_{22}$ | $p_{2\cdot}$ |
| $p_{\cdot 1}$ | $p_{\cdot 2}$ | 1 |

- What is the expected counts under **$H_0$**?
- Bicycle Helmet Example. **$H_0$**: Head injury rates are the same whether wearing helmet or not. That is, head injury and wearing helmet are independent. So $\frac{p_{11}}{p_{\cdot 1}} = \frac{p_{12}}{p_{\cdot 2}} \Leftrightarrow p_{11} = p_{1\cdot}\, p_{\cdot 1}$

# Module 10 Contingency Tables (Chapter 15)

- Generally for testing marginal independence

  **$H_0$**: $p_{ij}=p_{i\cdot}\, p_{\cdot j}$ for all i and j.

$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n}$, $\hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n}$ and the expected count for the (i,j)-th cell under **$H_0$** is $n\hat{p}_{i\cdot}\hat{p}_{\cdot j}=\frac{n_{i\cdot}n_{\cdot j}}{n}$.

**Observed**

| | | |
|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{1\cdot}$ |
| $n_{21}$ | $n_{22}$ | $n_{2\cdot}$ |
| $n_{\cdot 1}$ | $n_{\cdot 2}$ | n |

**Expected under $H_0$**

| | | |
|---|---|---|
| $\frac{n_{1\cdot}n_{\cdot 1}}{n}$ | $\frac{n_{1\cdot}n_{\cdot 2}}{n}$ | $n_{1\cdot}$ |
| $\frac{n_{2\cdot}n_{\cdot 1}}{n}$ | $\frac{n_{2\cdot}n_{\cdot 2}}{n}$ | $n_{2\cdot}$ |
| $n_{\cdot 1}$ | $n_{\cdot 2}$ | n |

# Module 10 Contingency Tables (Chapter 15)

- (1) Two independence population proportions.
- Bicycle Helmet Example.

**Expected under $H_0$**

| | | |
|---|---|---|
| $\frac{147\cdot 235}{793}=43.6$ | $\frac{646\cdot 235}{793}=191.4$ | |
| $\frac{147\cdot 558}{793}=103.4$ | $\frac{646\cdot 558}{793}=454.6$ | |

**Observed**

| | | |
|---|---|---|
| 17 | 218 | 235 |
| 130 | 428 | 558 |
| 147 | 646 | 793 |

$$\chi^2_{Obs}=\frac{(43.6-17)^2}{43.6}+\frac{(103.4-130)^2}{103.4}+\frac{(191.4-218)^2}{191.4}+\frac{(454.6-428)^2}{454.6}$$

$$=16.23 + 6.84 + 3.70 + 1.56 \quad =28.33$$

# Module 10 Contingency Tables (Chapter 15)

- (1) Two independence population proportions.
- Bicycle Helmet Example. $\chi^2_{Obs}$ =28.33,

d.f. = # of cells - # of est. para -1 = 4-2-1 =1. (est $\hat{p}_{1\cdot}$, $\hat{p}_{\cdot 1}$)

$\chi^2_{1,0.001}$ = 10.83 (Table A.8). Hence p-value<0.001

Reject $\mathbf{H_0}$ at $\alpha$=0.05 level.

Conclusion: Head injuries are *associated* with wearing helmets.

# $\chi^2$-test for marginal independence

- How does the $\chi^2$-test for marginal independence on a 2x2 table compare to the independent population proportion comparison test we covered in the last chapter?

**$\chi^2$-test $\Leftrightarrow$ 2-sided z-test** (last chapter)

- Bicycle Helmet Example. $\chi^2_{Obs}$ =28.33,

$$Z_{obs} = \left|\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\right| = \left|\frac{\frac{17}{147}-\frac{218}{646}}{\sqrt{\frac{235}{793}\left(1-\frac{235}{793}\right)\left(\frac{1}{147}+\frac{1}{646}\right)}}\right| = |-5.316|$$

$$(Z_{obs})^2 = (5.316)^2 = 28.3 = \chi^2_{Obs}$$

# $\chi^2$-test for marginal independence

- Notice that the $\chi^2$-distribution for a *continuous* random variable but the entries in a 2x2 table is *discrete* (count).
- **Continuity correction**: $\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$
- Bicycle Helmet Example.

$$\chi^2_{Obs} = \frac{(|43.6 - 17| - 0.5)^2}{43.6} + \ldots + \frac{(|454.6 - 428| - 0.5)^2}{454.6} = 27.27$$

Still p-value<0.001. Same inference as the one w/o correction.

- In practice, we should use the $\chi^2$-test with continuity correction. (so not exact match with the z-test.)

# $\chi^2$-test for marginal independence

- The $\chi^2$-test can be easily used on RxC table to test

**H$_0$**: $p_{ij} = p_{i.}\, p_{.j}$ for all i and j.

**Observed**

| | | | | |
|---|---|---|---|---|
| $n_{11}$ | $n_{12}$ | … | $n_{1C}$ | $n_{1.}$ |
| $n_{21}$ | $n_{22}$ | … | $n_{2C}$ | $n_{2.}$ |
| … | … | … | … | … |
| $n_{R1}$ | $n_{R2}$ | … | $n_{RC}$ | $n_{R.}$ |
| $n_{.1}$ | $n_{.2}$ | … | $n_{.C}$ | n |

**Expected under H$_0$**

| | | | | |
|---|---|---|---|---|
| $\frac{n_{1.}n_{.1}}{n}$ | $\frac{n_{1.}n_{.2}}{n}$ | … | $\frac{n_{1.}n_{.C}}{n}$ | $n_{1.}$ |
| … | … | … | … | … |
| $\frac{n_{R.}n_{.1}}{n}$ | $\frac{n_{R.}n_{.2}}{n}$ | … | $\frac{n_{R.}n_{.C}}{n}$ | $n_{R.}$ |
| $n_{.1}$ | $n_{.2}$ | … | $n_{.C}$ | n |

- $\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$

# $\chi^2$-test for marginal independence

- The $\chi^2$-test can be easily used on RxC table to test

  **$H_0$**: $p_{ij}=p_{i.}\, p_{.j}$ for all i and j.

- $\chi^2_{Obs} = \sum_{i=1}^{K} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$ compare with $\chi^2_{\alpha,df}$ where

  d.f. = # of cells - # of est. para -1

  = RC – [(R-1) + (C-1)] -1

  = RC – (R-$\cancel{1}$) - (C-1) -$\cancel{1}$

  = R(C-1) - (C-1)

so d.f. = (R-1)(C-1)

# Module 10 Contingency Tables (Chapter 15)

- (2) Paired two proportions. Recall last lecture

**Observed**

| | | |
|---|---|---|
| $n_{TT}$ | $n_{FT}$ | $X_2$ |
| $n_{TF}$ | $n_{FF}$ | $n-X_2$ |
| $X_1$ | $n-X_1$ | $n$ |

**Expected**

| | | |
|---|---|---|
| $n^*p_{TT}$ | $n^*p_{FT}$ | $n^*p_2$ |
| $n^*p_{TF}$ | $n^*p_{FF}$ | $n(1-p_2)$ |
| $n^*p_1$ | $n(1-p_1)$ | $n$ |

- What is the expected counts under $\mathbf{H_0}$?
- $\mathbf{H_0}$: $p_1=p_2=p \Leftrightarrow p_{TF} = p_{FT}$;

$\hat{p}_{ij}=\frac{n_{ij}}{n}$.

Pool $\hat{p}_{TF}=\hat{p}_{FT}=\frac{n_{TF}/n+n_{FT}/n}{2}$

**Expected under $H_0$**

| | | |
|---|---|---|
| $n_{TT}$ | $\frac{n_{FT}+n_{TF}}{2}$ | |
| $\frac{n_{FT}+n_{TF}}{2}$ | $n_{FF}$ | |
| | | |

# Module 10 Contingency Tables (Chapter 15)

- (2) Paired two proportions. Recall MI example

| MI | No MI: Diabetes | No MI: No Diabetes | Total |
|---|---|---|---|
| Diabetes | 9 | 37 | 46 |
| No Diabetes | 16 | 82 | 98 |
| Total | 25 | 119 | 144 |

Expected under $H_0$

| | |
|---|---|
| 9 | 26.5 |
| 26.5 | 82 |

$\chi^2_{Obs} = \frac{(26.5-16)^2}{26.5} + \frac{(26.5-37)^2}{26.5}$ =8.32. d.f. = 4-2-1 =1. (est $\hat{p}_{TT}$, $\hat{p}_{TF} = \hat{p}_{FT}$)

Use R: 1-pchisq(8.32, df=1) to get p-value=0.0039

- Reject $\mathbf{H_0}$ at $\alpha$=0.05 level.
- <u>Conclusion</u>: MI and Diabetes are associated.

# $\chi^2$-test for paired proportion comparison

- **$\chi^2$-test ⇔ 2-sided paired z-test** (last lecture)
- Recall $Z_{obs}=\dfrac{\hat{p}_{TF}-\hat{p}_{FT}}{\sqrt{\dfrac{\hat{p}_{TF}+\hat{p}_{FT}}{n}}}$ ~N(0,1), thus $Z_{obs}^2=\dfrac{(\hat{p}_{TF}-\hat{p}_{FT})^2}{\dfrac{\hat{p}_{TF}+\hat{p}_{FT}}{n}}\sim\chi_1^2$.
- In contrast,

$$\chi_{Obs}^2=\frac{\left(n_{FT}-\frac{n_{FT}+n_{TF}}{2}\right)^2}{\frac{n_{FT}+n_{TF}}{2}}+\frac{\left(n_{TF}-\frac{n_{FT}+n_{TF}}{2}\right)^2}{\frac{n_{FT}+n_{TF}}{2}}=2\frac{\left(\frac{n_{FT}-n_{TF}}{2}\right)^2}{\frac{n_{FT}+n_{TF}}{2}}$$

$$=\frac{(n_{FT}-n_{TF})^2}{n_{FT}+n_{TF}}=\frac{(n_{FT}-n_{TF})^2/n^2}{(n_{FT}+n_{TF})/n^2}=\frac{(\hat{p}_{TF}-\hat{p}_{FT})^2}{\frac{\hat{p}_{TF}+\hat{p}_{FT}}{n}}=Z_{obs}^2$$

# $\chi^2$-test for paired proportion comparison

- The McNemar's test in textbook is the $\chi^2$-test for paired proportion comparison with **continuity correction**:

$$\chi^2_{Obs}=\frac{(|n_{FT}-n_{TF}|-1)^2}{n_{FT}+n_{TF}} \text{ instead of } \frac{(n_{FT}-n_{TF})^2}{n_{FT}+n_{TF}}$$

- MI example $\chi^2_{Obs}=\frac{(|37-16|-1)^2}{37+16}$ = 7.547 instead of 8.32.

So p-value=0.006 instead of 0.0039.

Qualitatively the conclusion is the same as what we got earlier w/o correction.

# $\chi^2$-test on 2 by 2 contingency Tables

- (1) Two independence population proportions.

| | | Samples | | |
|---|---|---|---|---|
| | | A | B | |
| Factor | TRUE | $p_{11}$ | $p_{12}$ | $p_{1\cdot}$ |
| | FALSE | $p_{21}$ | $p_{22}$ | $p_{2\cdot}$ |
| | | $p_{\cdot 1}$ | $p_{\cdot 2}$ | 1 |

- **H$_0$**: The TRUE proportions are same in A and B$\Leftrightarrow$ $\boldsymbol{p_{11} = p_{1\cdot}\, p_{\cdot 1}}$
- The *unpaired* two proportions z-test is equivalent to the $\chi^2$-test on this table *w/o continuity correction*.
- Better test: $\chi^2$-test with *continuity correction*.

# $\chi^2$-test on 2 by 2 contingency Tables

- (2) Two paired population proportions.

| | | Factor in Sample A | | |
|---|---|---|---|---|
| | | TRUE | FALSE | |
| Factor in Sample B | TRUE | $p_{TT}$ | $p_{FT}$ | $p_2$ |
| | FALSE | $p_{TF}$ | $p_{FF}$ | $1-p_2$ |
| | | $p_1$ | $1-p_1$ | 1 |

- **$H_0$**: The TRUE proportions are same in A and B $\Leftrightarrow$ $p_{TF}=p_{FT}$
- The *paired* two proportions z-test is equivalent to the $\chi^2$-test on this table *w/o continuity correction*.
- Better test: McNemar test ($\chi^2$-test with *continuity correction*.)

# $\chi^2$-tests on 2 by 2 contingency Tables

- Which table to use? Recall the MI example

(1)

| | | MI | | |
|---|---|---|---|---|
| | | Yes | No | |
| Diabetes | Yes | 46 | 25 | 71 |
| | No | 98 | 119 | 217 |
| | | 144 | 144 | 288 |

(2)

| | | No MI | | |
|---|---|---|---|---|
| | | Diabetes | No Diabetes | |
| MI | Diabetes | 9 | 37 | 46 |
| | No Diabetes | 16 | 82 | 98 |
| | | 25 | 119 | 114 |

- MI and Diabetes are not associated

⇔Diabetes proportions in MI and No MI groups are the same.

- In table **(1)** **H$_0$**: $\frac{p_{11}}{p_{\cdot 1}} = \frac{p_{12}}{p_{\cdot 2}}$. In table **(2)** **H$_0$**: $\boldsymbol{p_{TF} = p_{FT}}$.

# Which table to use?

(1)

| | | MI | | |
|---|---|---|---|---|
| | | Yes | No | |
| Diabetes | Yes | 46 | 25 | 71 |
| | No | 98 | 119 | 217 |
| | | 144 | 144 | |

(2)

| | | No MI | | |
|---|---|---|---|---|
| | | Diabetes | No Diabetes | |
| MI | Diabetes | 9 | 37 | 46 |
| | No Diabetes | 16 | 82 | 98 |
| | | 25 | 119 | 114 |

- In table **(1)** **H$_0$**: $\frac{p_{11}}{p_{\cdot 1}} = \frac{p_{12}}{p_{\cdot 2}}$. In table **(2)** **H$_0$**: $\boldsymbol{p_{TF}}=\boldsymbol{p_{FT}}$.
- Mathematically both answers the same question.
- Which one is correct? Can we use both $\chi^2$-tests?
- <u>Answer</u>: **Can only use the $\chi^2$-test on table (2) H$_0$: $p_{TF}=p_{FT}$.**

# Which table to use?

(1)

| | | MI | |  |
|---|---|---|---|---|
| | | Yes | No | |
| Diabetes | Yes | 46 | 25 | 71 |
| | No | 98 | 119 | 217 |
| | | 144 | 144 | 288 |

(2)

| | | No MI | | |
|---|---|---|---|---|
| | | Diabetes | No Diabetes | |
| MI | Diabetes | 9 | 37 | 46 |
| | No Diabetes | 16 | 82 | 98 |
| | | 25 | 119 | 114 |

- **Can only use the $\chi^2$-test on table (2) H$_0$: $p_{TF}=p_{FT}$.**
- Model assumes that entries fall into the four cells **i.i.d**.
- In table (2) the 144 pairs do fall into the 4 cells **i.i.d**.
- In table (1), 144 pairs, within each pair one in the left 2 cells and one in the right 2 cells. Hence **NOT i.i.d**.

# $\chi^2$-tests on 2 by 2 contingency Tables

(1)

| | | MI | | |
|---|---|---|---|---|
| | | Yes | No | |
| Diabetes | Yes | 46 | 25 | 71 |
| | No | 98 | 119 | 217 |
| | | 144 | 144 | 288 |

(2)

| | | No MI | | |
|---|---|---|---|---|
| | | Diabetes | No Diabetes | |
| MI | Diabetes | 9 | 37 | 46 |
| | No Diabetes | 16 | 82 | 98 |
| | | 25 | 119 | 114 |

- MI and Diabetes are not associated ⇔ Table **(1)** **H$_0$**: $\frac{p_{11}}{p_{\cdot 1}}=\frac{p_{12}}{p_{\cdot 2}}$.

  ⇔ Table **(2)** **H$_0$**: $\boldsymbol{p_{TF}=p_{FT}}$.

- Can we use $\chi^2$-test on Table **(2)** **H$_0$**: $\frac{p_{11}}{p_{\cdot 1}}=\frac{p_{12}}{p_{\cdot 2}}$?
- Yes, but it is answering <u>*a different question*</u>!

# $\chi^2$-tests on 2 by 2 contingency Tables

(1)

| | | MI | | |
|---|---|---|---|---|
| | | Yes | No | |
| Diabetes | Yes | 46 | 25 | 71 |
| | No | 98 | 119 | 217 |
| | | 144 | 144 | 288 |

(2)

| | | No MI | | |
|---|---|---|---|---|
| | | Diabetes | No Diabetes | |
| MI | Diabetes | 9 | 37 | 46 |
| | No Diabetes | 16 | 82 | 98 |
| | | 25 | 119 | 114 |

- Table **(2)** $\mathbf{H_0}$: $\frac{p_{11}}{p_{\cdot 1}} = \frac{p_{12}}{p_{\cdot 2}}$ ⇔ Diabetes in the "No MI" group is independent of diabetes status of its paired person in "MI" group. ⇔ Pairing has no effect (thus not needed).

- $\chi^2$-test on Table **(2)** $\mathbf{H_0}$: $\frac{p_{11}}{p_{\cdot 1}} = \frac{p_{12}}{p_{\cdot 2}}$ test if pairing has no effect on Diabetes/MI. Not whether MI and Diabetes are associated.

# Summary

Today, we finished Module 10 contingency tables

- $\chi^2$-test is a general goodness-of-fit test.
- Using $\chi^2$-test on 2 by 2 tables can compare two populations proportions: paired or unpaired. They are equivalent to the z-tests in last lecture.
- Better to use the $\chi^2$-tests with continuity correction.
- Be careful about how tables are presented. The entries needs to be i.i.d. for usage of $\chi^2$-test.

- Homework 7 due in one week