Lecture 22: Chapter 12, Section 2
Two Categorical Variables
More About Chi-Square

- Variable Roles; Sample Sizes, Confidence Int.
- Hypotheses about Variables or Parameters
- Computing Chi-square Statistic
- Details of Chi-square Test
- Confounding Variables

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing (Lectures 4-8)
  - Probability (discussed in Lectures 9-14)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 14-16)
    - 1 quantitative (discussed in Lectures 16-18)
    - cat and quan: paired, 2-sample, several-sample (Lectures 19-21)
    - 2 categorical
    - 2 quantitative

Explanatory/Response: 2 Categorical Variables

- Roles impact what summaries to report
- Roles do *not* impact χ^2 statistic or *P*-value

Example: *Summaries Impacted by Roles*

- **Background**: Compared proportions alcoholic (resp) for smokers and non-smokers (expl).

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Nonsmoker	10	760	770
Total	40	960	1,000

$$\hat{p}_1 = \frac{30}{230} = 0.130$$

$$\hat{p}_2 = \frac{10}{770} = 0.013$$

$$\frac{30}{40} = 0.75 \qquad \frac{200}{960} = 0.21$$

- **Question:** What summaries would be appropriate if alcoholism is explanatory variable?
- **Response:** Compare proportions _____ (resp) for _____ (expl).

Example: *Comparative Summaries*

- **Background**: Calculated proportions for table:

	Alcoholic	Not Alcoholic	Total
Smoker	30	200	230
Nonsmoker	10	760	770
Total	40	960	1,000

$\hat{p}_1 = \frac{30}{230} = 0.130$

$\hat{p}_2 = \frac{10}{770} = 0.013$

$\frac{30}{40} = 0.75$ $\frac{200}{960} = 0.21$

- **Question:** How can we express the higher risk of alcoholism for smokers and the higher risk of smoking for alcoholics?
- **Response:** Smokers are ___ times as likely to be alcoholics compared to non-smokers. Alcoholics are _____ times as likely to be smokers compared to non-alcoholics.

Guidelines for Use of Chi-Square Procedure

- Need random samples taken independently from several populations.
- Confounding variables should be separated out.
- Sample sizes must be large enough to offset non-normality of distributions.
- Need populations at least 10 times sample sizes.

Rule of Thumb for Sample Size in Chi-Square

- Sample sizes must be large enough to offset non-normality of distributions.

Require expected counts all at least 5 in 2×2 table

(Requirement adjusted for larger tables.)

Looking Back: *Chi-square statistic follows chi-square distribution only if individual counts vary normally. Our requirement is extension of requirement for single categorical variables* $np \geq 10, n(1-p) \geq 10$ *with 10 replaced by 5 because of* ***summing*** *several components.*

Example: *Role of Sample Size*

- **Background**: Suppose counts in smoking and alcohol two-way table were 1/10th the originals:

	Alcoholic	Not Alcoholic	Total
Smoker	3	20	23
Nonsmoker	1	76	77
Total	4	96	100

- **Question:** Find chi-square; what do we conclude?
- **Response:** Observed counts 1/10th → expected counts 1/10th →chi-square _____ instead of 64.

  But the statistic does **not** follow χ^2 distribution because expected counts (0.92, 22.08, 3.08, 73.92) are _____; individual distributions are **not** normal.

Confidence Intervals for 2 Categorical Variables

Evidence of relationship→ to what extent does explanatory variable affect response?

Focus on **proportions**: 2 approaches

- Compare confidence intervals for population proportion in response of interest (one interval for each explanatory group)
- Set up confidence interval for difference between population proportions in response of interest, 1st group minus 2nd group

Example: *Confidence Intervals for 2 Proportions*

- **Background**: Individual CI's are constructed:
  - **Non-smokers** 95% CI for pop prop *p* alcoholic (0.005,0.021)
  - **Smokers** 95% CI for pop prop *p* alcoholic (0.09,0.17)
- **Question:** What do the intervals suggest about relationship between smoking and alcoholism?
- **Response:** Overlap?____

  Relationship between smoking and alcoholism?

  _____ (____ likely to be alcoholic if a smoker).

Practice: 12.36a-b p.621

Example: *Difference between 2 Proportions (CI)*

- **Background**: 95% CI for difference between population proportions alcoholic, smokers minus non-smokers is (0.088, 0.146)
- **Question:** What does the interval suggest about relationship between smoking and alcoholism?
- **Response:** Entire interval _____ suggests smokers _____ significantly more likely to be alcoholic → there _____ a relationship.

H_0 and H_a for 2 Cat. Variables *(Review)*

- In terms of variables
  - H_0: two categorical variables are not related
  - H_a: two categorical variables are related
- In terms of parameters
  - H_0: population proportions in response of interest are equal for various explanatory groups
  - H_a: population proportions in response of interest are not equal for various explanatory group

Word "not" appears in H_0 about variables, H_a about parameters.

Chi-Square Statistic

- Compute table of counts expected if H_0 true: each is

$$\text{Expected} = \frac{\text{Column total} \times \text{Row total}}{\text{Table total}}$$

  - Same as counts for which proportions in response categories are equal for various explanatory groups

- Compute **chi-square** test statistic χ^2

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

"Observed" and "Expected"

Expressions "observed" and "expected" commonly used for chi-square hypothesis tests.

More generally, "observed" is our sample statistic, "expected" is what happens on average in the population when H_0 is true, and there is no difference from claimed value, or no relationship.

Variable(s)	Observed	Expected
1 Categorical	$\hat{p}$	p_o
1 Quantitative	$\bar{x}$	μ_o
1 Cat & 1 Quan	$\bar{x}_d$ $\bar{x}_1 - \bar{x}_2$	0 0
2 Categorical	Observed Counts	Expected Counts

Example: *2 Categorical Variables: Data*

- **Background**: We're interested in the relationship between gender and lenswear.

	contacts	glasses	none	All
female	121	32	129	282
	42.91%	11.35%	45.74%	100.00%
male	42	37	85	164
	25.61%	22.56%	51.83%	100.00%
All	163	69	214	446

- **Question:** What do data show about sample relationship?
- **Response:** Females wear contacts more (_____ vs. _____); males wear glasses more (_____ vs. _____); proportions with none are close (_____ vs. _____).

Example: *Table of Expected Counts*

- **Background**: We're interested in the relationship between gender and lenswear.

Expected	Contacts	Glasses	None	Total
Female				282
Male				164
Total	163	69	214	446

- **Question:** What counts are expected if gender and lenswear are not related?
- **Response:** Calculate each expected count as

Example: *"Eyeballing" Obs. and Exp. Tables*

- **Background**: We're interested in the relationship between gender & lenswear.

Chi-square procedure: Compare counts observed to counts expected if null hypothesis were true

Observed	Contacts	Glasses	None	Total
Female	121	32	129	282
Male	42	37	85	164
Total	163	69	214	446

Expected	Contacts	Glasses	None	Total
Female	103	44	135	282
Male	60	25	79	164
Total	163	69	214	446

- **Question:** Do observed and expected counts seem very different?
- **Response:**

Example: *Components for Comparison*

- **Background**: Observed and expected tables:

Observed	Contacts	Glasses	None	Total
Female	121	32	129	282
Male	42	37	85	164
Total	163	69	214	446

Expected	Contacts	Glasses	None	Total
Female	103	44	135	282
Male	60	25	79	164
Total	163	69	214	446

- **Question:** What are the components of chi-square?
- **Response:** Calculate each

Example: *Components for Comparison*

- **Background**: Components of chi-square are

$$\frac{(121-103)^2}{103} = 3.1 \qquad \frac{(32-44)^2}{44} = 3.3 \qquad \frac{(129-135)^2}{135} = 0.3$$

$$\frac{(42-60)^2}{60} = 5.4 \qquad \frac{(37-25)^2}{25} = 5.8 \qquad \frac{(85-79)^2}{79} = 0.5$$

- **Questions:** Which contribute most and least to the chi-square statistic? What is chi-square? Is it large?
- **Responses:**
  - _____ largest: most impact from _____
  - _____ smallest: least impact from _____
  - _____
  - _____

Chi-Square Distribution *(Review)*

chi-square = sum of $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ follows predictable pattern known as

chi-square distribution with df = $(r\text{-}1) \times (c\text{-}1)$

- r = number of rows (possible explanatory values)
- c = number of columns (possible response values)

Properties of chi-square:

- Non-negative (based on squares) [= 0 when…?]
- Mean=df [=1 for smallest (2×2) table]
- Spread depends on df
- Skewed right

Example: *Chi-Square Degrees of Freedom*

- **Background**: Table for gender and lenswear:

Observed	Contacts	Glasses	None	Total
Female	121	32	129	282
Male	42	37	85	164
Total	163	69	214	446

- **Question:** How many degrees of freedom apply?
- **Response:** row variable (male or female) has $r =$ __, column variable (contacts, glasses, none) has $c =$ __. df = _____

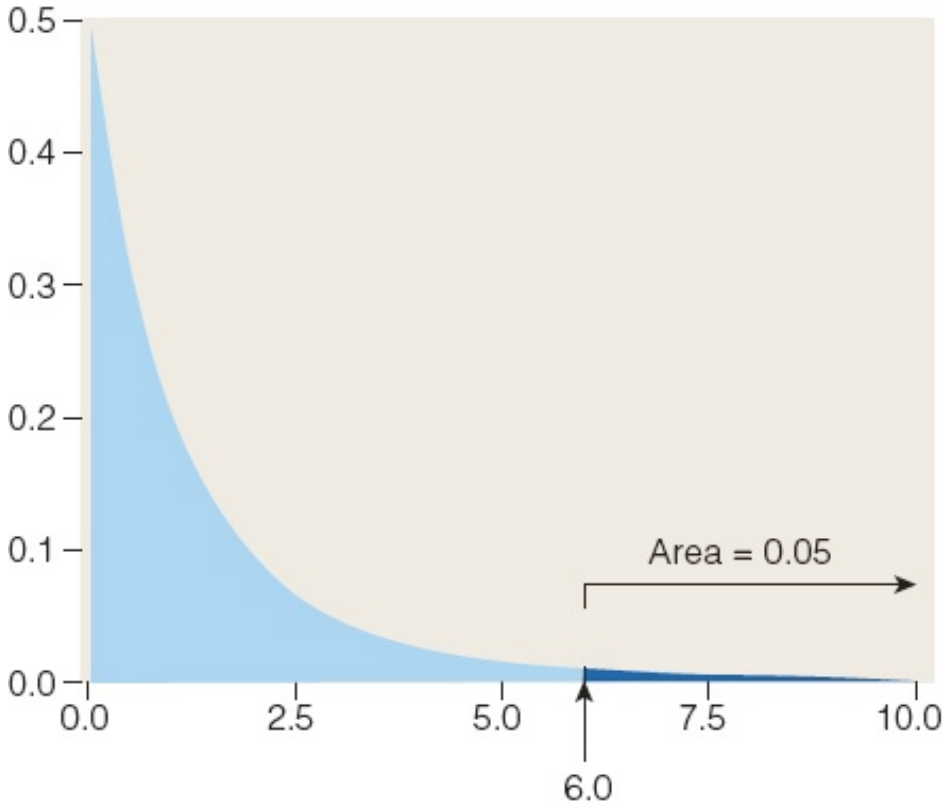
A Closer Look: *Degrees of freedom tell us how many unknowns can vary freely before the rest are "locked in."*

Practice: 12.42a p.623

Chi-Square Density Curve

For chi-square with 2 df, $P(\chi^2 \geq 6) = 0.05$

→ If χ^2 is more than 6, *P*-value is less than 0.05.



Chi-square with 2 df (for 2-by-3 table)

Example: *Assessing Chi-Square*

- **Background**: In testing for relationship between gender and lenswear in 2×3 table, found $\chi^2 = 18.4$.
- **Question:** Is there evidence of a relationship in general between gender and lenswear (not just in the sample)?
- **Response:** For df = (2-1)×(3-1) = 2, chi-square is considered "large" if greater than 6. Is 18.6 large?

  ____ Is the *P*-value small? ____

  Is there statistically significant evidence of a relationship between gender and lenswear? ____

Example: *Checking Assumptions*

- **Background**: We produced table of expected counts below right:

Observed	Contacts	Glasses	None	Total
Female	121	32	129	282
Male	42	37	85	164
Total	163	69	214	446

Expected	Contacts	Glasses	None	Total
Female	103	44	135	282
Male	60	25	79	164
Total	163	69	214	446

- **Question:** Are samples large enough to guarantee the individual distributions to be approximately normal, so the sum of standardized components follows a χ^2 distribution?
- **Response:**

Example: *Chi-Square with Software*

- **Background**: Some subjects injected under arm with Botox, others with placebo. After a month, reported if sweating had decreased.

```
Expected counts are printed below observed counts
     Decreased NotDecreased Total
Botox      121         40      161
         80.50      80.50
Placebo     40        121      161
         80.50      80.50
Total      161        161      322
Chi-Sq = 20.376 + 20.376 +
         20.376 + 20.376 = 81.503
DF = 1, P-Value = 0.000
```

- **Question:** What do we conclude?
- **Response:** Sample sizes large enough?____Proportions with reduced sweating _____ Seem different?

  ____*P*-value = ____ ➔diff significant? ____

  Conclude Botox reduces sweating? ____

Guidelines for Use of Chi-Square *(Review)*

- Need random samples taken independently from two or more populations.
- Confounding variables should be separated out.
- Sample sizes must be large enough to offset non-normality of distributions.
- Need populations at least 10 times sample sizes.

Example: *Confounding Variables*

- **Background**: Students of all years: $\chi^2 = 13.6, p = 0.000$

	On Campus	Off Campus	Total	Rate On Campus
Undecided	124	81	205	124/205=60%
Decided	96	129	225	96/225=43%

Underclassmen: $\chi^2 = 0.025, p = 0.873$

Underclassmen	On Campus	Off Campus	Total	Rate On Campus
Undecided	117	55	172	117/172=68%
Decided	82	37	119	82/119=69%

Upperclassmen: $\chi^2 = 1.26, p = 0.262$

Upperclassmen	On Campus	Off Campus	Total	Rate On Campus
Undecided	7	26	33	7/33=21%
Decided	14	92	106	14/106=13%

- **Question:** Are major (dec or not) and living situation related?
- **Response:**

Activity

- Complete table of total students of each gender on roster, and count those attending and not attending for each gender group. Carry out a chi-square test to see if gender and attendance are related in general.

90-707	Attend	Not Attend	Total
Female			
Male			
Total			

Lecture Summary

(Inference for Cat→Cat; Chi-Square)

- Explanatory/response roles in chi-square test
- Guidelines for use of chi-square
- Role of sample size
- Confidence intervals for 2 categorical variables

Lecture Summary
(Inference for Cat→Cat; More Chi-Square)

- Hypotheses about variables or parameters
- Computing chi-square statistic
  - Observed and expected counts
- Chi-square test
  - Calculations
  - Degrees of freedom
  - Chi-square density curve
  - Checking assumptions
  - Testing with software
- Confounding variables