Lecture 32: Chapter 12, Sections 1-2
Two Categorical Variables
Chi-Square

- Formulating Hypotheses to Test Relationship
- Test based on Proportions or on Counts
- Chi-square Test
- Confidence Intervals

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing (Lectures 5-12)
  - Probability (discussed in Lectures 13-20)
  - Statistical Inference
    - 1 categorical (discussed in Lectures 21-23)
    - 1 quantitative (discussed in Lectures 24-27)
    - cat and quan: paired, 2-sample, several-sample (Lectures 28-31)
    - 2 categorical
    - 2 quantitative

Inference for Relationship *(Review)*

- H_0 and H_a about variables: not related or related
  - Applies to all three C→Q, C→C, Q→Q
- H_0 and H_a about parameters: equality or not
  - C→Q: pop means equal?
  - C→C: pop proportions equal?
  - Q→Q: pop slope equals zero?

Example: *2 Categorical Variables: Hypotheses*

- **Background**: We are interested in whether or not smoking plays a role in alcoholism.
- **Question:** How would H_0 and H_a be written
  - in terms of variables?
  - in terms of parameters?
- **Response:**
  - in terms of variables
    - H_0: smoking and alcoholism related
    - H_a: smoking and alcoholism related
  - in terms of parameters
    - H_0: Pop proportions alcoholic for smokers, non-smokers
    - H_a: Pop. proportions alcoholic for smokers, non-smokers

The word "not" appears in Ho about variables, in Ha about parameters.

Practice: 12.24 p.616

Example: *Summarizing with Proportions*

- **Background**: Research Question: Does smoking play a role in alcoholism?
- **Question:** What statistics from this table should we examine to answer the research question?
- **Response:** Compare proportions ________ (response) for ________________ (explanatory).

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

Example: *Test Statistic for Proportions*

- **Background**: One approach to the question of whether smoking and alcoholism are related is to compare proportions.

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

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

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

- **Question:** What would be the next step, if we've summarized the situation with the difference between sample proportions 0.130-0.013?
- **Response:** ____________ the difference between sample proportions 0.130-0.013.

Stan. diff. is normal for large n:

Practice: 12.2a-b p.595

z Inference for 2 Proportions: Pros & Cons

Advantage:

Can test against *one-sided* alternative.

Disadvantage:

2-by-2 table: comparing proportions straightforward

Larger table: comparing proportions complicated, can't just standardize one difference $\hat{p}_1 - \hat{p}_2$

Another Comparison in Considering Categorical Relationships *(Review)*

- 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.
- Compared observed, expected counts in wasp study:

Obs	A	NA	T
B	16	15	31
U	24	7	31
T	40	22	62

Exp	A	NA	T
B	20	11	31
U	20	11	31
T	40	22	62

Inference Based on Counts

To test hypotheses about relationship in r-by-c table, compare counts observed to counts expected if H_0(equal proportions in response of interest) were true.

Example: *Table of Expected Counts*

- **Background**: Data on smoking and alcoholism:

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

- **Question:** What counts are expected if H_0 is true?
- **Response:** Overall proportion alcoholic is

If proportions alcoholic were same for S and NS, expect

- (40/1,000)(230)=_____ smokers to be alcoholic
- (40/1,000)(770)=_____ non-smokers to be alcoholic; also
- (960/1,000)(230)=_____ smokers not alcoholic
- (960/1,000)(770)=_____ non-smokers not alcoholic

Practice: 12.12d p.609

Example: *Table of Expected Counts*

- **Background**: If proportions alcoholic were same for S and NS, expect
  - (40/1,000)(230)= 9.2 smokers to be alcoholic
  - (40/1,000)(770)= 30.8 non-smokers to be alcoholic; also
  - (960/1,000)(230)= 220.8 smokers not alcoholic
  - (960/1,000)(770)= 739.2 non-smokers not alcoholic
- **Question:** Where do they appear in table of expected counts?
- **Response:**

	Alcoholic	Not Alcoholic	Total
Smoker			230
Nonsmoker			770
Total	40	960	1,000

Note:
9.2/230 =
30.8/770 =
40/1,000

Practice: 12.12e p.609

Example: *Table of Expected Counts*

	Alcoholic	Not Alcoholic	Total
Smoker	9.2	220.8	230
Non-smoker	30.8	739.2	770
Total	40	960	1000

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

  Expect:
  - (40)(230)/1,000 = 9.2 smokers to be alcoholic
  - (40)(770)/1,000 = 30.8 non-smokers to be alcoholic; also
  - (960)(230)/1,000 = 220.8 smokers not alcoholic
  - (960)(770)/1,000 = 739.2 non-smokers not alcoholic

Chi-Square Statistic

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

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

Components are **individual** standardized squared differences.

- **Chi-square** test statistic χ^2 combines all components by summing them up:

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

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

Example: *Chi-Square Statistic*

- **Background**: Observed and Expected Tables:

Obs	A	NA	Total
S	30	200	230
NS	10	760	770
Total	40	960	1000

Exp	A	NA	Total
S	9.2	220.8	230
NS	30.8	739.2	770
Total	40	960	1000

- **Question:** What is the chi-square statistic?
- **Response:** Find $\text{chi-square} = \text{sum of } \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

=

Example: *Assessing Chi-Square Statistic*

- **Background**: We found chi-square = 64.
- **Question:** Is the chi-square statistic (64) large?
- **Response:**

Chi-Square Distribution

$\text{chi-square} = \text{sum of } \frac{(\text{observed - expected})^2}{\text{expected}}$ follows a predictable pattern (assuming H_0 is true) known as **chi-square distribution** with df = (r-1) × (c-1)

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

Properties of chi-square:

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

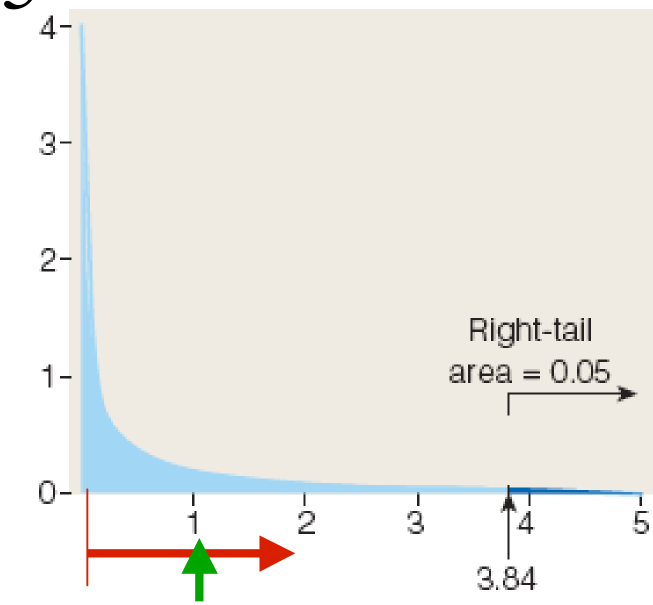
Chi-Square Density Curve

For chi-square with 1 df, $P(\chi^2 \geq 3.84) = 0.05$

→ If $\chi^2 > 3.84$, P-value < 0.05

Properties of chi-square:

- Non-negative
- Mean = df
  df=1 for smallest [2×2] table
- Spread depends on df
- Skewed right



Chi-square with 1 df (for 2-by-2 table)

Example: *Assessing Chi-Square (Continued)*

- **Background**: In testing for relationship between smoking and alcoholism in 2×2 table, found $\chi^2 = 64$
- **Question:** Is there evidence of a relationship in general between smoking and alcoholism (not just in the sample)?
- **Response:** For df=(2-1)×(2-1)=1, chi-square considered "large" if greater than 3.84
  →chi-square=64 large? *P*-value small?
  Evidence of a relationship between smoking and alcoholism?

Inference for 2 Categorical Variables; z or χ^2

For 2×2 table, $z^2 = \chi^2$

- z statistic (comparing proportions)→ combined tail probability=0.05 for z=1.96
- chi-square statistic (comparing counts)→ right-tail prob=0.05 for $\chi^2 = 1.96^2 = 3.84$

Example: *Relating Chi-Square & z*

- **Background**: We found chi-square = 64 for the 2-by-2 table relating smoking and alcoholism.
- **Question:** What would be the z statistic for a test comparing proportions alcoholic for smokers vs. non-smokers?
- **Response:**

Practice: 12.10 p.608

Assessing Size of Test Statistics *(Summary)*

When test statistic is "large":

- z: greater than 1.96 (about 2)
- t: depends on df; greater than about 2 or 3
- F: depends on DFG, DFE
- χ^2 depends on df=$(r-1)\times(c-1)$;

  greater than 3.84 (about 4) if df=1

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).

Practice: 5.98a,c p.215

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

$$\hat{p}_1 = \frac{30}{230} = 0.130$$
$$\hat{p}_2 = \frac{10}{770} = 0.013$$

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

- **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.

Practice: 12.12m p.609

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.

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

- Hypotheses in terms of variables or parameters
- Inference based on proportions or counts
- Chi-square test
  - Table of expected counts
  - Chi-square statistic, chi-square distribution
  - Relating z and chi-square for 2×2 table
  - Relative size of chi-square statistic
  - Explanatory/response roles in chi-square test
- Guidelines for use of chi-square
- Role of sample size
- Confidence intervals for 2 categorical variables