Lecture 10: Chapter 5, Section 2 Relationships (Two Categorical Variables)

- Two-Way Tables
- Summarizing and Displaying
- Comparing Proportions or Counts
- Confounding Variables

Looking Back: *Review*

- 4 Stages of Statistics
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing
    - Single variables: 1 cat,1 quan (discussed Lectures 5-8)
    - Relationships between 2 variables:
      - Categorical and quantitative (discussed in Lecture 9)
      - Two categorical
      - Two quantitative
  - Probability
  - Statistical Inference

Single Categorical Variables *(Review)*

- **Display:**
  - Pie Chart
  - Bar Graph
- **Summarize**:
  - Count or Proportion or Percentage

Add categorical explanatory variable →

display and summary of categorical responses are **extensions** of those used for single categorical variables.

Example: *Two Single Categorical Variables*

- **Background**: Data on students' gender and lenswear (contacts, glasses, or none) in two-way table:

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

- **Question:** What parts of the table convey info about the individual variables gender and lenswear?
- **Response:**
  - is about gender.
  - is about lenswear.

Example: *Relationship between Categorical Variables*

- **Background**: Data on students' gender and lenswear (contacts, glasses, or none) in two-way table:

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

- **Question:** What part of the table conveys info about the *relationship* between gender and lenswear?
- **Response:** is about relationship.

Summarizing and Displaying Categorical Relationships

- Identify variables' roles (explanatory, response)
- Use rows for explanatory, columns for response
- Compare proportions or percentages in response of interest *(conditional proportions or percentages)* for various explanatory groups.
- Display with bar graph:
  - Explanatory groups identified on horizontal axis
  - Conditional percentages or proportions in response(s) of interest graphed vertically

Definition

- A **conditional** percentage or proportion tells the percentage or proportion in the response of interest, given that an individual falls in a particular explanatory group.

Example: *Comparing Counts vs. Proportions*

- **Background**: Data on students' gender and lenswear (contacts, glasses, or none) in two-way table:

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

- **Question:** Since 129 females and 85 males wore no lenses, should we report that fewer males wore no lenses?
- **Response:**
  - proportion of females with no lenswear:
  - proportion of males with no lenswear:

Practice: 5.26 p.162

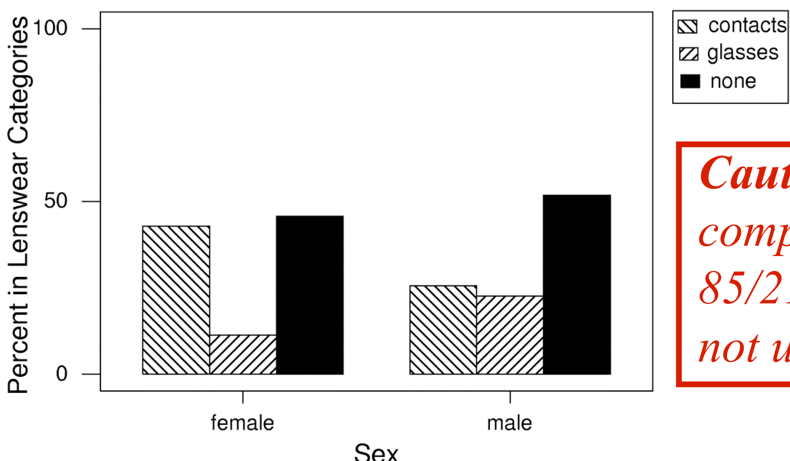
Example: *Displaying Categorical Relationship*

- **Background**: Counts and conditional percentages produced with software:

Rows: Gender Columns: 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:** How can we display this information?
- **Response:**



Caution: *If we made lenswear explanatory, we'd compare 129/214 =**60%** with no lenses female, 85/214= **40%** with no lenses male, etc. Why is this not useful?*

Example: *Interpreting Results*

- **Background**: Counts and conditional percentages produced with software:

```
Rows: Gender      Columns: 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
```

- **Questions:** Are you convinced that, in general,
  - all females wear contacts more than males do?
  - all males are more likely to wear no lenses?
- **Responses:** Consider *how* different sample percentages are:
  - Contacts:
  - No lenses

Looking Ahead: *Inference will let us judge if sample differences are large enough to suggest a general trend. For now, we can guess that the first difference is "real", due to different priorities for importance of appearance.*

Practice: 5.20d-g p.161

Example: *Comparing Proportions*

- **Background**: An experiment considered if wasp larvae were less likely to attack an embryo if it was a brother:

	Attacked	Not attacked	Total
Brother	16	15	31
Unrelated	24	7	31
Total	40	22	62

- **Question:** What are the relevant proportions to compare?
- **Response:**
  - Brother: ____________________ were attacked
  - Unrelated: were attacked

→ likely to attack a brother wasp

Practice: 5.32 p.164

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.

Example: *Expected Counts*

- **Background**: Experiment considered if wasp larvae were less likely to attack embryo if it was a brother:

	Attacked	Not attacked	Total
Brother	16	15	31
Unrelated	24	7	31
Total	40	22	62

- **Question:** What counts would we expect to see, if being a brother had no effect on likelihood of attack?
- **Response:** Overall 40/62 attacked→ expect

brothers,

unrelated to be attacked; expect

remaining brothers and unrelated not to be attacked.

Practice: 5.29a p.163

Example: *Comparing Counts*

- **Background**: Tables of observed and expected counts in wasp aggression experiment:

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

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

Looking Ahead: *Inference (Part 4) will help decide if these differences are large enough to provide evidence that kinship and aggression are related.*

Practice: 5.29b p.163

Example: *Expected Counts in Lenswear Table*

- **Background**: Data on students' gender and lenswear (contacts, glasses, or none) in two-way table:

	C	G	N	Total
F	121	32	129	282
M	42	37	85	164
Total	163	69	214	446

- **Question:** What counts would we expect to wear glasses, if there were no relationship between gender and lenswear?
- **Response:** Altogether, 69/446 wore glasses. If there were no relationship, we'd expect females and males with glasses.

Practice: 5.31d p.164

Example: *Observed vs. Expected Counts*

- **Background**:If gender and lenswear were unrelated, we'd expect 44 females and 25 males with glasses.

	C	G	N	Total
F	121	32	129	282
M	42	37	85	164
Total	163	69	214	446

- **Question:** How different are the observed and expected counts of females and males with glasses?
- **Response:** Considerably females and males wore glasses, compared to what would be expected if there were no relationship.

Practice: 5.31e p.164

Confounding Variable in Categorical Relationships

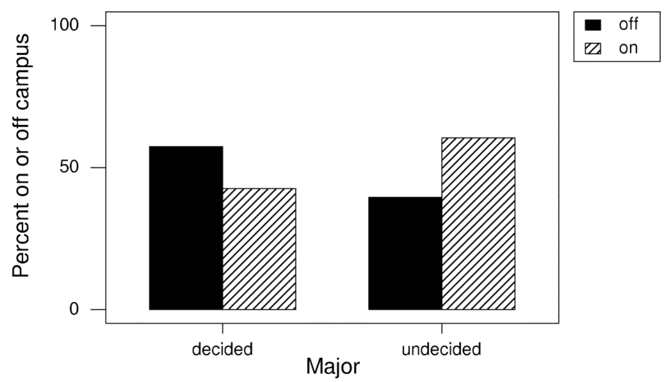
- If data in two-way table arise from an observational study, consider possibility of confounding variables.

Looking Back: *Sampling and Design issues should always be considered before reporting summaries of single variables or relationships.*

Example: *Confounding Variables*

- **Background**: Survey results for full-time students:

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

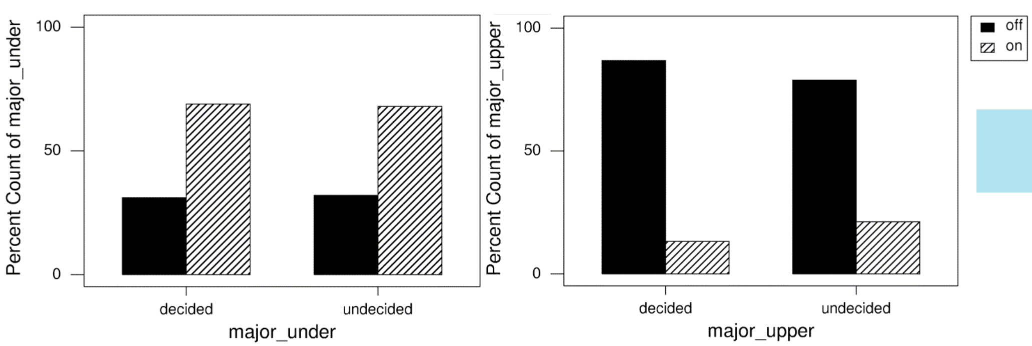


- **Question:** Is there a relationship between whether or not major is decided and living on or off campus?
- **Response:**

Example: *Handling Confounding Variables*

- **Background**: Year at school may be confounding variable in relationship between major decided or not and living situation.
- **Question:** How should we handle the data?
- **Response:**

Underclassmen	On Campus	Off Campus	Total	Rate On Campus
Undecided	117	55	172	117/172=68%
Decided	82	37	119	82/119=69%
Upperclassmen	On Campus	Off Campus	Total	Rate On Campus
Undecided	7	26	33	7/33=21%
Decided	14	92	106	14/106=13%



Underclassmen (1st&2nd yr): proportions on campus are for those with major decided or not. Upperclassmen (3rd &4th yr): proportions are

Practice: 5.32c p.164

Simpson's Paradox

If the nature of a relationship changes, depending on whether groups are combined or kept separate, we call this phenomenon "Simpson's Paradox".

Example: *Considering Confounding Variables*

- **Background**: Suppose that boys, like Bart, tend to eat a lot of sugar and they also tend to be hyperactive. Girls, like Lisa, tend not to eat much sugar and they are less likely to be hyperactive.
- **Question:** Why would the data lead to a misperception that sugar causes hyperactivity?
- **Response:**

Lecture Summary
(Categorical Relationships)

- **Two-Way Tables**
  - Individual variables in margins
  - Relationship inside table
- **Summarize:** Compare (conditional) proportions.
- **Display:** Bar graph
- **Interpreting Results:** How different are proportions?
- **Comparing Observed and Expected Counts**
- **Confounding Variables**