# DATA ANALYSIS

Week 14: Chi-square tests

# upcoming **review** sessions

- Sunday (Yanevith): 3.30 pm - 5 pm
- Tuesday (Whitt): 4.15 pm - 5.45 pm
- Wednesday (in class)
- Wednesday (Prof. Kumar): 2 - 5 pm
- Thursday (Prof. Kumar): 10 – 4 pm
- Thursday (Yanevith): 7.30 pm – 9 pm
- poll for submitting questions

| | | |
|---|---|---|
| 14 | F: April 26, 2024 | W14 continued… |
| 15 | T: April 30, 2024 | **Problem Set 7 due / Opt-out Deadline** |
| 15 | W: May 1, 2024 | W15: Odds and Ends |
| 15 | T: May 2, 2024 | **Data Around Us / Practice Questions due** |
| 15 | F: May 3, 2024 | **Conceptual Final (In Class)** |
| 16 | T: May 7, 2024 | **Computational Final Computational due** |
| 16 | T: May 7, 2024 | **Last Class Survey due** |
| 16 | W: May 8, 2024 | **Wrapping Up!** (Last Class) |
| 17 | T: May 14, 2024 | **PS7 Revisions due** |
| 17 | M: May 14, 2024 | **ALL late work due** |

# parametric vs. non-parametric tests

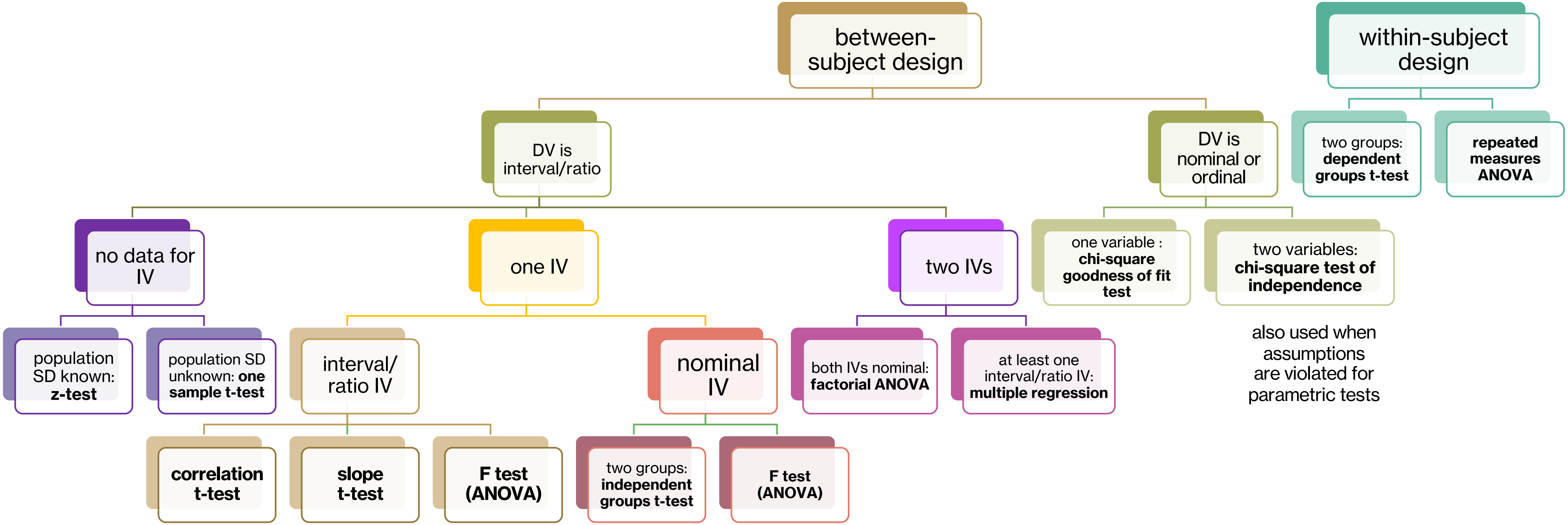
**parametric tests**

- interval/ratio DVs
- involve estimating parameters
- assumptions about the underlying sampling distribution
- if assumptions are violated, these tests may not be appropriate

**non-parametric tests**

- assume no underlying distributions (”distribution-free”)
- typically used for nominal/ordinal DVs that yield counts
- no assumptions about underlying population
- most parametric tests have a non-parametric alternative

# final hypothesis chart

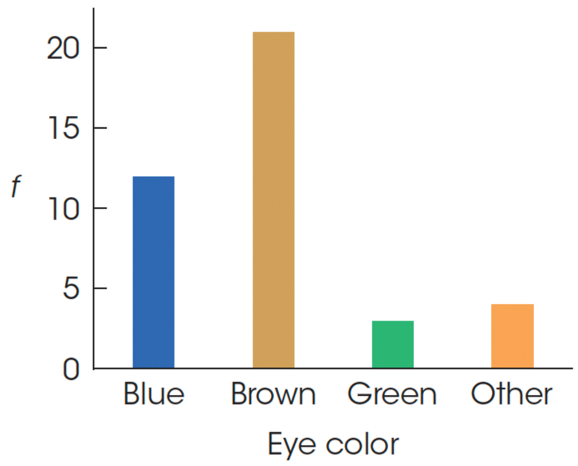


# chi-square tests

- chi-square goodness of fit test
  - one nominal/ordinal variable
  - asks whether observed distribution of responses matches hypothesized distribution
- chi-square test of independence
  - two nominal/ordinal variables
  - asks whether observed distribution of responses on one variable depends on responses on other variable

# example: eye color

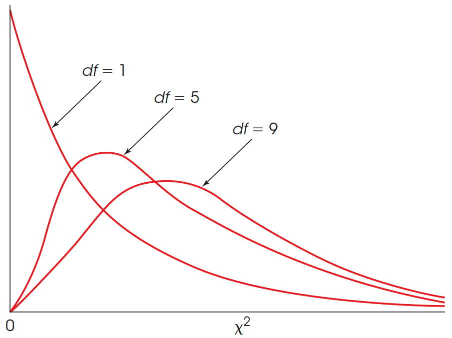
- eye color counts for 40 students
- can be represented in a bar graph or frequency distribution table
- counts typically converted to a table
- observed values/counts are then compared to expected values/counts via a ratio
- asking: how extreme are the differences between what is expected and what is observed?



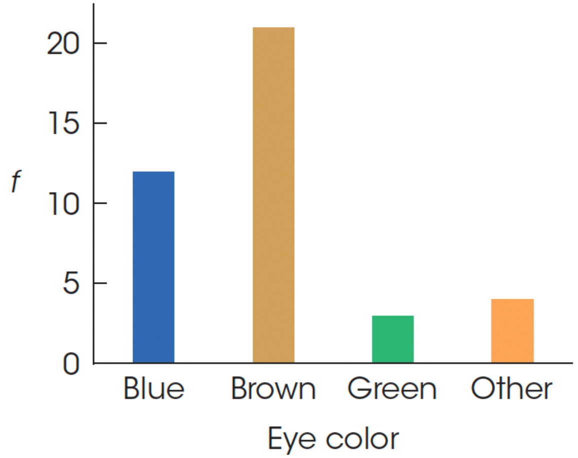

| Eye color (*X*) | *f* |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | | | | |

# chi-square goodness of fit test




- $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$
- the "expected" frequencies form the null hypothesis ($H_0$)
  - equal preference (all counts equal)
  - known population (specific distribution)
- observed $\chi^2$ statistic is then compared to the expected distribution for a set degrees of freedom based on number of categories C
  - $df = C - 1$




| Eye color (X) | f |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | 10 | 10 | 10 | 10 |

$$f_e = \frac{N}{C} \quad for\ equal\ preference$$

# NHST for chi-square goodness of fit test

**step 1: state the hypotheses**

$H_0: equal\ preference\ OR\ known\ distribution$

$H_1: distribution\ does\ not\ match\ expected\ distribution$

**step 2: set criteria for decision**

$\alpha = .05$

find $\chi^2{}_{critical}$ based on **<u>right tailed</u>** test and degrees of freedom

$df = C - 1$

**step 3: collect data**

(1) find observed frequencies $f_o$

(2) find expected frequencies $f_e$

$f_e = \frac{N}{C}\ for\ equal\ preference$

$f_e = N\ (p_k)\ for\ expected\ proportions$

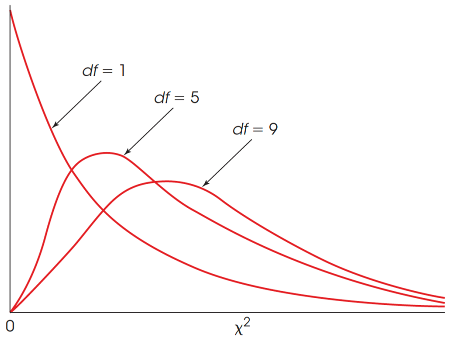
(3) compute $\chi^2{}_{observed} = \sum \frac{(f_o - f_e)^2}{f_e}$

(4) find p-value for $\chi^2{}_{observed}$

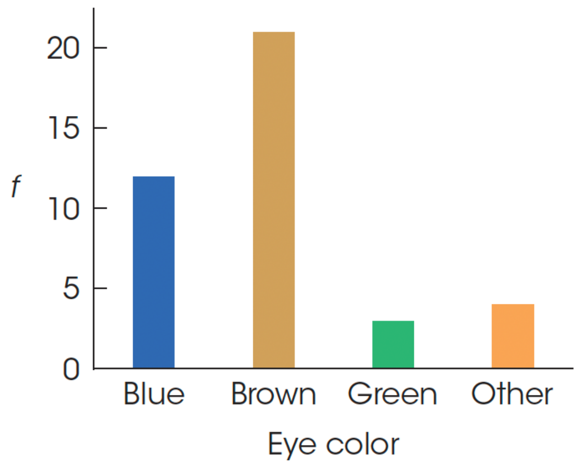
**step 4: make a decision!**

check whether $\chi^2{}_{observed}$ is beyond $\chi^2{}_{critical}$ and p-value < .05. if so, reject null hypothesis!

# chi-square goodness of fit test




- conduct the test
- $C = 4$
- $df = C - 1 = 3$
- $\chi^2{}_{critical}\ (3) = 7.8147$
- $\chi^2{}_{observed} = \sum \frac{(f_o - f_e)^2}{fe} = 21$
- p-value < .0001
- APA reporting: A significant difference was observed in eye color distributions, $\chi^2\ (3, n = 40) = 21, p < \ .0001$



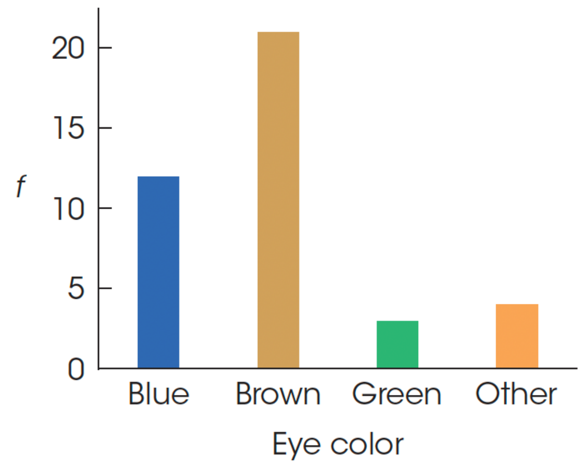

| Eye color (X) | f |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | 10 | 10 | 10 | 10 |

# known distribution

| Eye Color | U.S. Population | World Population |
|---|---|---|
| Gray and other | Less than 1% | Less than 1% |
| Green | 9% | 2% |
| Hazel/amber | 18% | 10% |
| Blue | 27% | 8% to 10% |
| Brown | 45% | 55% to 79% |

- has eye color significantly changed in the US population since 2000?
- our hypothesis is no longer about equal preference, but instead about a known population distribution
- $f_e = N\ (p_k)\ for\ expected\ proportions$
- $f_e\ (blue) = 40\ (.27) = 10.8$
- $f_e\ (other) = 40\ (.18 +\ .01) = 7.6$

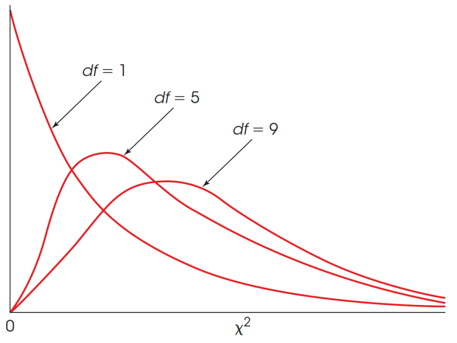



| Eye color (X) | f |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

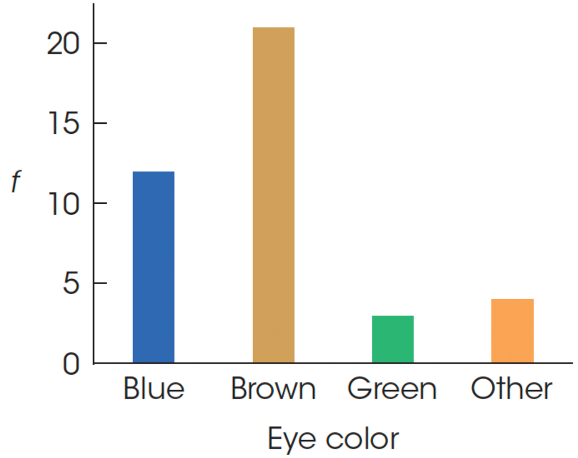
| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | 10.8 | 18 | 3.6 | 7.6 |

$$f_e = N\ (p_k)\ for\ expected\ proportions$$

# chi-square goodness of fit test




- conduct the test
- $C = 4$
- $df = C - 1 = 3$
- $\chi^2{}_{critical}\ (3) = 7.8147$
- $\chi^2{}_{observed} = \sum \frac{(f_o - f_e)^2}{fe} = 2.438$
- p-value = 0.4865
- APA reporting: Eye color distributions have not significantly changed since 2000, $\chi^2\ (3, n = 40) = 2.43, p = .49$




| Eye color (X) | f |
|---|---|
| Blue | 12 |
| Brown | 21 |
| Green | 3 |
| Other | 4 |

| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | 10.8 | 18 | 3.6 | 7.6 |

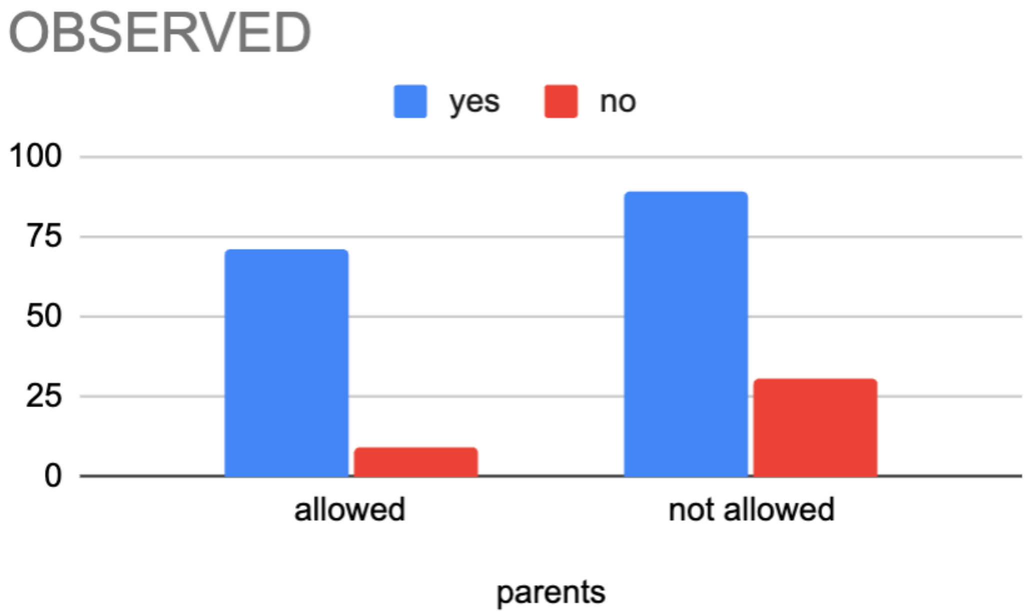
# chi-square test for independence

- is parent-allowed alcohol use related to how many alcohol-related problems are experienced?
- typically, this is a situation where there is no clear IV/DV but a relationship needs to be tested
- note that variables are no longer interval/ratio: these are COUNTS

| OBSERVED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | |
| parents allowed alcohol use | allowed | 71 | 9 | |
| | not allowed | 89 | 31 | |
| | | | | |

# chi-square test for independence

- is parent-allowed alcohol use related to how many alcohol-related problems are experienced?
- typically, this is a situation where there is no clear IV/DV but a relationship needs to be tested
- note that variables are no longer interval/ratio: these are COUNTS



# chi-square test for independence

- is parent-allowed alcohol use related to how many alcohol-related problems are experienced?
- typically, this is a situation where there is no clear IV/DV but a relationship needs to be tested
- note that variables are no longer interval/ratio: these are COUNTS

| OBSERVED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | |
| parents allowed alcohol use | allowed | 71 | 9 | |
| | not allowed | 89 | 31 | |
| | | | | |

# chi-square test for independence

- we first count up the totals to get how many people were sampled and how many were in each level

| OBSERVED frequencies | | experienced alcohol-related problems | | |
| --- | --- | --- | --- | --- |
| | | yes | no | total |
| parents allowed alcohol use | allowed | 71 | 9 | 80 |
| | not allowed | 89 | 31 | 120 |
| | total | 160 | 40 | N = 200 |

# expected frequencies

- what proportion of students experienced problems?
  - 160 / 200 = .80
- if problems experienced **are not related to whether parents allowed alcohol use or not**, then 80% of the students should experience problems and 20% shouldn't
  - expected (allowed-yes) = .80 * 80 = 64
  - expected (allowed-no) = .20*80 = 16

| EXPECTED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | total |
| parents allowed alcohol use | allowed | | | 80 |
| | not allowed | | | 120 |
| | total | 160 | 40 | N = 200 |
| | | .80 | .20 | |

# expected frequencies

- what proportion of students experienced problems?
  - 160 / 200 = .80
- if problems experienced **are not related to whether parents allowed alcohol use or not**, then 80% of the students should experience problems and 20% shouldn't
  - expected (allowed-yes) = .80 * 80 = 64
  - expected (allowed-no) = .20*80 = 16

| EXPECTED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | total |
| parents allowed alcohol use | allowed | 64 | 16 | 80 |
| | not allowed | | | 120 |
| | total | 160 | 40 | N = 200 |
| | | .80 | .20 | |

# expected frequencies

- what proportion of students experienced problems?
  - 160 / 200 = .80
- if problems experienced **are not related to whether parents allowed alcohol use or not**, then 80% of the students should experience problems and 20% shouldn't
  - expected (not allowed-yes) = .80 * 120 = 96
  - expected (not allowed-no) = .20*120 = 24

| EXPECTED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | total |
| parents allowed alcohol use | allowed | 64 | 16 | 80 |
| | not allowed | | | 120 |
| | total | 160 | 40 | N = 200 |
| | | .80 | .20 | |

# expected frequencies

- what proportion of students did NOT experience problems?
  - 40 / 200 = .20
- if problems experienced **are not related to whether parents allowed alcohol use or not**, then 80% of the students should experience problems and 20% shouldn't
  - expected (not allowed-yes) = .80 * 120 = 96
  - expected (not allowed-no) = .20*120 = 24

| EXPECTED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | total |
| parents allowed alcohol use | allowed | 64 | 16 | 80 |
| | not allowed | 96 | 24 | 120 |
| | total | 160 | 40 | N = 200 |
| | | .80 | .20 | |

# NHST for chi-square test of independence

**step 1: state the hypotheses**

$H_0$: *no relationship between variables*

$H_1$: *there is a relationship between variables*

**step 2: set criteria for decision**

$\alpha = .05$

find $\chi^2_{critical}$ based on **right tailed** test and degrees of freedom

$df = (R-1)(C-1)$

**step 3: collect data**

(1) find observed frequencies $f_o$

(2) find expected frequencies $f_e$ based on proportions

(3) compute $\chi^2_{observed} = \sum \frac{(f_o - f_e)^2}{f_e}$

(3) find p-value for $\chi^2_{observed}$

**step 4: make a decision!**

check whether $\chi^2_{observed}$ is beyond $\chi^2_{critical}$ and p-value < .05. if so, reject null hypothesis!

# activity

- compute the expected frequencies

# chi-square test

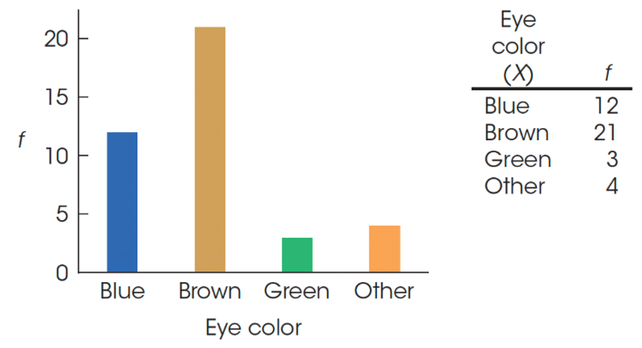
- $df = (R-1)(C-1)$
- $df = (2-1)(2-1) = 1$
- $\chi^2{}_{critical}\ (1) = 3.84$
- $\chi^2{}_{observed} = \sum \frac{(f_o - f_e)^2}{f_e} = 6.38$
- p-value = 0.0115

| OBSERVED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | total |
| parents allowed alcohol use | allowed | 71 | 9 | 80 |
| | not allowed | 89 | 31 | 120 |
| | total | 160 | 40 | N = 200 |

| EXPECTED frequencies | | experienced alcohol-related problems | | |
|---|---|---|---|---|
| | | yes | no | |
| parents allowed alcohol use | allowed | 64 | 16 | 80 |
| | not allowed | 96 | 24 | 120 |
| | | 160 | 40 | N = 200 |

# chi-square test: assumptions

- independence of observations (between-subject measurements)
- expected frequencies in each cell > 5
- typically categories are merged if counts are low




| | blue | brown | green | other |
|---|---|---|---|---|
| observed ($f_o$) | 12 | 21 | 3 | 4 |
| expected ($f_e$) | 10.8 | 18 | 3.6 | 7.6 |

$$f_e = N\,(p_k)\ for\ expected\ proportions$$