chi-square independence test

- two categorical variables,
  at least 1 with >2 levels

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

obesity and marital status

- A study reported in the medical journal *Obesity* in 2009 analyzed data from the National Longitudinal Study of Adolescent Health.
- Obesity was defined as having a BMI of 30 or more.
- The research subjects were followed from adolescence to adulthood, and all the people in the sample were categorized in terms of whether they were obese and whether they were dating, cohabiting, or married.

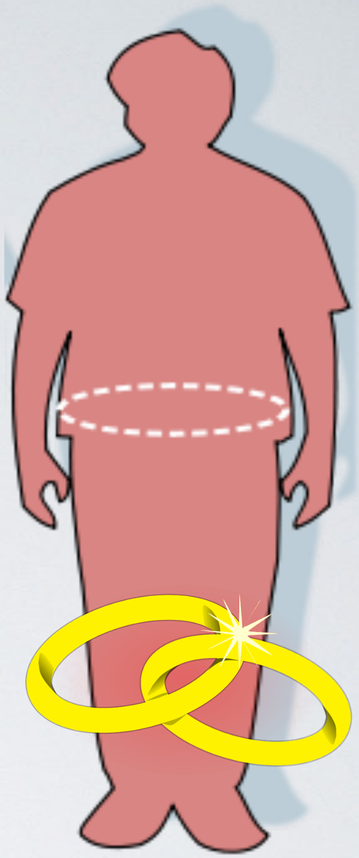

Image source:

Overweight: http://commons.wikimedia.org/wiki/File:Obesity-waist_circumference.PNG

Rings: http://openclipart.org/detail/182973/wedding-rings-by-Jarno-182973

study results

	dating	cohabiting	married	total
obese	81	103	147	331
not obese	359	326	277	962
total	440	429	424	1293

Does there appear to be a relationship between weight and relationship status?

hypotheses

H_0 (nothing going on): Weight and relationship status are independent.
Obesity rates do not vary by relationship status.

H_A (something going on): Weight and relationship status are dependent.
Obesity rates do vary by relationship status.

evaluating the hypotheses

- quantify how different the observed counts are from the expected counts
- large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- called an independence test since we're evaluating the relationship between two categorical variables

chi-square test of independence

χ^2 test of independence:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

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

O : observed E : expected

k : number of cells

R : number of rows

C : number of columns

Conditions for the chi-square test:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
   - each case only contributes to one cell in the table
2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

practice

What is the overall obesity rate in the sample?

331 / 1293 = 0.256

	dating	cohabiting	married	total
obese	81	103	147	331
not obese	359	326	277	962
total	440	429	424	1293

If in fact weight and relationship status are independent (i.e. if in fact H_0 is true) how many of the dating people would we expect to be obese? How many of the cohabiting and married?

dating: $440 \times 0.256 \approx 113$

cohabiting: $429 \times 0.256 \approx 110$

married: $424 \times 0.256 \approx 108$

$113 + 110 + 108 = 331$ ✔

expected counts in two-way tables

$$expected\ count\ =\ \frac{(row\ total) \times (column\ total)}{table\ total}$$

practice

Test the hypothesis that relationship status and obesity are associated at the 5% significance level.

	dating	cohabiting	married	total
obese	81 **(113)**	103 **(110)**	147 **(108)**	331
not obese	359 **(327)**	326 **(319)**	277 **(316)**	962
total	440	429	424	1293

$$\chi^2 = \frac{(81-113)^2}{113} + \frac{(103-110)^2}{110} + \frac{(147-108)^2}{108} + \frac{(359-327)^2}{327} + \frac{(326-319)^2}{319} + \frac{(277-316)^2}{316}$$

$$= 31.68$$

$$df = (2-1) \times (3-1) = 1 \times 2 = 2$$

R

```
> pchisq(31.68, 2, lower.tail = FALSE)
[1] 1.320613e-07
```

Can we conclude from these data that living with someone is making some people obese, and that marrying someone is making people even more obese?

No!

chi-square tests

- goodness of fit: comparing the distribution of one categorical variable (with more than 2 levels) to a hypothesized distribution
- independence: evaluating the relationship between two categorical variables (at least one with more than 2 levels)