5.8: Putting It Together- Relationships in Categorical Data with Intro to Probability

Let's Summarize

To summarize the relationship between two categorical variables, use:

- A data display: A two-way table
- Numerical summaries: Conditional percentages

When we investigate the relationship between two categorical variables, we use the values of the explanatory variable to define the comparison groups. We then compare the distributions of the response variable for values of the explanatory variable. In particular, we look at how the pattern of conditional percentages differs between the values of the explanatory variable.

> For example, we investigated the relationship between body image and gender. We compared males to females. For each gender, we determined the percentage who felt their body weight was about right, overweight, or underweight. *P*(body image "about right" | male) is compared to *P*(body image "about right" | female).

Keys Ideas from Our Work with Probability

We defined three kinds of probabilities related to a two-way table:

- A **marginal probability** is the probability of a categorical variable taking on a particular value *without regard to the other categorical variable*. For example, *P*(Health Sciences) is the probability that a student is enrolled in the Health Sciences program. In calculating the probability, we use overall student data contained in the margins of the table. A marginal probability is a row or column total divided by the table total.
- A **conditional probability** is the probability of a categorical variable taking on a particular value *given the condition that the other categorical variable has some particular value*. For example, *P*(Health Sciences **given** female) means we look first at all females, then identify the female students who are Health Science students. In calculating the probability, we use only a subset of the data. The condition determines the subset of data we use. If our condition relates to female students, then we consider only the information in the table pertaining to females.
- A **joint probability** is the probability that the *two categorical variables each take on a specific value*. For example: *P*(male **and** Info Tech) is the probability that a student is both a male and in the Info Tech program. In calculating this probability, we divide the count from one inner cell of the table by the overall total count (in the lower right corner.)

When we calculate the probability of a **negative outcome**, we often refer to the probability as a **risk**. We compare risk by calculating the percentage change (divide difference in risks by risk in placebo group).

Finally, we created hypothetical two-way tables to compute complex probabilities, such as the probability of a positive drug test for someone who does not use drugs.

Contributors and Attributions

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*