3.1 Association on Two Categorical Variables

- association : when two variables are related in some way.
- response variable : is the outcome variable on which comparisons are made.
- explanatory variable : defines the groups to be compared with respect to values on the response variable
- grade school goals : students were surveyed (1992) in Michigan for a study. students indicated which of grades, popularity and sports was most important, students were in grades 4 - 6.

	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

given gender, how likely is each of the goal categories? in this case, goal is the response variable and gender the explanatory variable.

	Grades	Popular	Sports
Boy	0.515	0.22	0.264
Girl	0.518	0.362	0.119

given the goal category, how likely is each gender? in this case, gender is the response variable and goal is the explanatory variable.

	Grades	Popular	Sports
Boy	0.473	0.354	0.667
Girl	0.526	0.645	0.333

what would the table look like if there was no association?

- problems 3.1 and 3.2

3.2 Association on Two Quantitative Variables

- three scenarios
  - both categorical variables
    - use contingency tables as in section 3.1
  - one quantitative variable and one categorical variable
    - compare categories using summaries of center and spread (i.e. $\bar{x}$, s, quartiles, etc.)
    - graphics such as side by side boxplots

  - both variables are quantitative
    - analyze how the outcome on the response variable tends to change as the value of the explanatory variable changes. the subject of the rest of the chapter.

- scatterplots a graphical display for two quantitative variables. the explanatory variable is plotted on the x-axis and the response variable on the y-axis. explanatory variables are sometimes called the independent variable whereas the response variables are called the dependent variable.
- positive association is when high values of x occur with high values of y
- negative association is when high values of x occur with low values of y
- correlation : summarizes the direction of the association and the strength of its linear (straight line) trend.
  - positive values for r indicate positive association, likewise for negative values.
  - a correlation value $r = 1, -1$ indicates a perfect linear relationship in the data.
  - $r = 0$ indicates a very weak linear relationship.
  - correlation does not depend on the variables' units
  - correlation does not care which variable is treated as the response variable.
- examples (transparency)
- we say that there is association in cases where the correlation is close to 1 or -1
- equation to calculate correlation is related to the z score. the z score tells us how many standard deviations an observation is from the mean.

  $$z_x = \frac{x - \bar{x}}{s_x}$$

  then the correlation is given as

  $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- the equation highlights some properties of the correlation
  - remember that the z score is positive when an observation is above the mean and negative when an observation is below the mean
  - this means that when z_x and z_y are both positive or both negative then the contribution to the sum is positive
  - when the z scores have opposite sign, the contribution is negative
- roughly speaking, the correlation tells us how often the observations are above or below their means together.
- always check the data with a scatterplots when calculating the correlation. correlation will give bad values for anything but a linear relationship.

3.3 Prediction

- we have shown that association between two quantitative variables can follow a linear relationship
- we may want to use this relationship to predict values of the response (y) variable for some value(s) of the explanatory variable. (example)
- review of the equation for a line
  - the equation of a line is given as

    $$\hat{y} = a + bx$$

  - we will refer to this as the prediction or regression equation (because of the hat)
  - a is called the intercept
  - b is called the slope
- example 8
- interpretation
  - a may not have any interpretive value. it represents the value of $\hat{y}$ when $x = 0$.
  - b is the amount $\hat{y}$ changes when x changes by one unit.
- example 9
- prediction error
- residuals : the vertical distance between an observation and its predicted value

  $$e_i = y_i - \hat{y}_i$$

- residuals are positive when the observation lies above the regression line and negative when below
- similarity to categorical analysis : regression predicts y at a given value of x similar to analyzing a contingecy table where we study the response variable conditional on a given explanatory variable.
- finding the slope and intercept is done using the method of least squares. the idea is to minimize the residual sum of squares. the exact derivation is done using calculus, so you'll just have to take my word for it.
- the equations are

  $$a = \bar{y} - b\bar{x}$$

  $$b = r\left(\frac{s_y}{s_x}\right)$$

- why use regression instead of some other prediction method based on $\bar{y}$ and s_y? if there is an association between x and y, regression will give a much better prediction of the response variable.

- slope, correlation, units

  - correlation is unit-less, regression (slope) will depend on the units used in the explanatory variable.
  - correlation coefficient is the same regardless of the choice of response vs explanatory variable. regression coefficients will be different with different choices.
  - both appropriate when looking for a linear relationship.
  - slope and correlation will have the same sign
  - correlation falls between -1 and 1, slope can be any real number.

- r^2 is used to determine the strength of the regression line (called fit)

- interpretation of r^2 : prediction error is r^2% smaller than the prediction error using $\bar{y}$ to predict y.

- compares the variance of the y values to the variance around the regression line.

- r^2 is more difficult to interpret because it is related to variance which does not have the same units as the y variable. it is brought up here to make you aware of it and to have a working knowledge of what it means. it will be discussed in more detail in ch 11 (if we get there).

3.4 Cautions

- extrapolation

- cautious of influential outliers

- correlation $\neq$ causation

- lurking variables