correlation

- Pearson product-moment correlation coefficient
    - a scatterplot shows paired data, e.g., X and Y
    - correlation coefficient: ρ (population), r (sample); range is -1 to 1
    - direction of correlation can be positive, negative
    - strength of correlation: weak (near zero), strong (near ±1)
        - up to 0.2, weak; 0.2 to 0.5, moderate; 0.5 or more, strong

$$\rho = (1/N) \sum z_x * z_y \qquad \text{(population)}$$

$$r = (1/(n-1)) \sum z_x * z_y \qquad \text{(sample)}$$

    - correlation does not imply causation
    - Pearson correlation captures <u>linear</u> relationships between variables

- coefficient of determination: r^2
    - measures "shared variance", or "proportion of variance explained"
    - so in a sense, r^2 tells us how much of the variation in X is due to, or "can be explained by" variation in Y

- statistically significant correlations
    - null hypothesis: $\rho = 0$; alternative hypothesis: $\rho \neq 0$
    - use t distribution
        s_r = sqrt((1 - r^2) / (n - 2))
        t = (r - ρ) / s_r = r / s_r (assuming $\rho = 0$)
        df = n - 2

- example (in box on significance testing)
    - test correlation between hours of sunlight (0-24) and mood (1-10)
    - r = 0.25, n = 100
        s_r = sqrt((1 - 0.25^2) / 98) = 0.0978
        t = (0.25 - 0) / 0.0978 = 2.5562
        df = 100 - 2 = 98
    - critical t value (two-tailed, alpha = 0.05, df = 98): $t_c = 2.00$
    - reject the null hypothesis
    - note! a weak correlation can be highly statistically significant

- other types of correlation coefficients
    - Pearson coefficient requires interval or ratio scales
    - point-biserial correlation: one dichotomous variable (two categories), one continuous variable
    - phi: two dichotomous variables
    - Spearman ρ: two rank variables