IX. **Bivariate Methods:** \((X_i, Y_i) \sim f_{X,Y}(x, y)\); determine whether there is association between these variables.

A. Test whether \(f_{X,Y}(x, y) = f_X(x) f_Y(y)\).

1. Want test not to require knowledge of \(f_{X,Y}\), or even null values.
2. Against alternative that, vaguely, large values of \(X\) are associated with large values of \(Y\) (or vice versa).

B. Also, measure strength of association.

C. Parametric Approach: Pearson Correlation

\[
\rho_p = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.
\]

This gives the slope of the least squares line fitting \(Y\) to \(X\), after scaling both variables by standard deviation.

3. Cauchy-Schwartz says that this is always in \([-1, 1]\).
   a. Perfect positive or negative linear association is reflected in 1 or -1 resp.

4. Inference:
   a. \((X_j, Y_j)\) independent bivariate normal, unit variance, \(j = 1, n\).
   b. Numerator of \(\rho^2_p\) is multiple of \(\chi^2\).
   c. Denominator of \(\rho^2_p\) is multiple of \(\chi^2_{n-1}\).
   d. Numerator square is one of squares in denominator
   e. So \(t = \sqrt{n-2}r_p / \sqrt{1-r^2_p} \sim t_{n-2}\), even when \(\rho \neq 0\).

D. Variance of \(r_p\):

1. Fix the \(X\)'s, and rearrange the \(Y\)'s.
2. For permutation distribution of \(Y\)'s, note \(\text{Cov}[Y_1, Y_2] = \sum_{i \neq j} \frac{1}{n(n-1)}(y_i - \bar{y})(y_j - \bar{y}) = \sum_{i,j} \frac{1}{n(n-1)}(y_i - \bar{y})(y_j - \bar{y}) - \sum_{i} \frac{1}{n(n-1)}(y_i - \bar{y})^2 = \frac{1}{n-1}\sigma^2_Y\), for \(\sigma^2_Y = \sum \frac{1}{n} (y_i - \bar{y})^2\).

3. So \(\text{Var} \left[\sum_j (X_j - \bar{X})Y_j\right] = \sum_j (X_j - \bar{X})^2 \text{Var}[Y_j] + \sum_{j \neq i} (X_i - \bar{X})(X_j - \bar{X}) \text{Cov}[Y_j, Y_i] = \sum_j (X_j - \bar{X})^2 \sigma^2_Y = \frac{\sum_j (X_i - \bar{X})(X_j - \bar{X}) - \sum_i (X_i - \bar{X})^2}{n-1} \sigma^2_Y = \frac{\sum_j (X_i - \bar{X})^2 + \sum_i (X_i - \bar{X})^2 \sigma^2_Y / (n-1) = n^2 \sigma^2_Y / (n-1) - \frac{n-1}{2} \sigma^2_Y}{n-1} = 1 - \frac{4Q^1}{n(n-1)}\), \(Q^1 = n(n-1)/2 - Q\) is number of discordant pairs; equals number of rearrangements necessary to make all pairs concordant.
   a. \(s(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{otherwise} \end{cases}\)
   b. Number of discordant pairs is \(n(n-1)/2 - Q^1\).
   c. Estimate \(\tau\) by the excess of concordant over discordant pairs, divided by maximum:
   \[
r_{\tau} = \frac{(n(n-1)/2 - Q^1 - Q^1)}{n(n-1)/2} = 1 - \frac{4Q^1}{n(n-1)}.
\]

3. Then \(E[Q^1] = \frac{1}{2} n(n-1)p_1\), for \(p_1 = P[(X_1 - X_2)(Y_2 - Y_1) > 0]\).

4. \(E[r_{\tau}] = 1 - 2p_1\).

5. Null value of \(p_1\) is half.
6. \[ \text{Var}[Q^1] = \sum_{i<j} \text{Var}[Z_{ij}] + \sum_{i<j}^* \text{Cov}[Z_{ij}, Z_{kj}] \]
   a. \( \sum^* \) the sum over \( i < j, k < l, 3 \) distinct indices.
   i. Consists of \( \frac{1}{4}n(n-1)^2 - \frac{1}{2}n(n-1) - \frac{1}{4}n(n-1)(n-2)(n-3) = \frac{1}{4}n(n-1)^2 - \frac{1}{2}n(n-1) + \frac{1}{4}n = n(n-1)(n-2) \) terms.
   b. \( Z_{ij} = (X_i - X_j)(Y_j - Y_i) \).
   c. \[ \text{Var}[Q^1] = \sum_{i<j} \text{Var}[Z_{ij}] + \sum_{i<j}^* \text{Cov}[Z_{12}, Z_{13}] \]
      \[ = \sum_{i<j} (p_1 - p_1^2) + \sum_{i<j}^* (p_3 - p_1^2) \]
      \[ = \frac{1}{2}n(n-1)(p_1 - p_1^2) + n(n-1)(n-2)(p_4 - p_1^2) \]

X. Bivariate semi-parametric methods:

A. Recall Spearman’s Rank Correlation \( r_s \) as correlation of ranks of \( X \)’s and \( Y \)’s. Motivation was under the alternative large \( X \)’s would be associated with large \( Y \)’s, or with small \( Y \)’s.

B. Assume a linear model \( Y_j = \beta_1 + \beta_2 X_i + e_i, e_i \) has some distribution with median 0 (making \( \beta_1 \) identifiable).

C. Two objectives:
   1. We already know how to test \( \beta_2 = 0 \);
   2. Now we estimate \( \beta_2 \).

D. Get confidence interval by inverting test of \( \tau = \tau(\beta_2) = 0 \), based on \( r_\tau \).
   1. Could also do this using Spearman.
   2. Could also do this using Pearson correlation under permutation distribution.
   3. Could also do this using permutation distribution for \( \sum_{j=1}^n (X_j - \bar{X}) R_j \) for \( R_j \) the rank of the residuals.

E. Steps to invert the test using correlation between \( X_i \) and \( Y_i - \beta_2 X_i \), as a function of \( \beta_2 \)
   1. Calculate \( q_{\alpha/2} \) and \( q_{1-\alpha/2} \): lower and upper 2.5% points of distribution of \( r, r_S, r_\tau \).
      a. Since we know null mean and variance, this could be approximated using normal distribution.
   2. Solve \( r(\beta_2) = q_{\alpha/2} \quad r(\beta_2) = q_{1-\alpha/2} \).

F. Both parts are tricky for Pearson correlation (with permutation distribution).
   1. Distribution is not on a unit lattice, so that exact calculations are next to impossible.
   2. \( r \) has steps of irregular size, and so translating \( q_{\alpha/2} \) and \( q_{1-\alpha/2} \) into values of \( \beta_2 \) is hard.

This page intentionally left blank.