Wilks, Chapter 3: Exploratory data analysis

Robustness (not sensitive to assumptions), resistance (to outliers)

Mean, not a robust/resistant measure
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

Median, quartiles, quantiles are robust/resistant

Spread Sample Standard Deviation, not robust
\[ s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

Inter Quartile Range, robust \[ IQR = q_{0.75} - q_{0.25} \]

Skewness (lack of symmetry) \[ \gamma = \frac{1}{n-1} \sqrt{(x_i - \bar{x})^3 / s^3} \sim 0 \text{ if symmetric, is not robust.} \]
\[ \tilde{\gamma} = \left[ (q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25}) \right] / IQR \]
\[ = \left[ q_{0.75} - 2q_{0.5} + q_{0.25} \right] / IQR \]
is robust.

Frequency distribution (histograms), cumulative frequency distribution
Data pairs – Exercise: do scatterplots like Fig. 3.11/3.15

Covariance
\[ C(x, y) = \sum_{i=1}^{n} \frac{1}{n-1} (x_i - \bar{x})(y_i - \bar{y}) \]

Correlation
\[ r(x, y) = \rho_{xy} = \frac{C(x, y)}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{x_i'y_i'}{\sqrt{x_i'^2 y_i'^2}} \]

Rank Correlation: Rank (order) them first, then correlate rank. This will give \( \rho_1 \approx 1, \rho_2 \approx 0 \), physically much more meaningful: it is robust/resistant.

Lag-correlation: \( r_k = \rho(x_t, y_{t-k}) \), Lag auto-correlation \( r_k = \rho(x_t, x_{t-k}) \)

Higher dimensional data

Correlation matrix

Covariance matrix

Correlation maps (Fig. 3.19/3.27)

Teleconnections (Fig. 3.20/3.28)