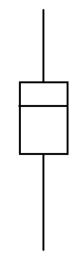
Wilks, Chapter 3: Exploratory data analysis

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

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

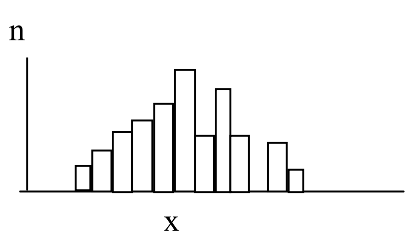
Inter Quartile Range, robust $IQR = q_{0.75} - q_{0.25}$

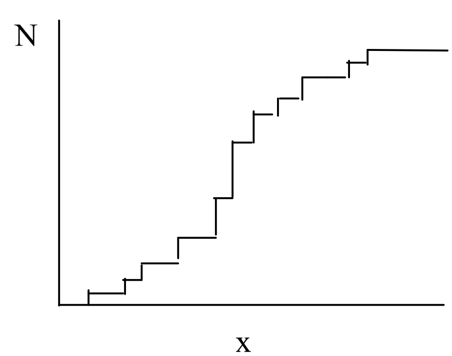
Skewness (lack of symmetry) $\gamma = \frac{1}{n-1}\sqrt{(x_i - \bar{x})^3} \,/\, s^3$ ~zero if symmetric, is not robust.

$$\tilde{\gamma} = \left[(q_{.75} - q_{.5}) - (q_{.5} - q_{.25})\right] / IQR = \left[q_{.75} - 2q_{.5} + q_{.25})\right] / IQR$$

is robust.

Frequency distribution (histograms), cumulative frequency distribution

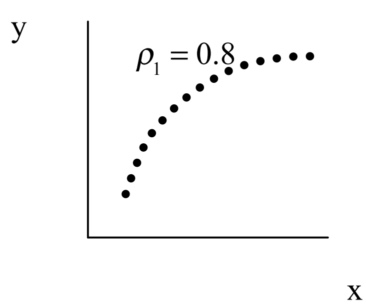


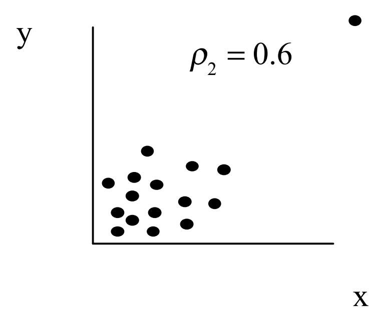


Data pairs – Exercise: do scatterplots like Fig. 3.11/3.15

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

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





Rank Correlation: Rank (order) them first, then correlate rank. This will give $\rho_1 \sim 1, \quad \rho_2 \sim 0$, 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)