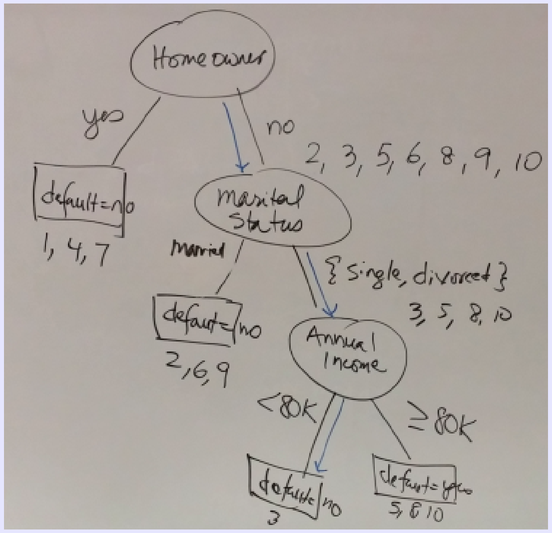
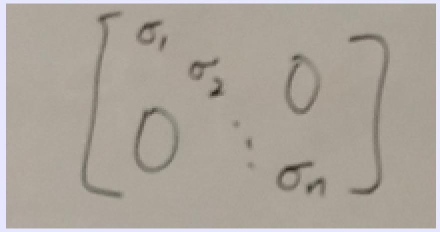
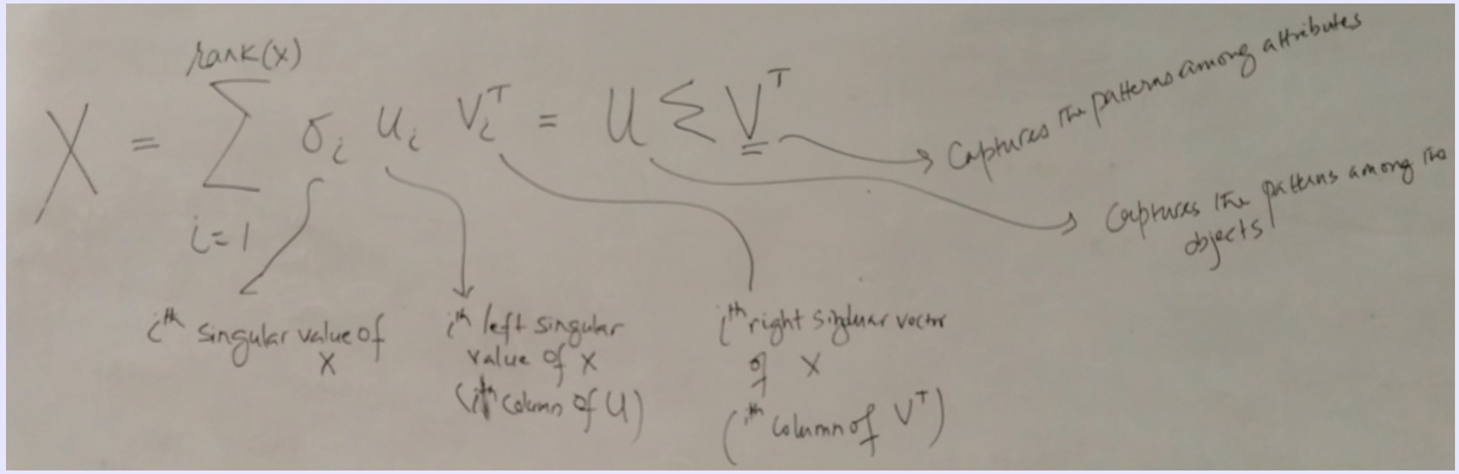
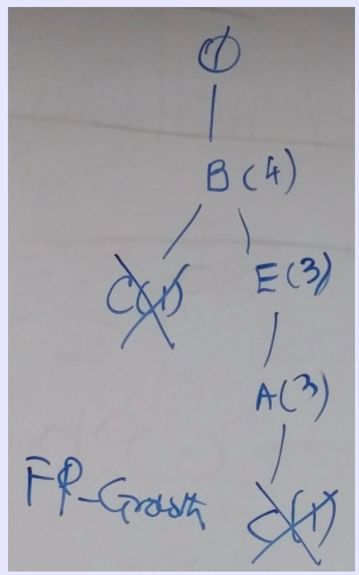
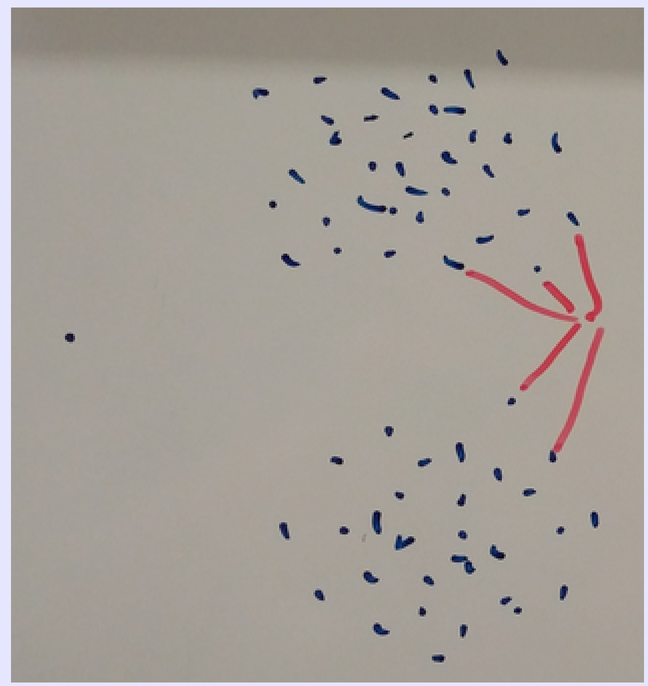










CS 422: Data Mining

Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

Lecture 2: **Random variables, measures of central tendency and distributions**

Random variables

- A random variable, *X*, is a variable whose possible values are drawn from the outcome of a random phenomenon.
  - Tossing a coin
  - Tossing a die
- Two types:
  - Discrete
  - Continuous
- We will use mostly discrete r.v.

Random variables

- Population versus sample
  - Consider a column vector, D:

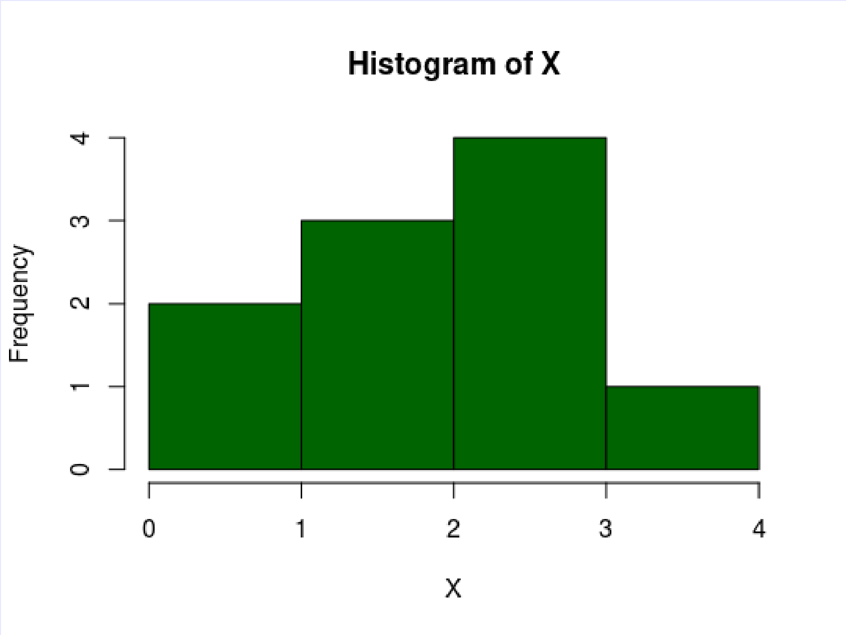
$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad X \in \mathbb{R}^n$$

  - We assume that the observed data is a random sample drawn from *X*, each x_i is *iid.*
  - In general, the distribution from which *X* is drawn is unknown, as are the moments.
    - All we have is the sample, from which we will derive the distribution and moments, which (hopefully) are close to the population distribution and moments.

Random variables

- Suppose a discrete variable *X* can take the values 1, 2, 3, 4 with the following probabilities:
  - 1 : 0.2; 2 : 0.3; 3 : 0.4; 4: 0.10

Then the probability mass function can be described by the following equation and histogram:



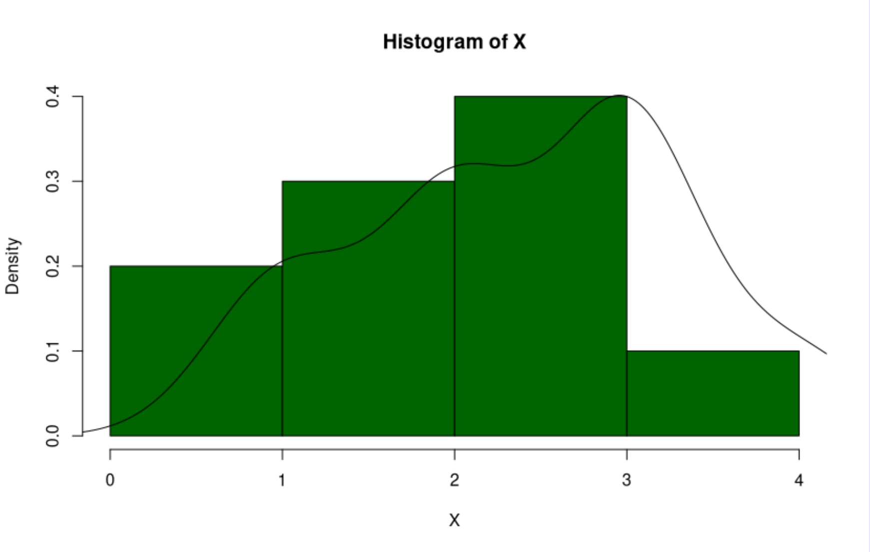
$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

Random variables

- Associated with a PMF (discrete) is a PDF (continous).
- A density plot visualizes the underlying probability distribution of the data by drawing an appropriate continuous curve.
  - Curve estimated from data using *kernel density* estimation.




```
> hist(X, c(0.0, 1.0, 2.0, 3.0, 4.0), prob=T, col="darkgreen")
> lines(density(X))
```

Measures of central tendency

- Mean (sample):
  - $\hat{\mu} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$
- Is the mean robust (or stable)?
  - We define *robustness* as the tendency not to be affected by extreme values.

Measures of central tendency

- Mean (sample):
  - $\hat{\mu} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$
- Is the mean robust (or stable)?
  - We define *robustness* as the tendency not to be affected by extreme values.
  - Generally, the mean is not robust.
  - A robust measure is *trimmed mean*, which occurs after extreme values on either side are discarded.

Measures of central tendency

- Expectation of a r.v. (related to mean, but conceptually different).
  - $$E[X] = \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$
- Properties of expectation:
  - E[X+Y] = E[X]+E[Y] (Linearity of expectation)
  - E[*a*X] = *a*E[X] *a is a constant*
  - E[XY] = E[X] * E[Y] *iff X and Y are iid*
  - E[E[X]] = E[X]

Measures of central tendency

- Median (sample):
  - $P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$
- Is the median robust (or stable)?

Measures of central tendency

- Median (sample):
  - $P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$
- Is the median robust (or stable)?
  - Yes.
    - Not affected by extreme values.
    - Also, an actual value that a r.v. takes.

Measures of central tendency

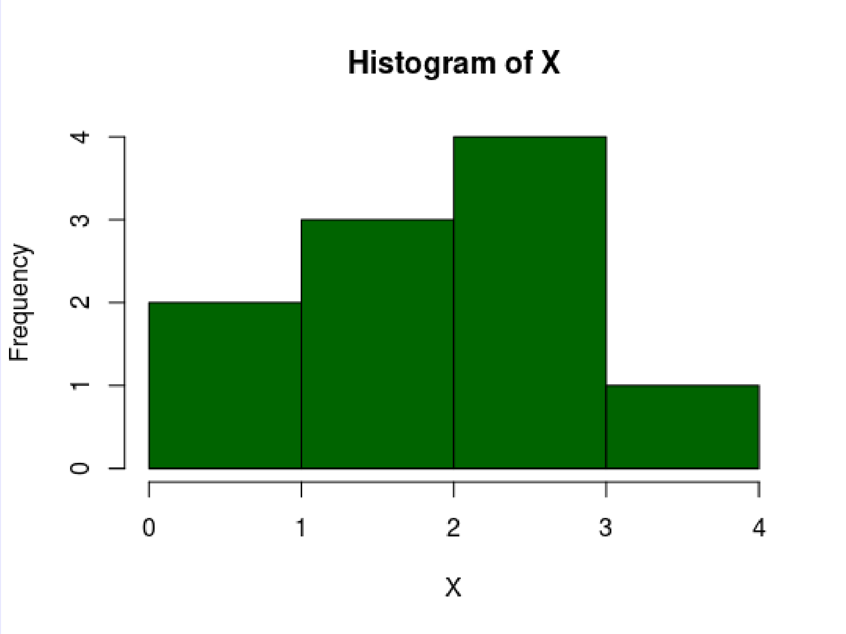
- Mode (sample): mode of a r.v. *X* is the value at which the PMF attains its maximum value.
  - $\text{mode}(X) = \arg\max_x \hat{f}(x)$

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

- May not be a useful measure of central tendency.

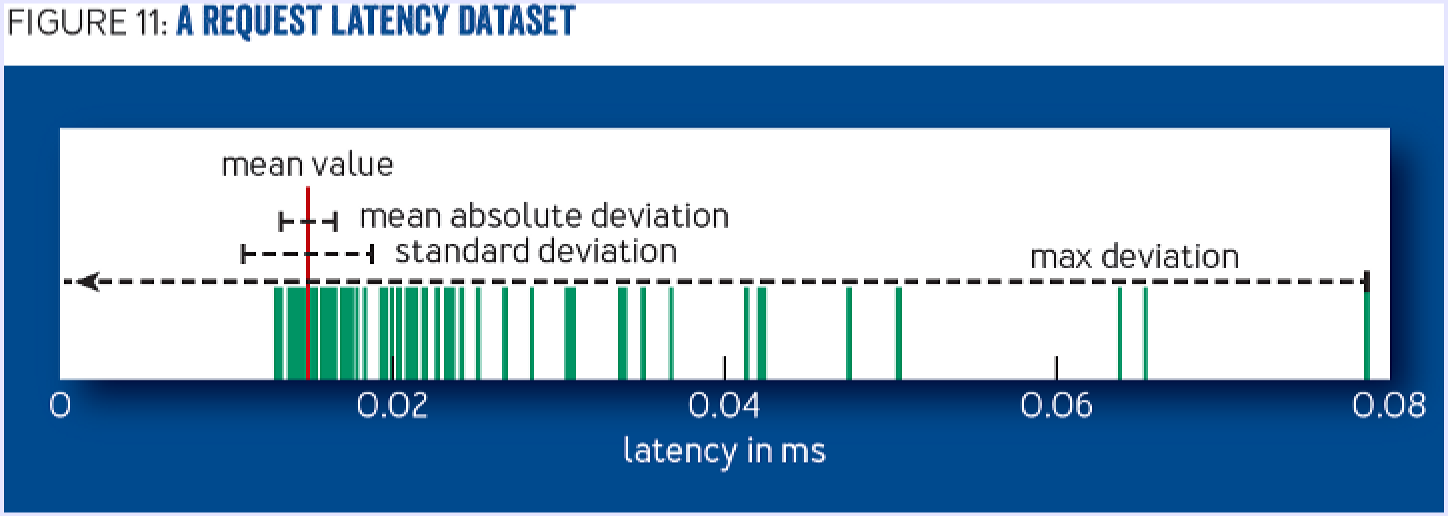


Measures of central tendency

- Measures of dispersion: Variance and standard deviation.
  - Variance: A measure of how much the values of *X* deviate from the expected (mean) value of *X*.
  - Sample variance: var(*X*) = $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$
  - Sample standard deviation is the squared root of the sample variance.
  - Sample standard deviation: $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$

Measures of central tendency

- Standard deviation is not the only game in town.
  - Maximal deviation: $\mathrm{maxdev(X)} = max(|x_i - \mu|)\ \forall x_i \in X$
  - Mean absolute deviation: $\mathrm{mad(X)} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu|\,, \forall x_i \in X$



Source: H. Hartmann, "Statistics for Engineers," CACM 50(7), July 2016.

Measures of central tendency of multivariates

- We will consider bivarate analysis for discussion. (Results generalize to n-D).
- We seek to understand the association or dependence of two attributes X_1 and X_2.

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

- Geometrically, we can view them as vectors in 2-D space (row view), or vectors in an n-D space (column view).

Measures of central tendency of multivariates

- The first and second moments (mean and variance, respectively) are computed in the same manner, except a *vector* is returned.

$$\mathbf{D} = \left(\begin{array}{cc} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{array}\right)$$

- The variance can be computed for each attribute, σ_1^2 for X_1 and σ_2^2 for X_2.
- The total variance is given by:

$$var(D) = \sigma_1^2 + \sigma_2^2$$

Measures of central tendency of multivariates

- Measure of association: Covariance.
- Covariance is the measure of association or linear dependence between two variables, X_1 and X_2. $$cov(X_1, X_2) = \hat{\sigma}_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

Measures of central tendency of multivariates

- Measure of association: Covariance.
- Covariance is the measure of association or linear dependence between two variables, X_1 and X_2. $$cov(X_1, X_2) = \hat{\sigma}_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$
- The variance-covariance information for the two attributes can be summarized by a square (nxn) covariance matrix: $$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$

Measures of central tendency of multivariates

- The covariance matrix is: $\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$
  - Square
  - Symmetrical
  - Attribute specific covariances on main diagonal; covariance between attributes on off-diagonal elements.
  - Total variance of the attributes is the sum of the diagonal elements (sometimes called the **trace** of the Σ).

Measures of central tendency of multivariates

- Example:

$$A = \begin{pmatrix} 3 & 6 & 0 \\ 6 & 12 & 16 \\ 5 & 10 & 59 \end{pmatrix}$$

- Cov(A) = $\begin{pmatrix} 2.3 & 4.6 & 20.5 \\ 4.6 & 9.3 & 41.0 \\ 20.5 & 41.0 & 931.0 \end{pmatrix}$

Measures of central tendency of multivariates

- Related to covariance is correlation.
- Correlation between two variables, X_1 and X_2 is the *standardized covariance* obtained by normalizing the covariance with the std. dev. of each variable:

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$$

Measures of central tendency of multivariates

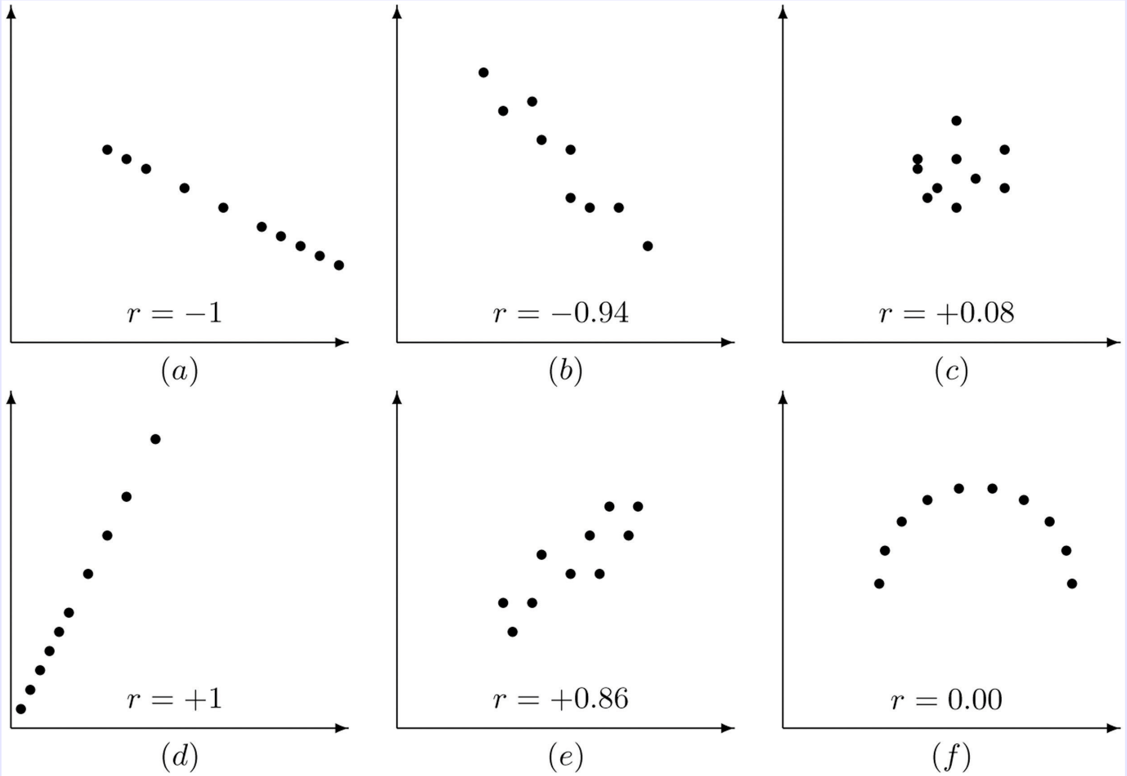
- Example:

$$A = \begin{pmatrix} 3 & 6 & 0 \\ 6 & 12 & 16 \\ 5 & 10 & 59 \end{pmatrix}$$

- Cor(A) = $\begin{pmatrix} 1.00 & 1.00 & 0.44 \\ 1.00 & 1.00 & 0.44 \\ 0.44 & 0.44 & 1.00 \end{pmatrix}$

Measures of central tendency of multivariates

- Why have both covariance and correlation?
  - The range of covariance: [-∞, ∞]
  - The range of correlation: [-1, 1]
  - Correlation is dimensionless.
- And remember: Correlation != Causation!!



Code: cars.r

Measures of central tendency of multivariates

- Correlation and collinearity
  - Correlation measures the relationship between two variables.
  - Collinearity occurs when the two variables are so highly correlated that we can use one to predict another; i.e., one variable is a linear combination of the other variable.
    - Multicollinearity occurs when > 2 predictor variables are are inter-correlated.