# CEM1002

Neil Montgomery

2015-09-28

# Today

## Graphical/numerical summary wrap-up

- (Side-by-side) boxplots
  - Plots five numbers e.g.
  - min, 1st quartile, median, 3rd quartile, max
  - no standard definition
  - some implementations will try to suggest "outliers"
- (Observed) (sample) correlation (coefficient)

## Essentials of probability

- Random variable, distribution, etc.
- The Normal distributions
- Normal quantile plot

# Correlation coefficient

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

# Random variables and distributions

Some analogies/correspondences:

| Theoretical Model | Observed Data |
|---|---|
| Random variable $X$ | Observation |
| Distribution | ??? |
| Sample $X_1, X_2, \ldots, X_n$ | Dataset $x_1, x_2, \ldots, x_n$ |
| Probability Density Function | Histogram/density plot |
| Expected value $E(X) = \int x f_X(x)\, dx$ | ??? |
| Sample average $\overline{X} = \frac{\sum X_i}{n}$ | Obs. sample average $\overline{x} = \frac{\sum x_i}{n}$ |
| Variance $E((X - E(X))^2)$ | ??? |
| Sample Variance $\frac{\sum (X_i - \overline{X})^2}{n-1}$ | Obs. sample variance $\frac{\sum (x_i - \overline{x})^2}{n-1}$ |