GENE:6234 Intro

Class Overview

Here are some of the goals of our class this semester:

- Familiarize ourselves with basic concepts in statistics
- Be able to read and interpret data
- Understand the specific goals of statistical analysis
- Get context in which various statistical tools are used, particularly in genetics
- Have fun

Goals for Today

We are going to start with a review of common (little s) statistics

- What are statistics?
- What is data reduction?
- Important data reductions:
    - Measures of centrality
    - Measures of dispersion
    - Measures of association

What is Statistics

(Big S) Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

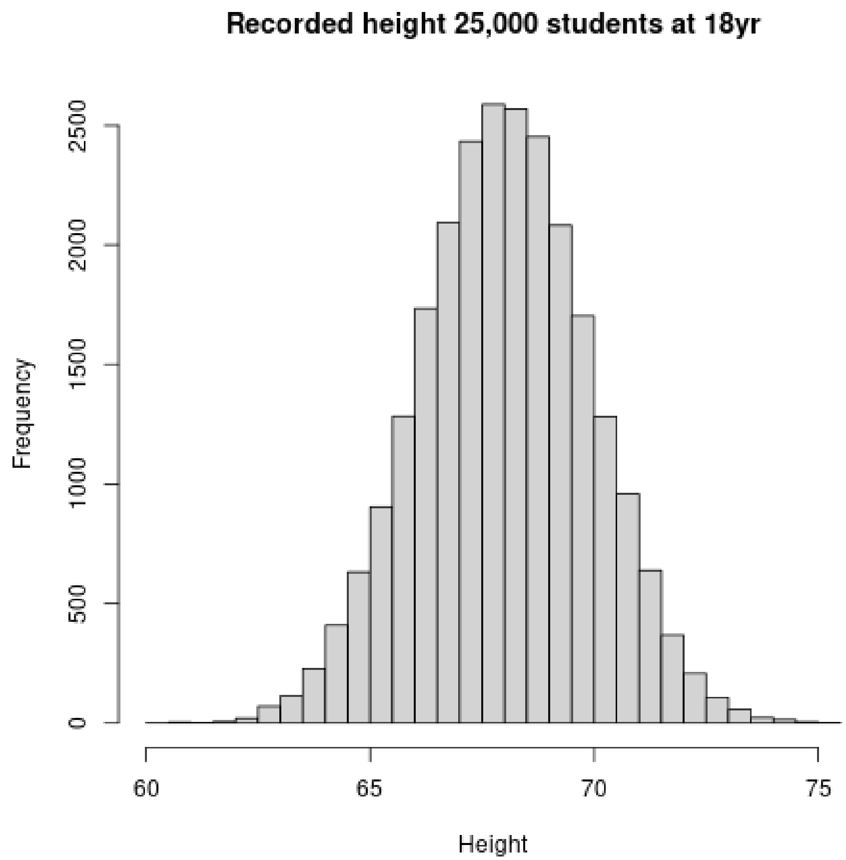
- Study design
- Analysis
- Presentation

Types of data

Type of Observation	Distinguishing Characteristics	Examples
Discrete	**Observations in discrete classes**	
A. Nominal	Distinct classes do not have any natural order or ranking	Sex, treatment group, presence or absence of disease
B. Ordinal	Distinct classes have a predetermined or natural ordering	Classificatino of disease by severity, scales of degree for agreement, plaque index
Continuous	**Observations assume any value on continuous scale**	
A. Interval	Scale is defined in terms of differences between observations; *zero point is arbitrary*	Temperature in degrees, IQ measurements
B. Ratio	Scale differences represent real realtionships in the items measured; *zero point represents total absence of the attribute being measured*	Height, weight, income, cytokine levels

Data Reduction

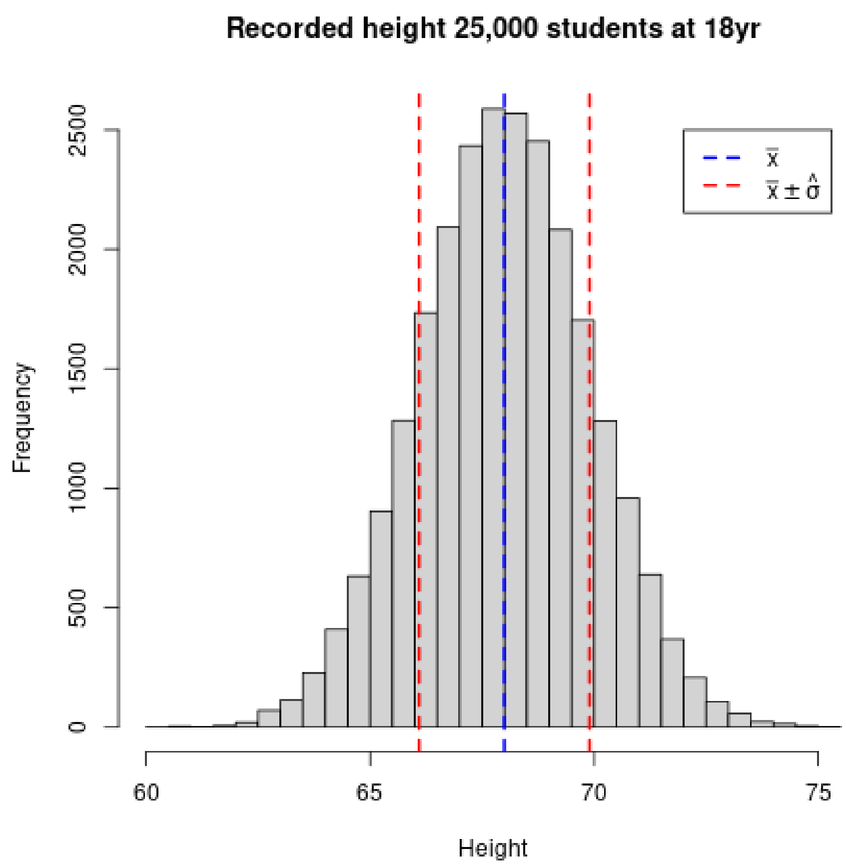
Consider SOCR data on the height of 25,000 children recruited at 18 years of age



- visual representation
- measures of centrality
- measures of dispersion
- measure of association

Data Reduction cont.

(Little s) statistics are the outcomes of a numeric reduction to a dataset



- Mean value is $\bar{x} = 67.99$
- Standard deviation is $\hat{\sigma} = 1.9017$
- $\bar{x} \pm \hat{\sigma}$ is (66.091, 69.895)
- 68, 95, 99.7 rule
- This range includes 17,089 individuals, or 68.36%

A note on X vs x

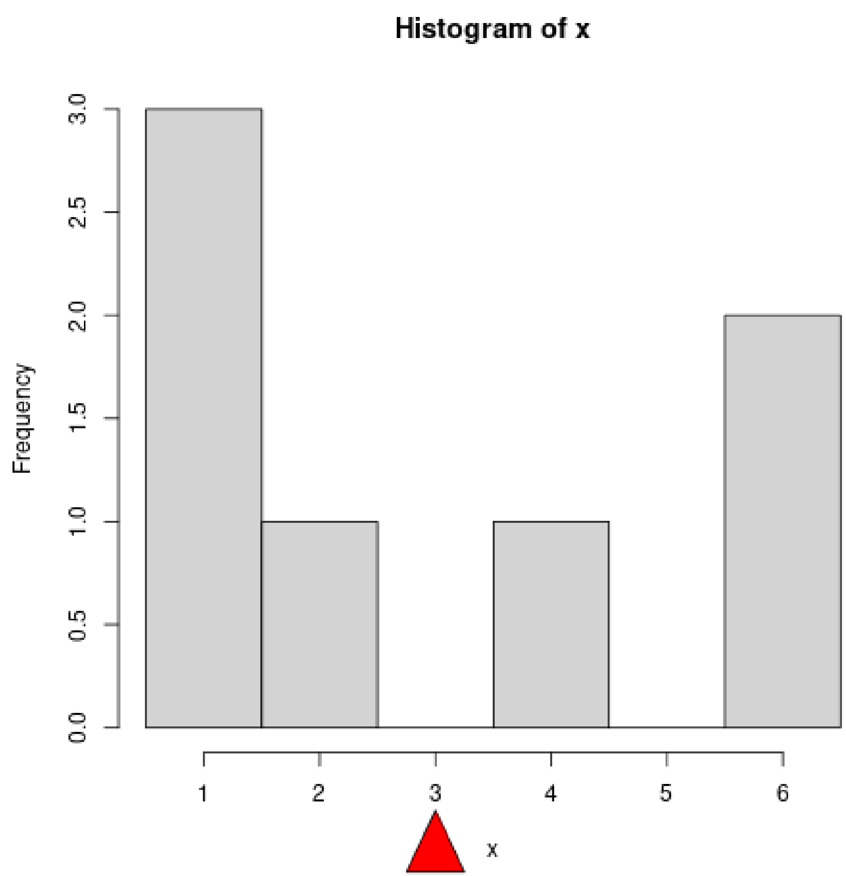
Throughout this course (and in Statistics in general), we differentiate between a hypothetical random variable with captial X, and a specific, realized observation of a sample with lowercase x

- $\overline{X}$ may represent the average height of students in any statistics course at UI, though we have yet to measure it
- $\bar{x}$ represents the average of *this specific class* which we have defined and measured

Measure of Centrality

Mean

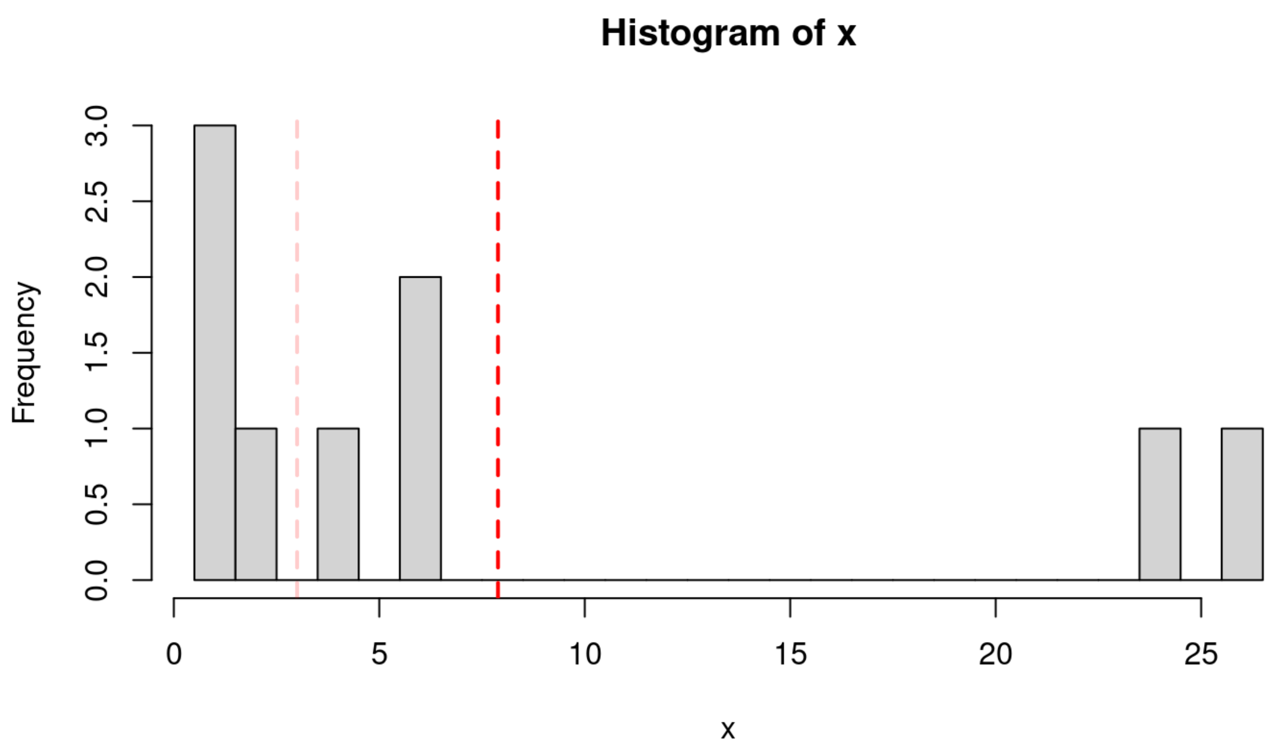
$x = \{1, 1, 1, 2, 4, 6, 6\}$ (in mg)



- Arithmetic mean, average
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\bar{x} = 3$ mg
- Imagine as fulcrum
- “center of mass”

Mean Continued

Mean can be highly sensitive to outliers



Median

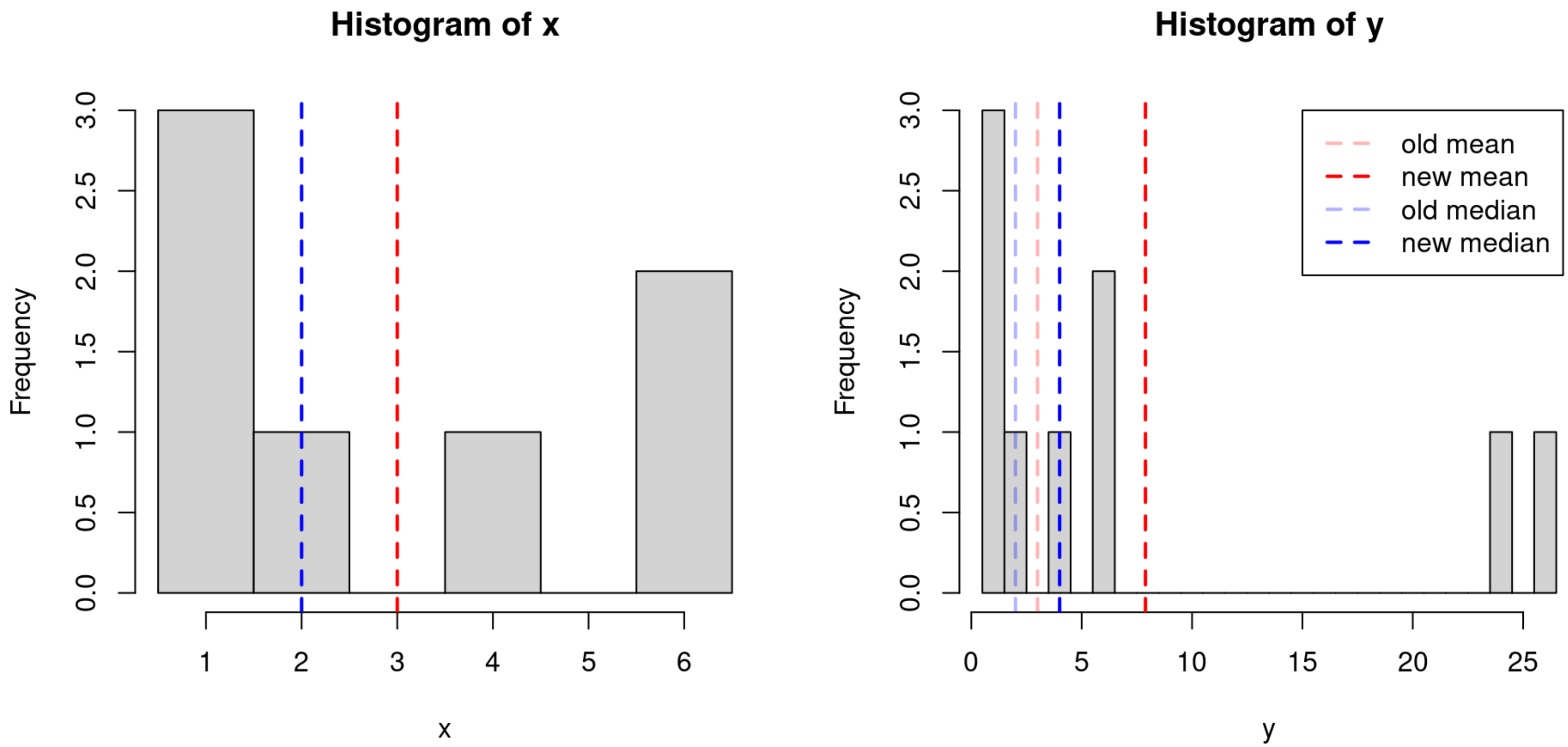
The median is taken to be the center value of the observed data, ranked from smallest to largest. In the event that n is even, the average of the center two observations is used.

$$X_{odd} = \{1, 1, 1, \textcolor{red}{2}, 4, 6, 6\}, \qquad \text{median} = 2$$

$$X_{even} = \{1, 1, 1, \textcolor{red}{2}, \textcolor{red}{3}, 4, 6, 6\}, \quad \text{median} = \frac{2+3}{2} = 2.5$$

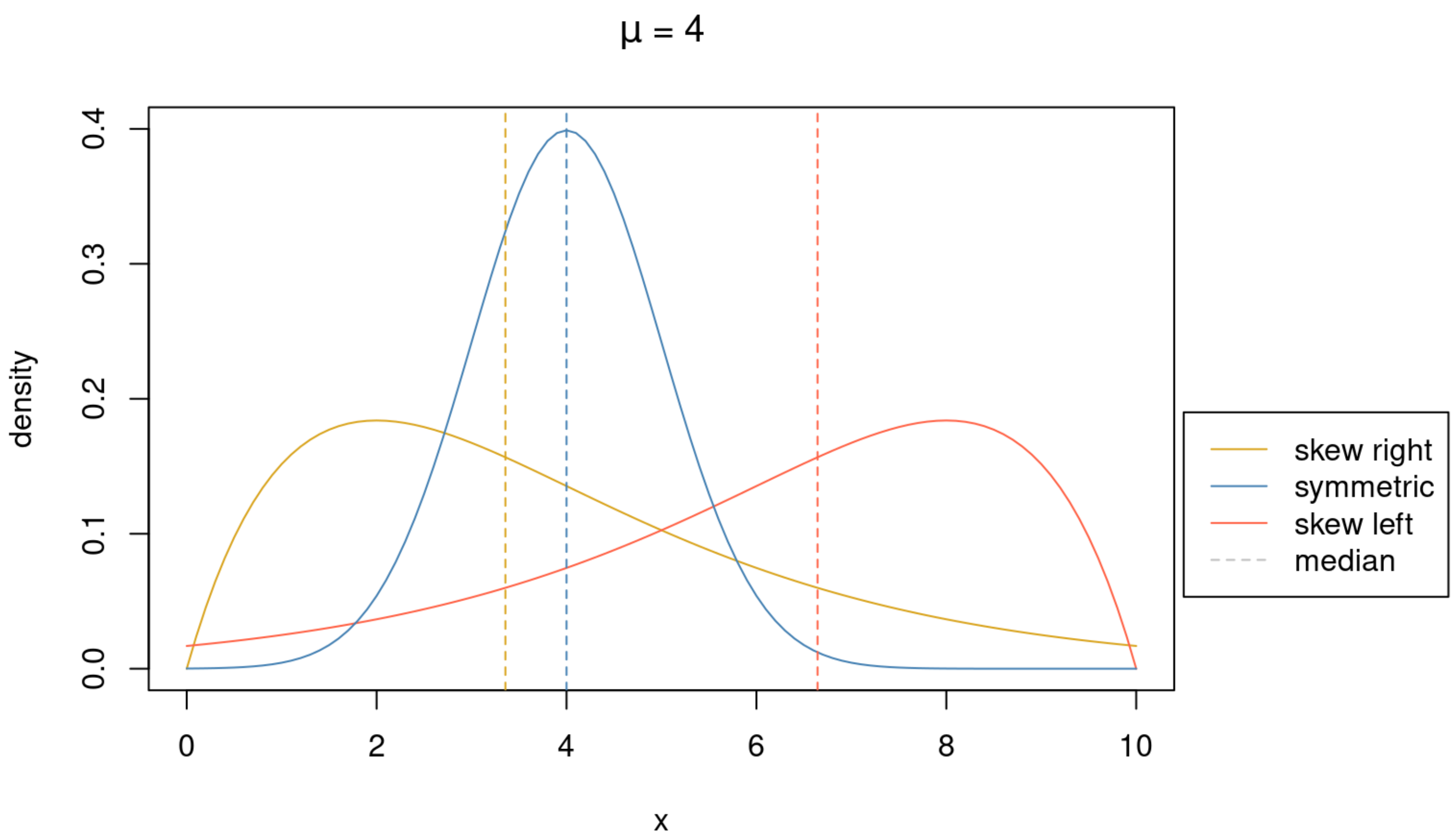
Median continued

Unlike the mean, the median is more robust to outliers.



Skewness

Each of these curves have the same mean

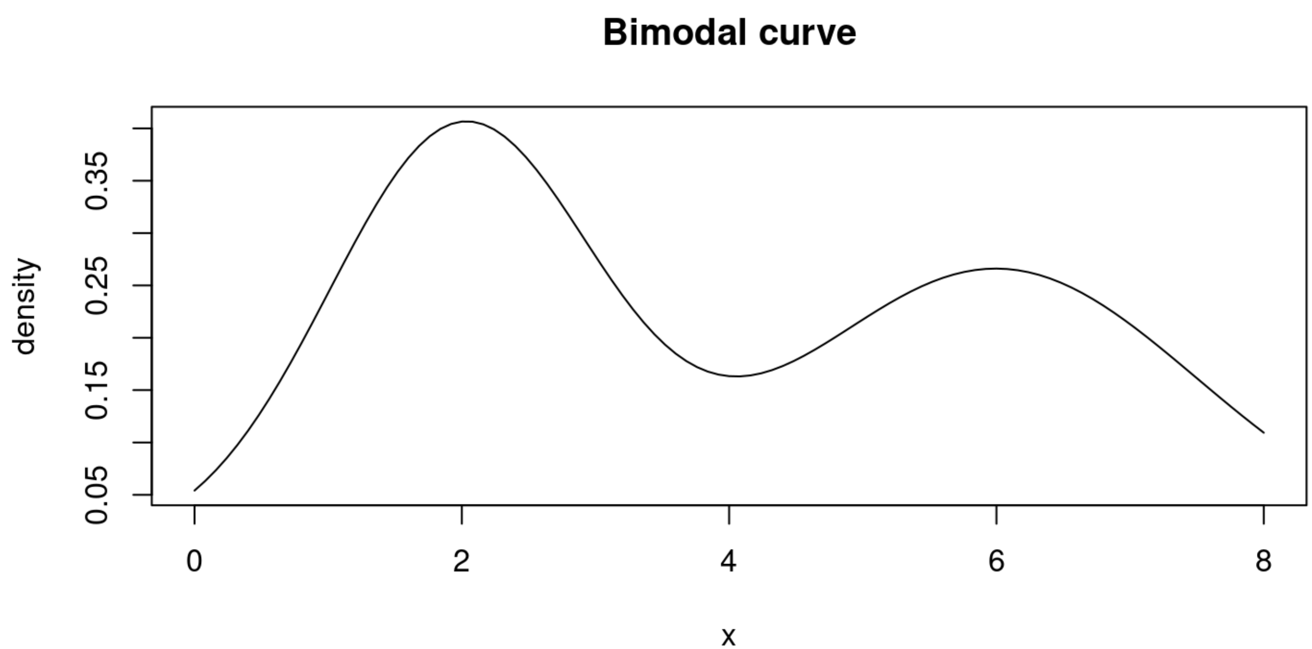


Mode

The mode is determined by the value that occurs most frequently

$$x = \{1, 1, 1, 2, 4, 6, 6\}$$

More frequently, we use it to describe a value whose frequency is larger than the values of either side of it. For continuous data, this looks like a “hump”



Mode

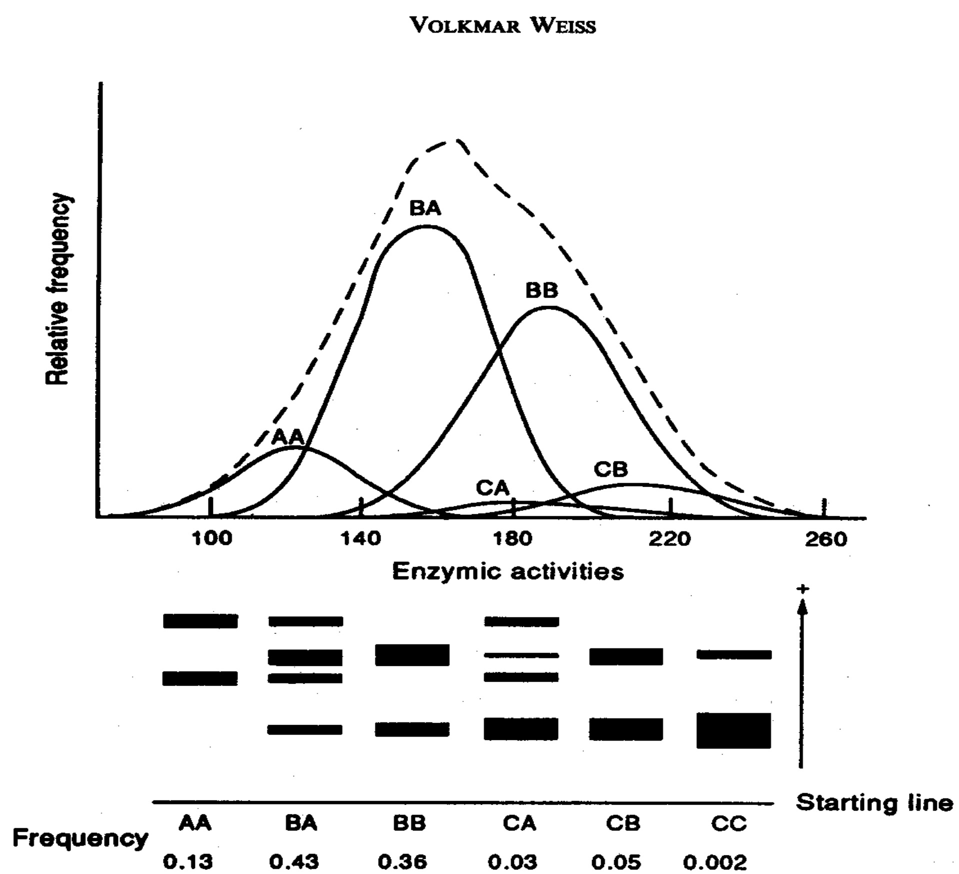


Fig. 2. Genotypes separated qualitatively by electrophoresis (below) and corresponding quantitative distribution (above) of their enzymic activities (human red cell acid phosphatase from Harris, 1966).

From: http://www.v-weiss.de/majgenes-full.html

Original illustration from Harris H. (1966) Enzyme polymorphisms in man. Proc. Roy. Soc. B 164, 298-310.

Measures of Dispersion

Variance

The variance is defined as a (kind of) average of squared deviations from the mean

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

For our sample $x = \{1, 1, 1, 2, 4, 6, 6\}$, where we found $\bar{x} = 3$ mg, we have

$$s^2 = \frac{1}{7-1}\sum_{i=1}^{n}(x_i - 3)^2 = \frac{32}{6} = 5.33 \text{ mg}^2$$

Standard Deviation and Coef of Variance

Standard Deviation

The standard deviation is simply the square root of the variance, $s = \sqrt{s^2}$. For our sample, this gives

$$s = \sqrt{5.33 \text{ mg}^2} = 2.31 \text{ mg}.$$

Coefficient of Variation

$$C.V = \frac{s}{\overline{x}}$$

Percentiles and IQR

We will frequently consider the percentiles of a sample. For any whole number r between 1 and 99, the rth percentile, $X_{\{r\}}$ for a sample is value for which at most r percent of observations are less than $X_{\{r\}}$ and at most $(100 - r)$ percent are larger than $X_{\{r\}}$. Some common percentiles include:

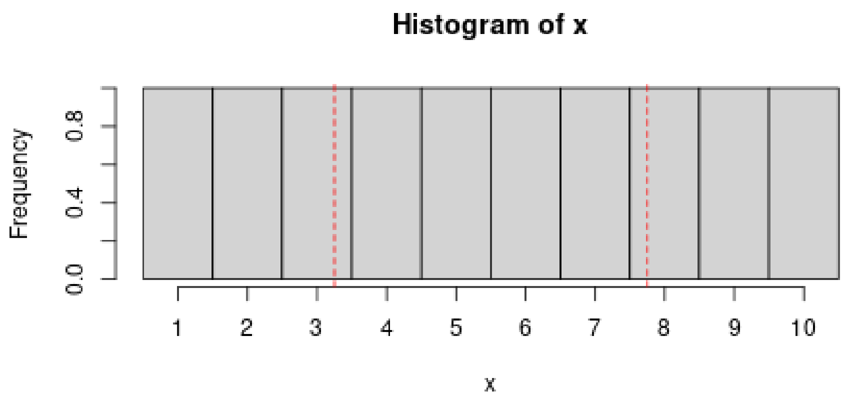
- Median, $X_{\{50\}}$
- 1st or lower quartile - $X_{\{25\}}$
- 3rd or upper quartile - $X_{\{75\}}$

These last two values are used to compute the *interquartile range*, which gives upper and lower bounds for the middle 50% of the data.

IQR Cont.

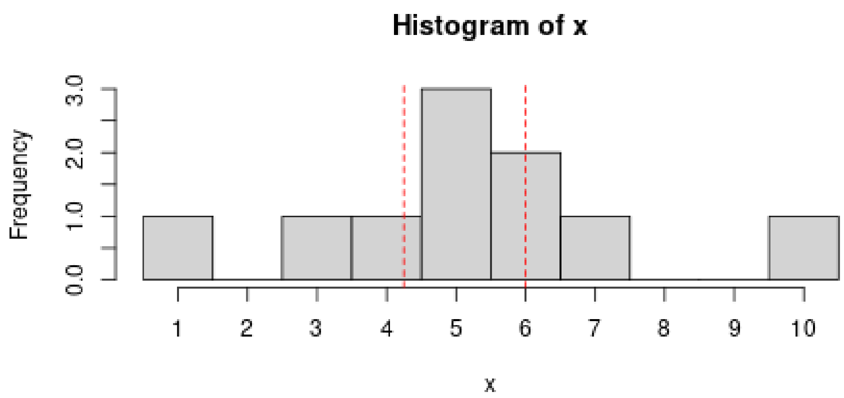
$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

- $x_{\{25\}} = 3.25$
- $x_{\{75\}} = 7.75$

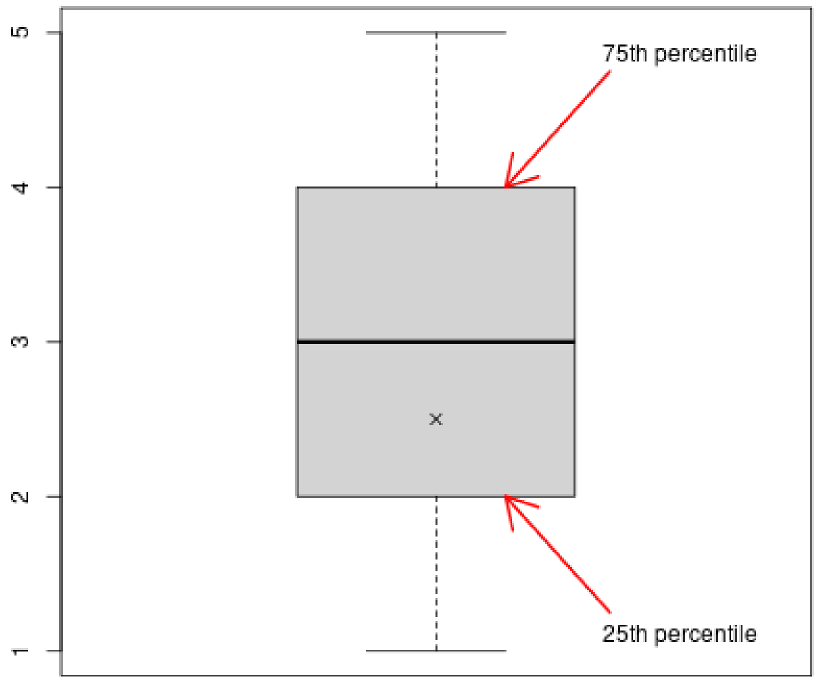


$x = \{1, 3, 4, 5, 5, 5, 6, 6, 7, 10\}$

- $x_{\{25\}} = 4.25$
- $x_{\{75\}} = 6$

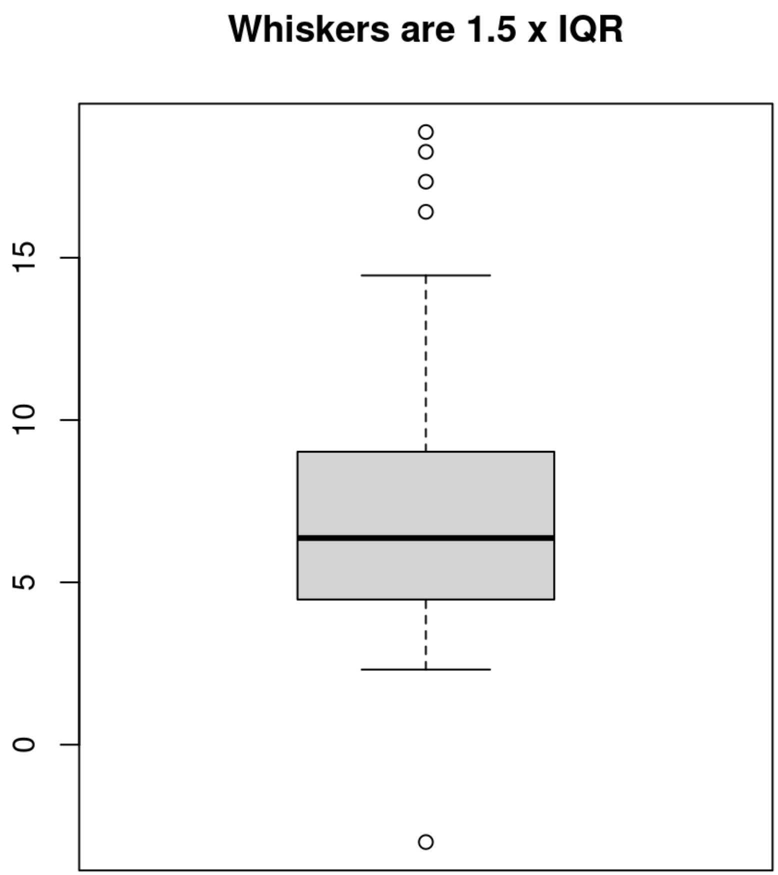


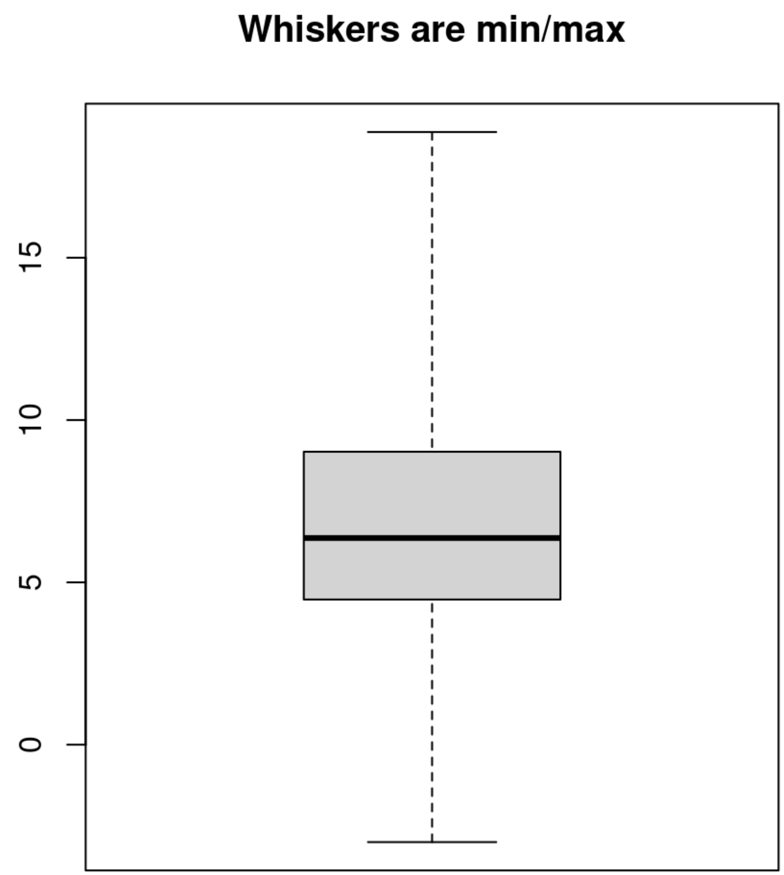
Box Plot



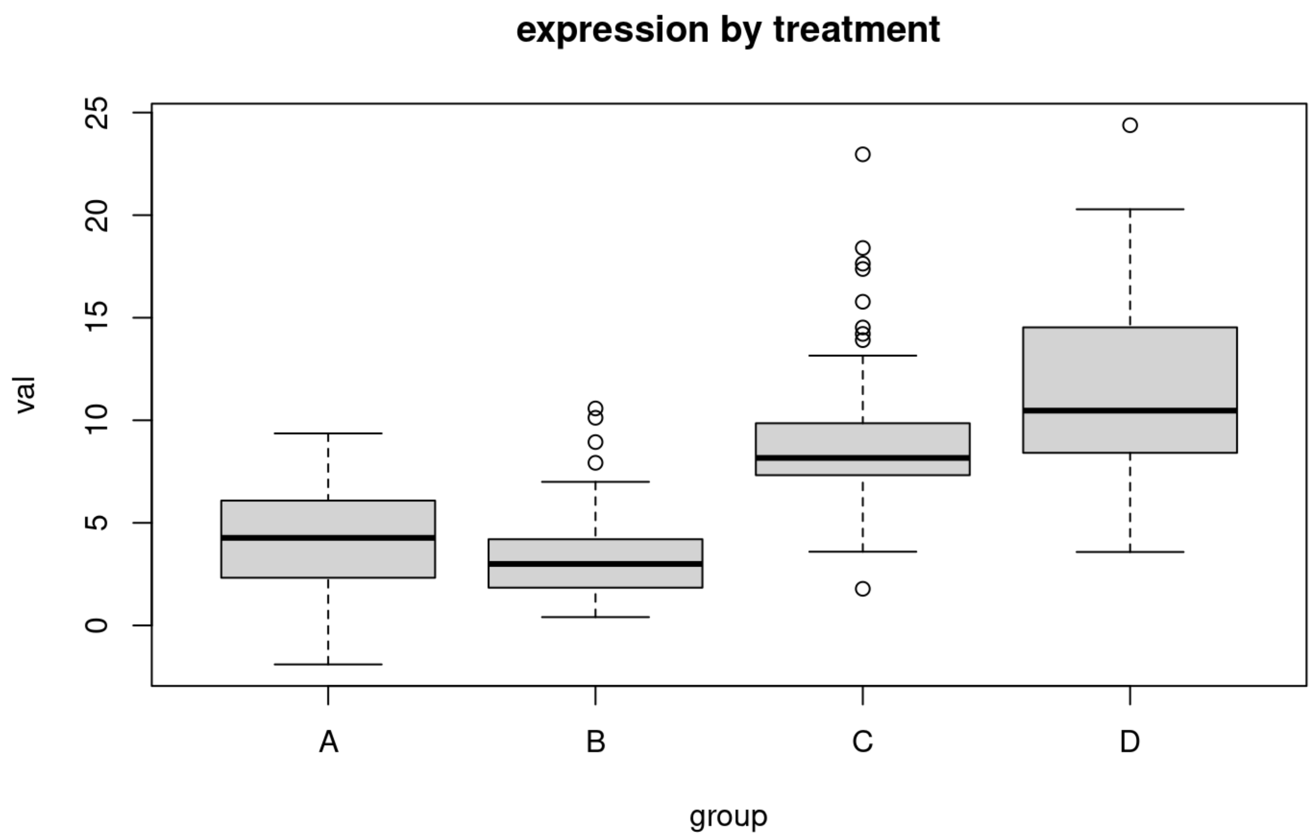
- Center line is median
- Gray box is IQR
- Mean indicated with $\times$ or $*$
- Five-number summary
  - minimum
  - lower quartile
  - median
  - upper quartile
  - maximum

Box Plot



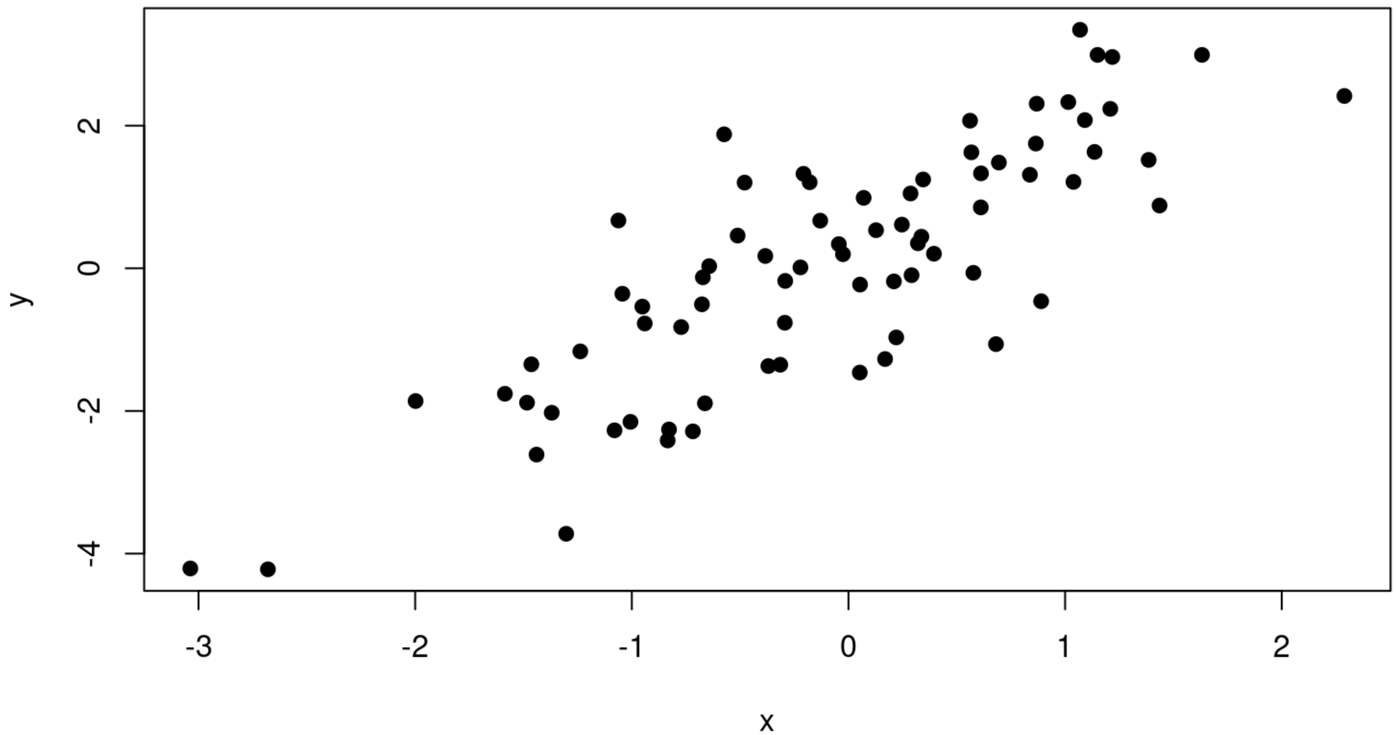


Box plots



Measures of Association

Scatterplot

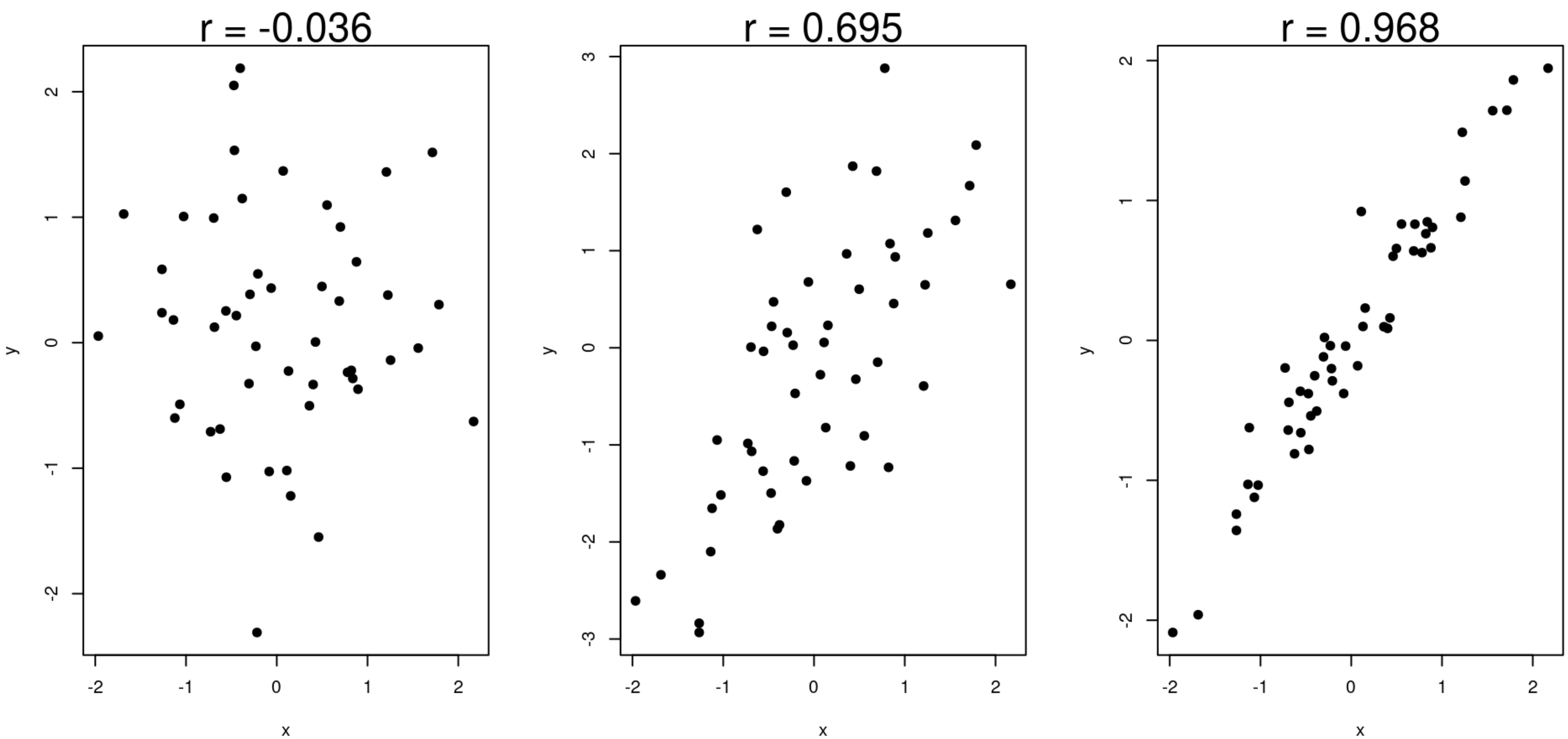


Correlation

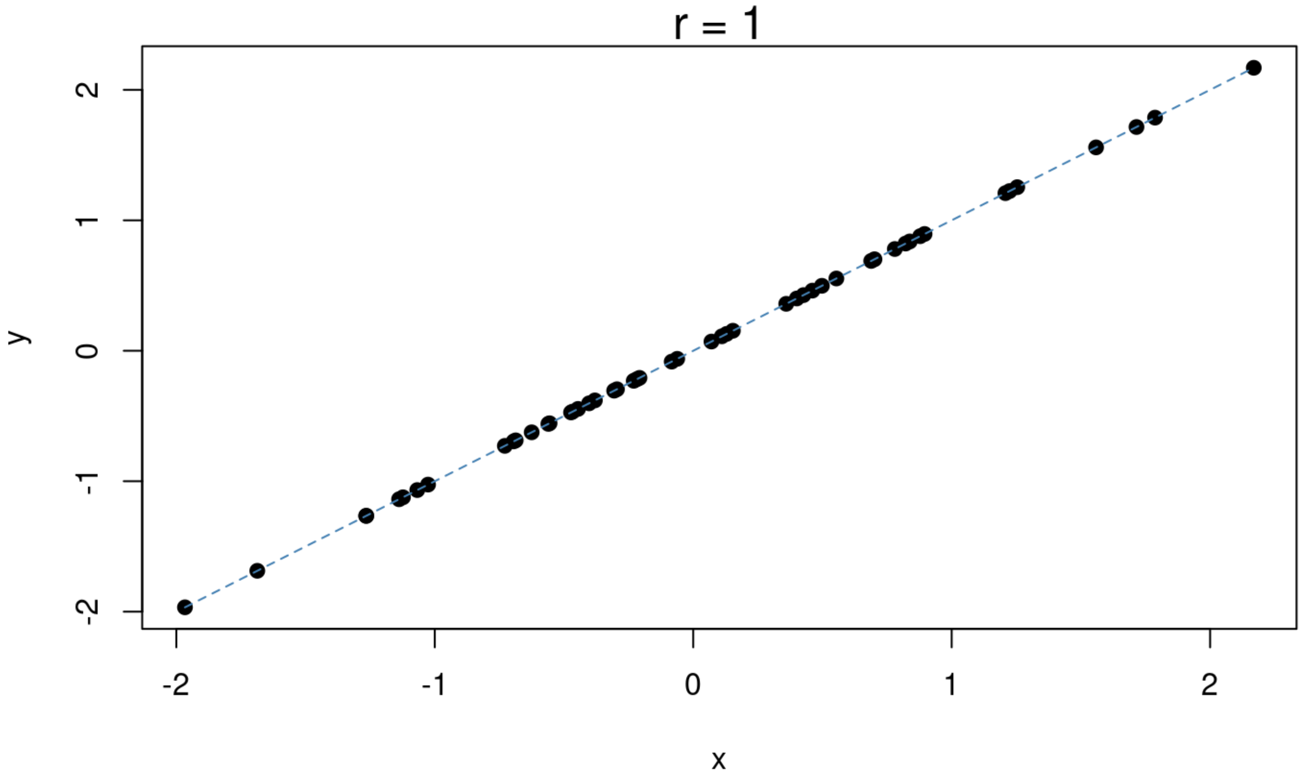
Pearson correlation r measures the **linear** association between two variables, (X, Y)

- Unitless measure
- $-1 \leq r \leq 1$
- $r = 0$ indicates no linear association

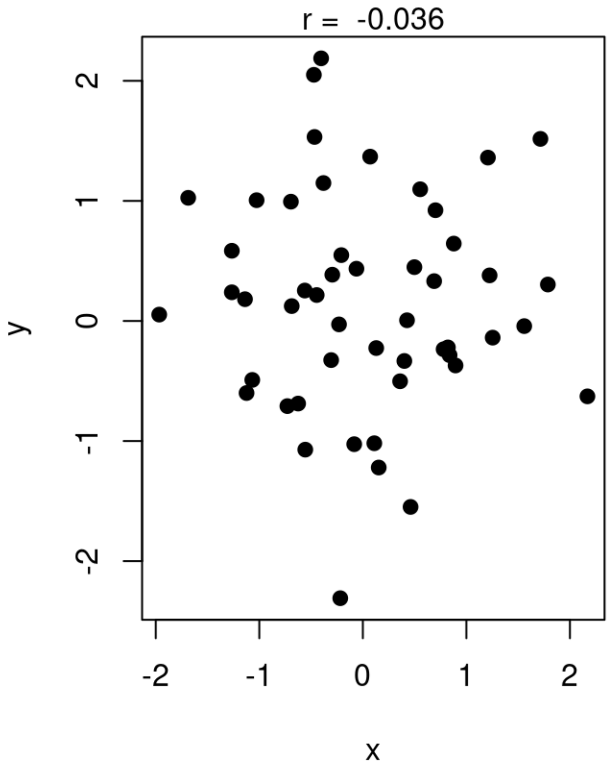
Pearson Correlation

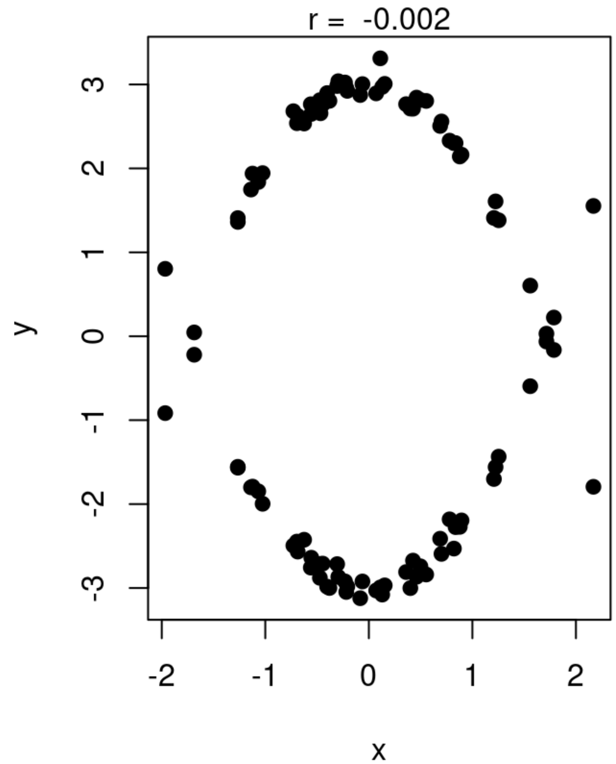


Pearson Correlation



Pearson Correlation





Rank Correlation

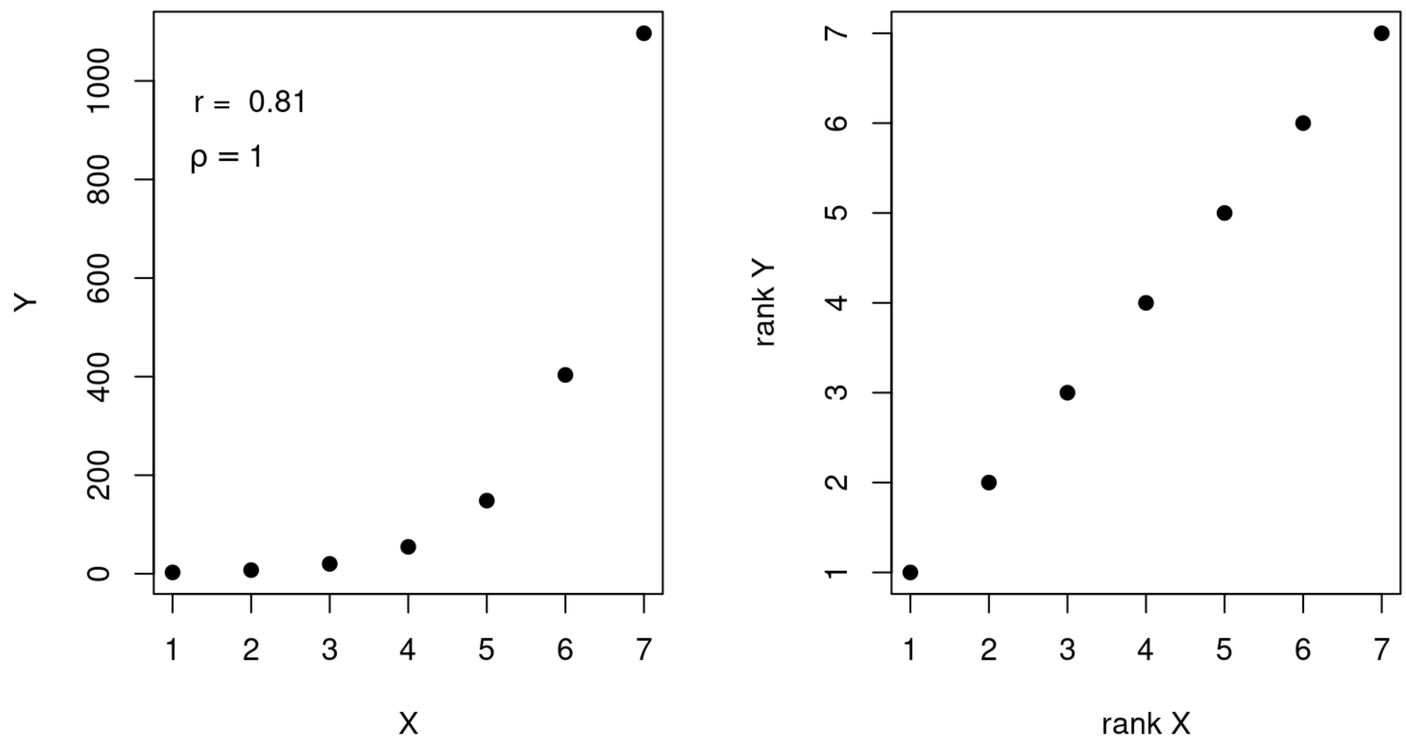
In addition to Pearson, we have *Spearman's* ρ *coefficient*, where the values of X and Y are replaced with their rank ordering values

$$\begin{aligned} X &= \{2, 4, 6, 10, 8\} \\ Y &= \{7, 4, 1, 5, 3\} \end{aligned} \implies \begin{aligned} X_{rank} &= \{1, 2, 3, 5, 4\} \\ Y_{rank} &= \{5, 3, 1, 4, 2\} \end{aligned}$$

Where Pearson's r measures the *linear* association, Spearman's ρ measures the *monotonic* association

Rank Correlation

$$y = e^x$$



Review

- Measures of centrality
  - mean, center of mass
  - median, middle observation
  - mode, humps in curve
- Measures of dispersion
  - variance, average squared error
  - standard deviation
  - interquartile range
- Measures of association
  - pearson's r, linear association
  - spearman's ρ, monotonic association