# STAT 270- Chapter 2

May 13, 2012

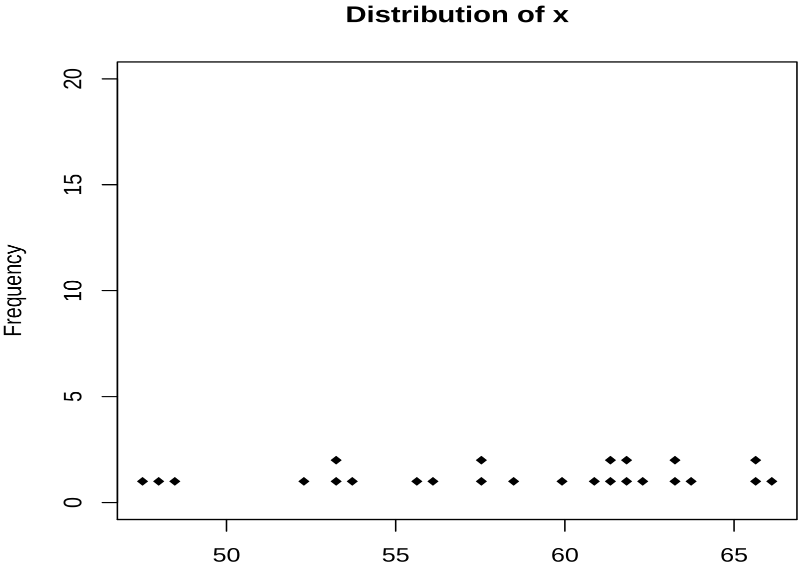
# Descriptive Statistics

- Numerical
- Graphical

**Dotplots** $\rightarrow$ Graphical
Used for univariate data, i.e., single measurments on subjects.

$$x_1, x_2, \ldots, x_n$$

# Dotplots



# Dotplots

Characteristics that can be detected from dotplots:

- Outliers (extreme observations)
- Centrality (concentration of data in the middle portion)
- Dispersion (Spread of data along the axis)

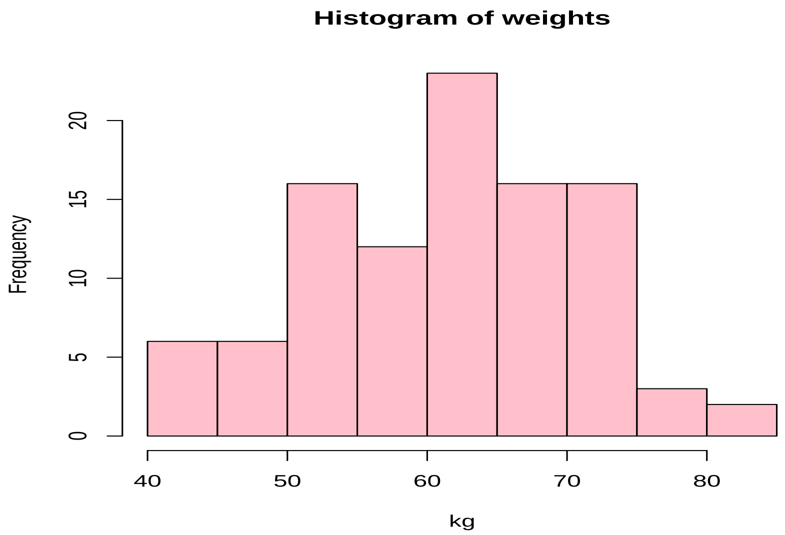
Not useful for large data sets; not very common in general!

# Histograms $\rightarrow$ graphical

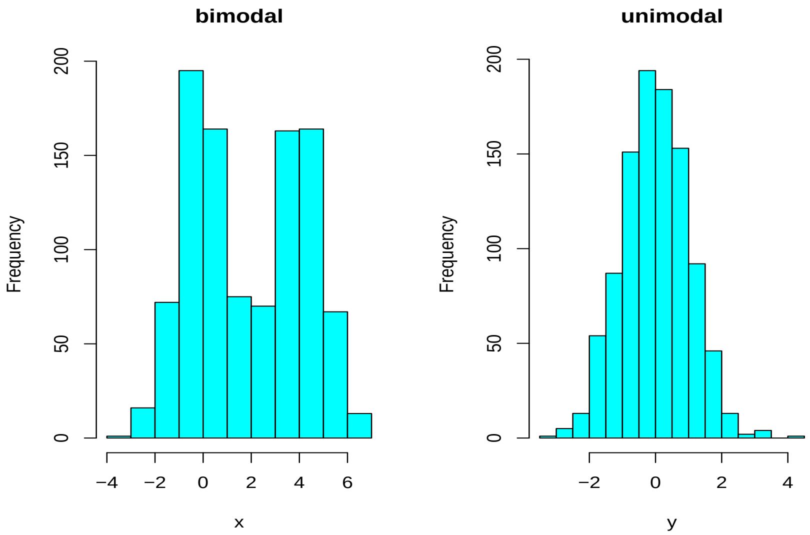
Used to describe univariate data
Constructed by statistical software for large data sets
Characteristics that can be detected form the histogram:

- Outliers
- Centrality
- Dispersion
- Modality
- Skewness and symmetry

# Histogram

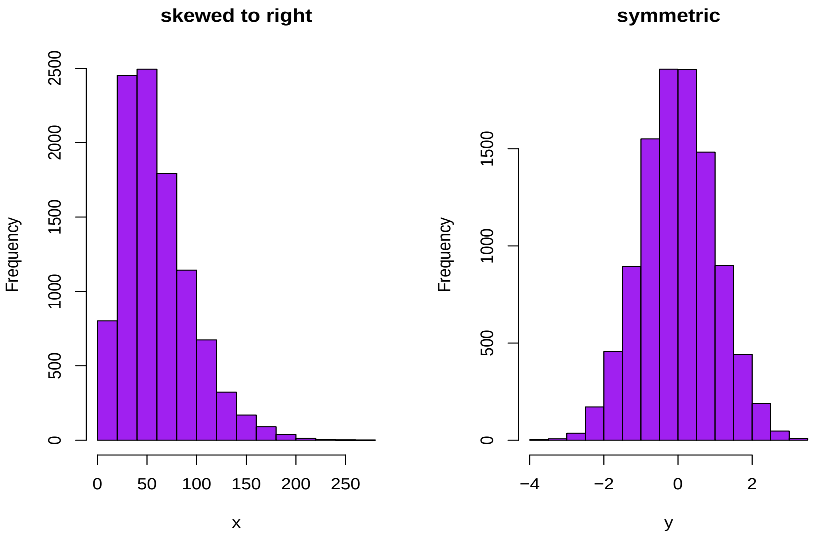


# Modality



bimodal
Frequency
0
50
100
150
200
−4
−2
0
2
4
6
x
unimodal
Frequency
0
50
100
150
200
−2
0
2
4
y

# Skewness and Symmetry



# Histograms

How to construct histograms:

- Construct consecutive intervals of equal size
- Calculate frequencies and relative frequencies
- Label the axes and provide a title

Example: Weight data,

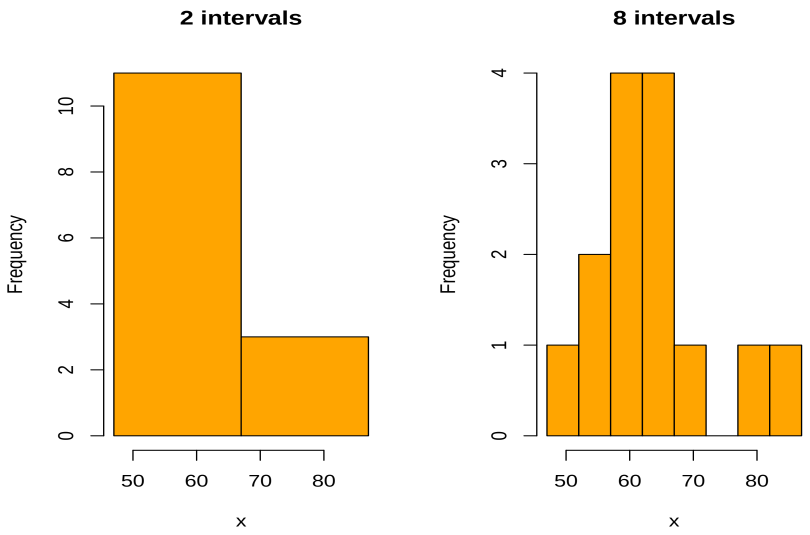
$$47, 55, 79, 63, 64, 67, 54, 59, 58, 84, 70, 61, 65, 59$$

How many intervals?
Extreme cases:

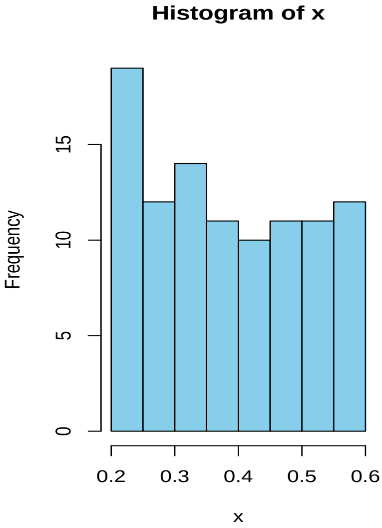
- A few long intervals $\rightarrow$ too much summarization
- Many short intervals $\rightarrow$ not enough summarization

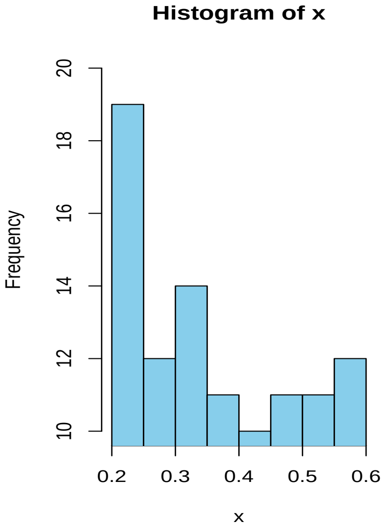
A rule of thumb: number of intervals$= \sqrt{n}$
$n$: sample size

# Number of Intervals

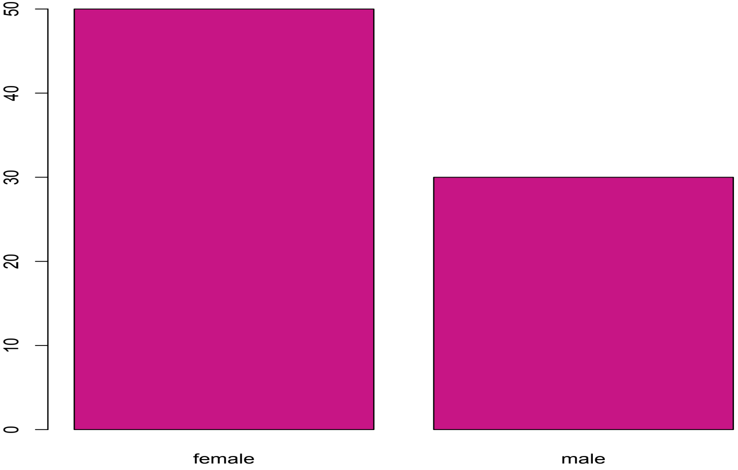


# Vertical axis scale in a histogram



# Histogram for categorical data - Barplot/Bar graph



# Unequal intervals

Intervals of equal size are recommended.
For reasonable visual understanding when unequal intervals are used:
vertical axis:

$$\frac{\text{relative frequency}}{\text{length of interval}}$$

# Measures of location

Describe centrality of univariate data

$$x_1, x_2, \ldots, x_n$$

**Sample mean** $\rightarrow$ Numerical

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

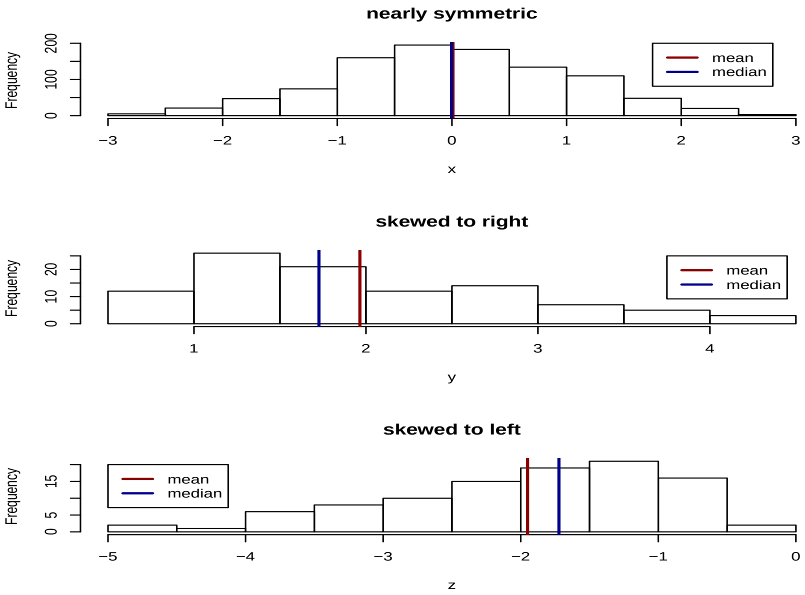
**Sample median** $\rightarrow$ Numerical
sorted data

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})/2 & \text{if } n \text{ is even} \end{cases}$$

Sample mean is more sensitive to outliers that the sample median.

# Comparability of mean and median



# Measures of variability (dispersion)

Why does variability matter?

- Range

$$R = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$$

Depends only on two data values $\rightarrow$ inefficient (like median)

- Variance

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

# Variance

- $s^2 \geq 0$
- $s^2 = 0$ when $x_1 = x_2 = \ldots = x_n$ (no variability)
- Denominator is $(n-1)$ NOT $n$
- $s^2$ is in squared units
- More convinient formula for $s^2$:

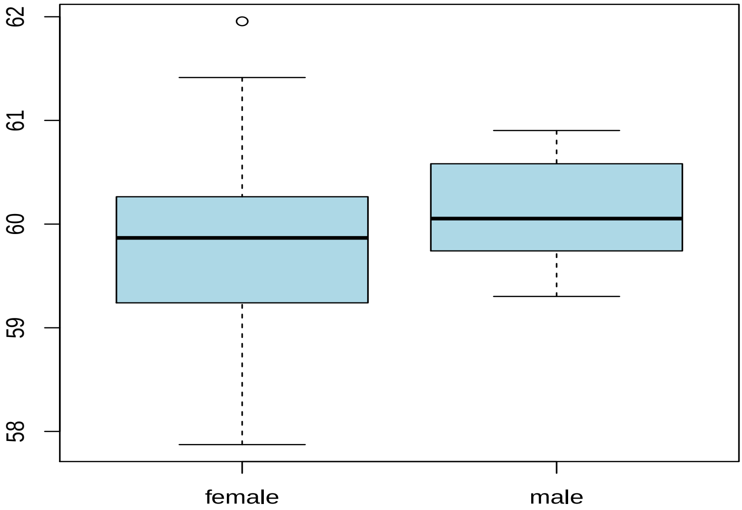
# Standard deviation (sd)

$s = \sqrt{s^2} \rightarrow$ same units as the data
3-sigma rule: roughly 99% of the data falls into $(\bar{x} - 3s, \bar{x} + 3s)$

# Boxplots → graphical

- Useful for **grouped** univariate data
- Constructed by statistical softwares; we will focus on interpretation
- Variation and skewness of the data can be detected from boxplots

# Boxplots



# Paired data (bivariate data)

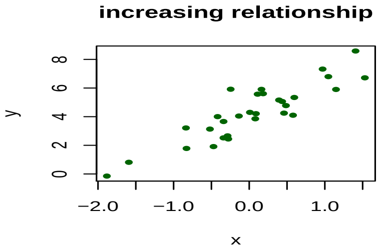
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

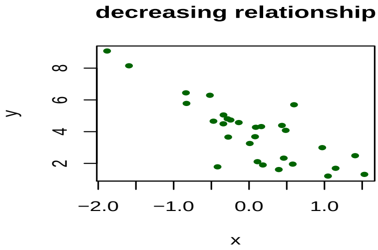
We want to study the relationship between $x$ and $y$ that can be

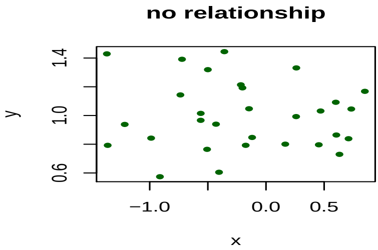
- No relationship
- Association
- Causal

# Scatterplots

First thing to look at to detect a relationship between two variables



# Prediction

- model the relationship between two variables $x$ and $y$
  $\rightarrow$ predict $y$ at a new point $x = x^*$
- Be causious of **extrapolation**

# Correlation Coefficient (correlation/sample correlation)

is used to study paired data,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $r$ is dimensionless
- $-1 \leq r \leq 1$
- $r \approx 1 \rightarrow$ strong positive correlation
- $r \approx -1 \rightarrow$ strong negative correlation
- $r \approx 0$ does NOT imply no relationship, it implies no **linear** relationship

# correlation coefficient

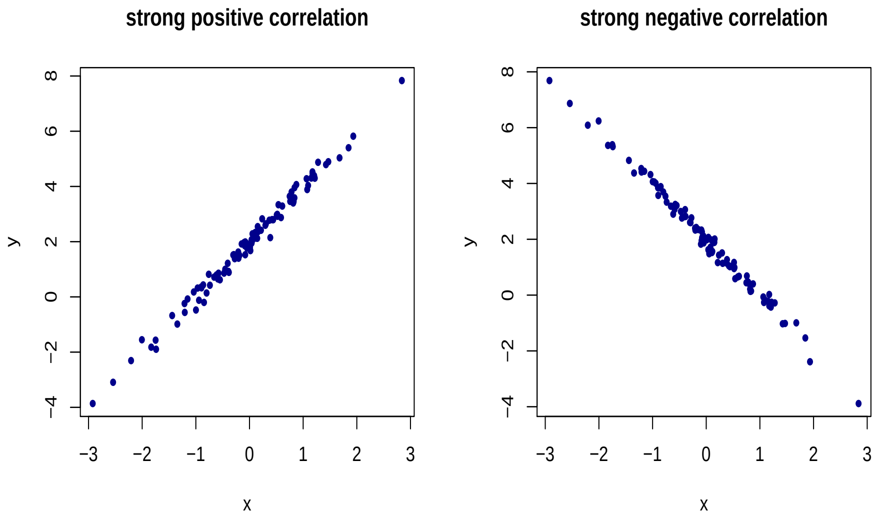
- $r$ measures the degree of linear association,
  If $y_i = a + bx_i$ (exact linear relationship) then

$$r = \begin{cases} 1 & \text{if } b > 0 \\ -1 & \text{if } b < 0 \end{cases}$$

- The intuition behind the sign of $r$:
- Easier to calculate formula:

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

# correction



strong positive correlation
strong negative correlation
y
x
−4
−2
0
2
4
6
8
−3
−2
−1
0
1
2
3