Data Analysis Basics

Prof. Dr. Renato Pajarola
Visualization and MultiMedia Lab
Department of Informatics
University of Zürich

Copyrights

- Most figures of these slides are copyright protected and come from the various indicated sources
- You understand that the slides contain copyright protected material and therefore the following conditions of use apply:
  - The slides may be used for personal teaching purposes only
  - Publishing the slides to any public web site is not allowed
  - Sharing the slides with other persons or institutions is prohibited

Overview

1. Review of variable types
2. Univariate data analysis
3. Distance metrics
4. Dissimilarity and similarity measures

Review of Variables Types

- Continuous variables
  - Positive or negative numeric values
    - age in years, weight, wind speed, temperature, concentrations of pollutants and other measurements
  - Always of ordinal character
- Categorical variables
  - Information that can be sorted into categories
  - Categorical variables can be
    - ordinal
    - nominal
    - dichotomous (binary)

Ordinal Variables

- A variable with some intrinsic order or numeric value
  - Can be categorical if some canonic order can be given
- Examples of ordinal variables:
  - Education
    - no high school degree, high school degree, some university education, university degree
  - Agreement
    - strongly disagree, disagree, neutral, agree, strongly agree
  - Rating
    - excellent, good, fair, poor
  - Frequency
    - always, often, sometimes, never
  - Age group
    - ≤10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, >70

Nominal Variables

- Categorical variable without an intrinsic order
- Examples of nominal variables:
  - Where a person lives (Switzerland, Italy, France, Germany, etc.)
  - Gender (male, female)
  - Employment sector (Industry, Finance, ICT, Agriculture, etc.)
  - Favorite pet (dog, cat, horse, fish, etc.)

Dichotomous Variables

- Categorical (binary) variable with only 2 levels of categories
  - Often represents the answer to a yes or no question
- For example:
  - “Did you attend the information event on May 24?”
  - “Did you eat potato salad at the dinner?”
  - Anything with only 2 categories

Data Cleaning

- One of the first steps in analyzing data is to “clean” it of any obvious data entry errors:
  - Outliers? (e.g. really high or low numbers)
    - Example: Age = 110 (really 10 or 11?)
  - Value entered that doesn’t exist for variable?
    - Example: 2 entered where 1=male, 0=female
  - Missing values?
    - Did the person not give an answer? Was answer accidentally not entered into the database?

Data Cleaning (cont.)

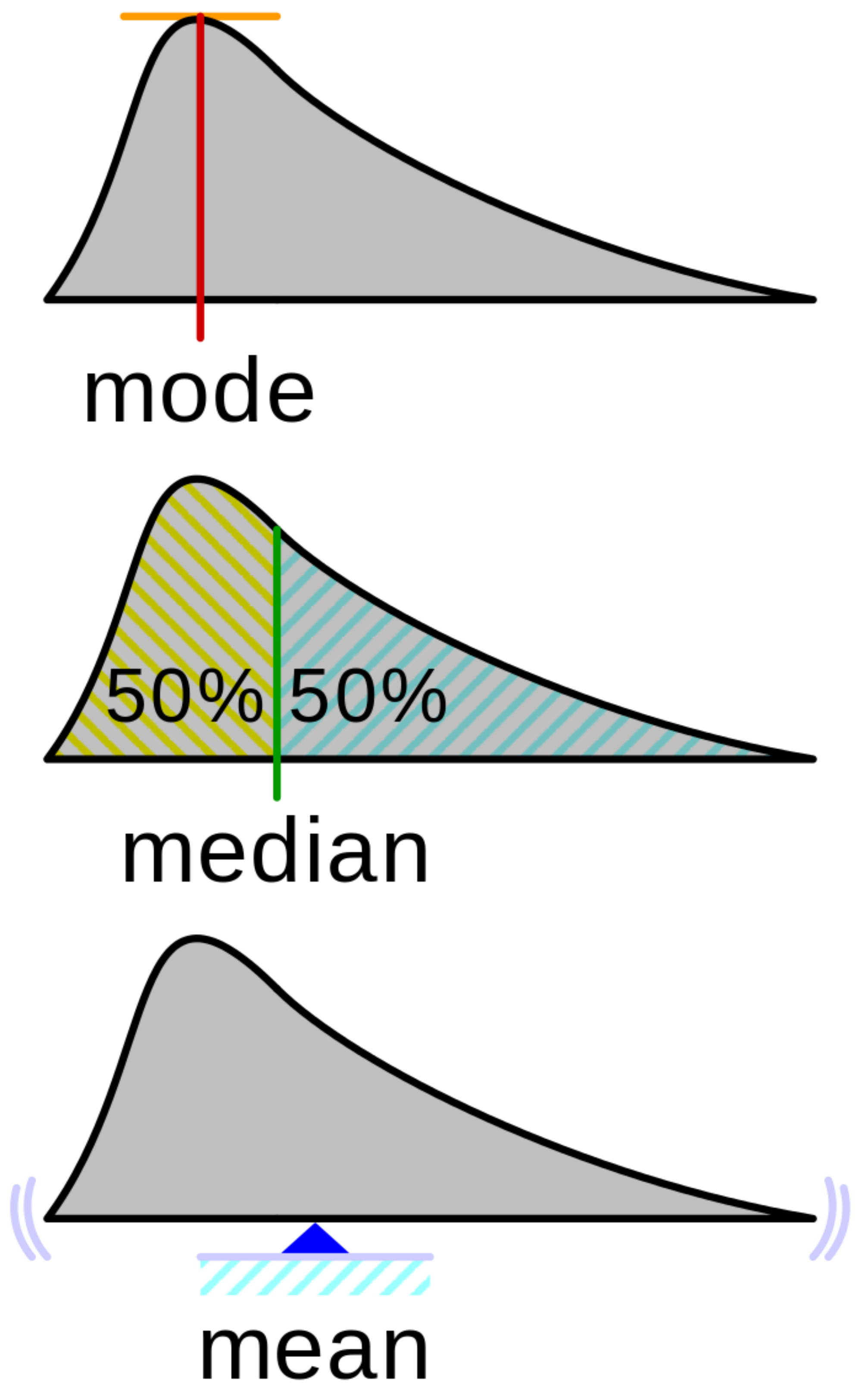
- May be able to set defined limits when entering data
  - Prevents entering a 2 when only 1, 0, or missing are acceptable values
- Limits can be set for continuous and nominal variables
  - Examples:
    - Only allowing 3 digits for age
    - Limiting words that can be entered
    - Assigning field types (e.g. formatting dates as mm/dd/yyyy or specifying numeric values or text)
- Many data entry systems allow “double-entry”
  - i.e., entering the data twice and then comparing both entries for discrepancies
- Univariate data analysis is a useful way to check the quality of the data

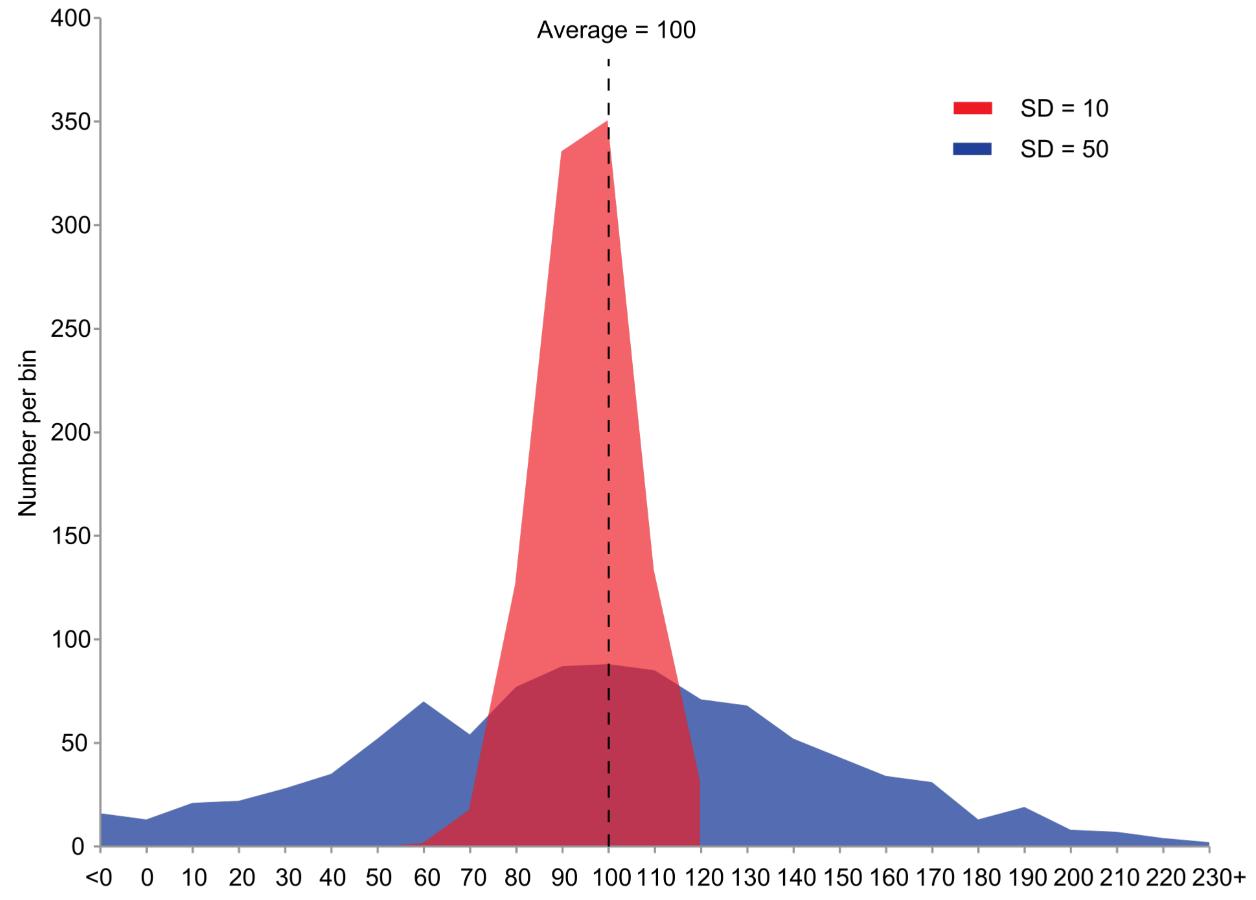
Univariate Data Analysis

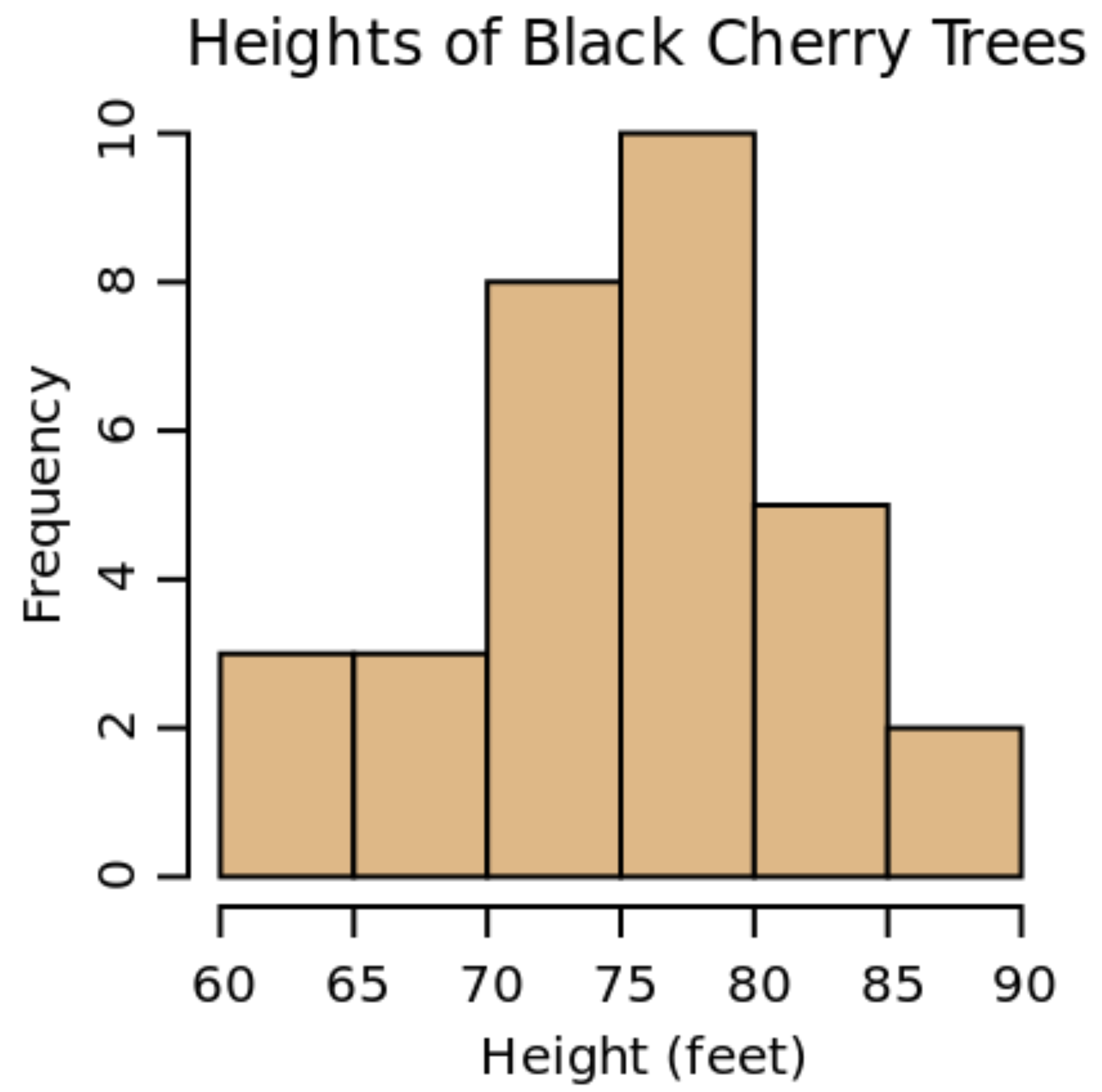
- Univariate data analysis explores each variable, attribute in a data set separately
  - Serves as a good method to check the quality of the data
  - Inconsistencies or unexpected results should be investigated using the original data as the reference point
    - can help you identify data cleaning problems
- Examining variables can give you important information
  - Do all subjects have data, or are values missing?
  - Are most values clumped together, or is there a lot of variation?
  - Are there outliers?
  - Do the minimum and maximum values make sense, or could there be mistakes in the coding?

Common Descriptive Statistics

- Measures of central tendency
  - Mean
  - Median
  - Mode
  - Range
- Measures of data variation
  - Standard deviation
  - Variance
  - Kurtosis, Skewness
  - Percentiles
- Counting data
  - Histogram
  - Frequency distributions
  - Percentage distributions

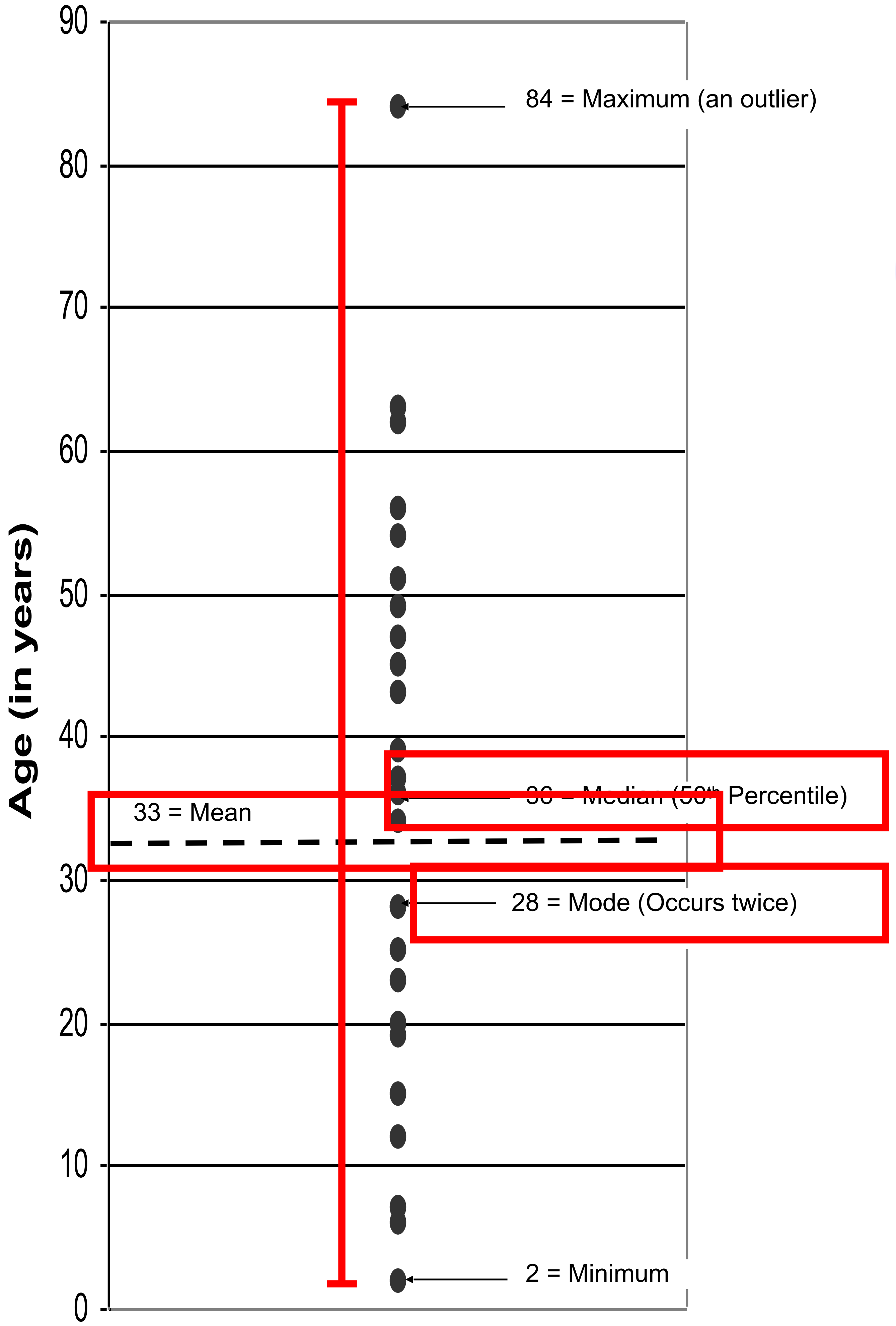


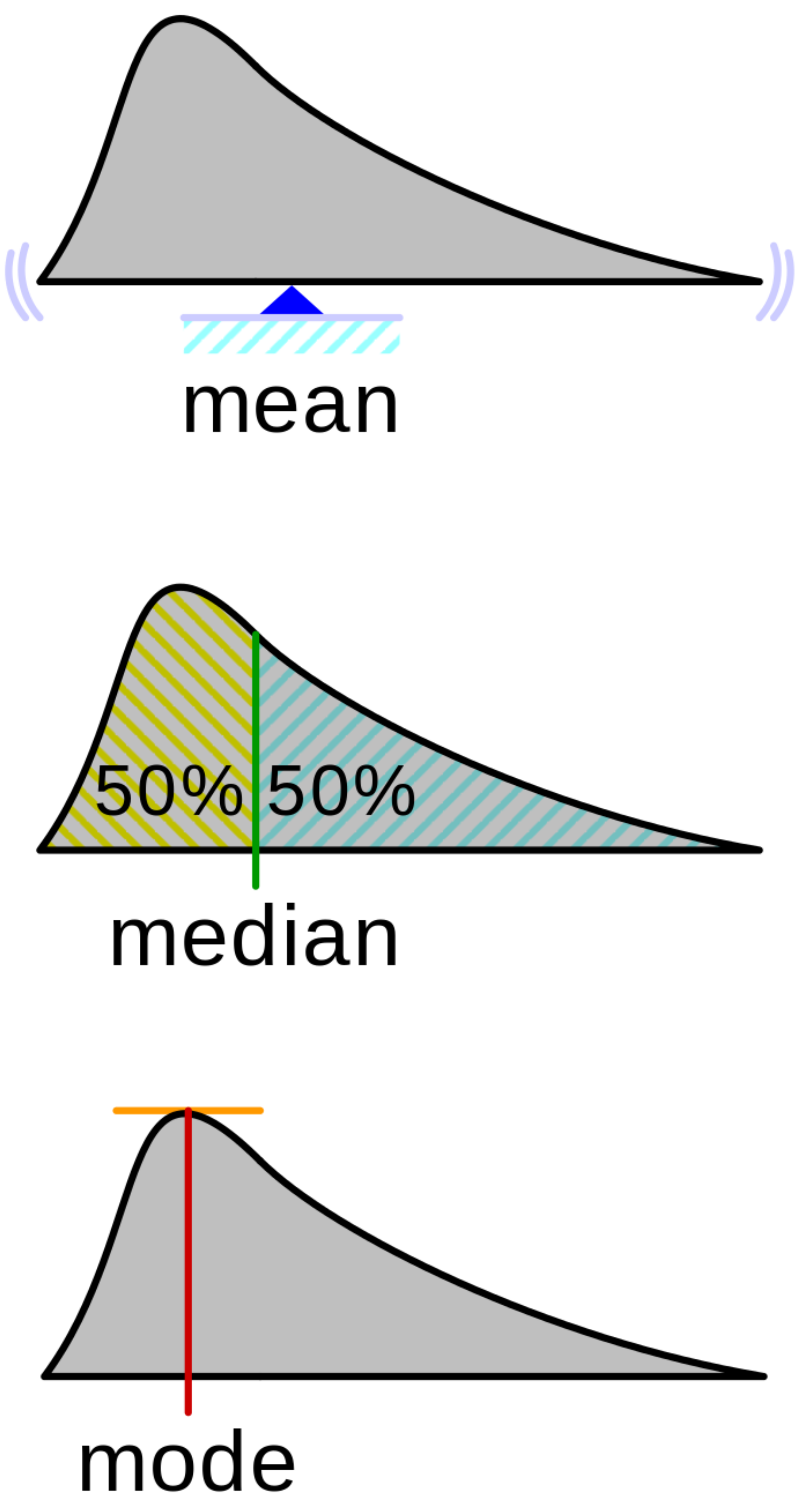




Measures of Central Tendency

- Commonly used statistics with univariate analysis of continuous variables
  - Mean – average of all values x_i of this variable in the dataset
    $$\mu = \frac{1}{n}\sum_{i=0}^{n-1} x_i$$
  - Median – the middle of the distribution, the number where half of the values are above and half are below
    - for n ordered values x_i with i=0…n-1
    $$\tilde{x} = \begin{cases} x_{\frac{n-1}{2}} & ,\text{if } \frac{n}{2} \neq 0 \\ (x_{\frac{n}{2}-1} + x_{\frac{n}{2}})/2 & ,\text{else} \end{cases}$$
  - Mode – the value that occurs the most times
    - has highest probability $x_{Mo} = \underset{i}{\operatorname{argmax}}\, P(x_i)$
  - Range of values – from minimum value to maximum value





Means and Medians

Math	98
English	96
History	95
Music	94
Biology	93
Latin	92
Gym	40

40	92	93	94	95	96	98

Mean = 87

Median = 94

Means and Medians

Math	100
English	94
History	55
Music	100
Biology	100
Latin	50
Gym	40

40	50	55	94	100	100	100

Mean = 77

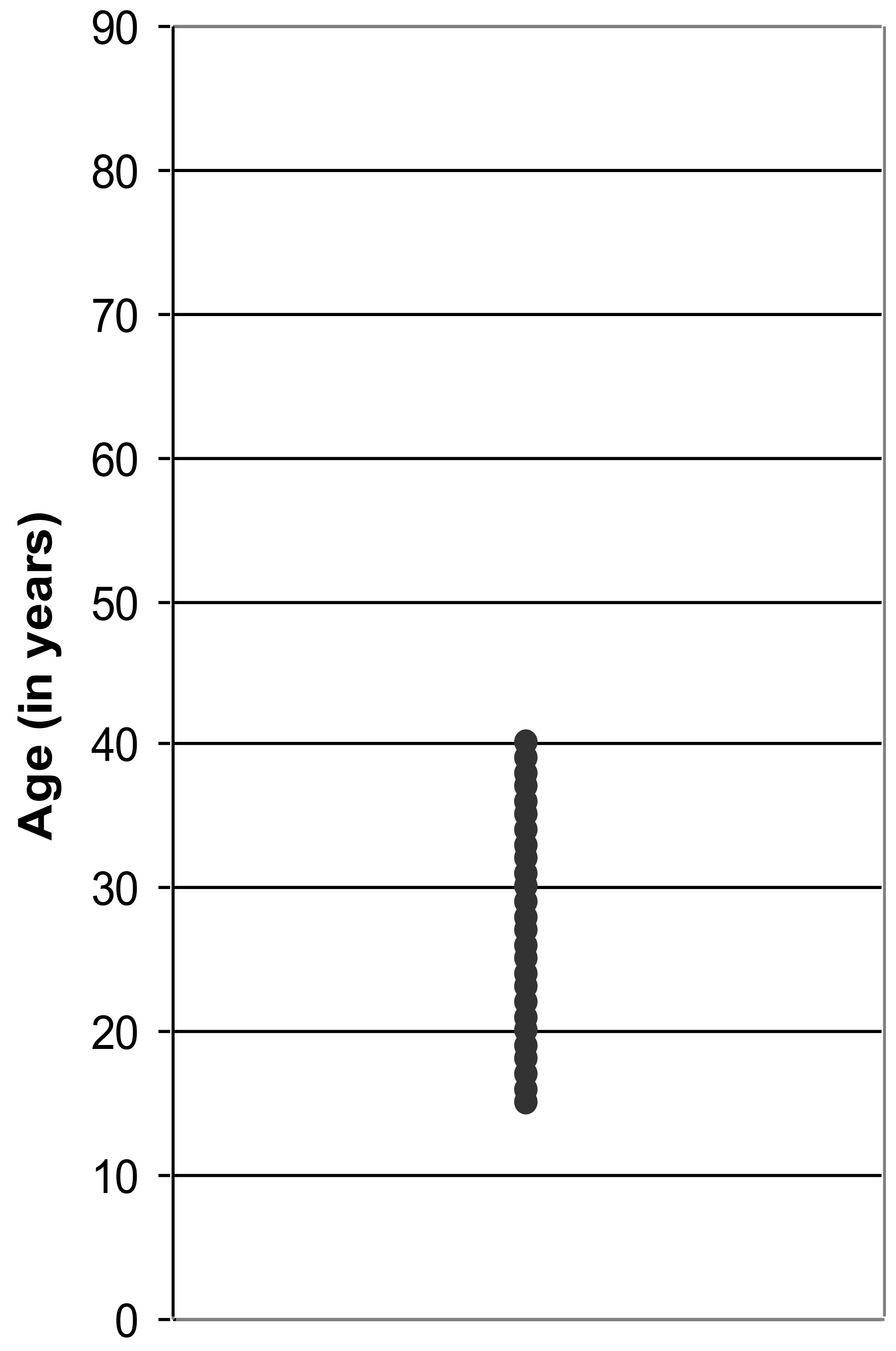
Median = 94

Means and Medians

Math	100	98
English	94	96
History	55	95
Music	100	94
Biology	100	93
Latin	50	92
Gym	40	40

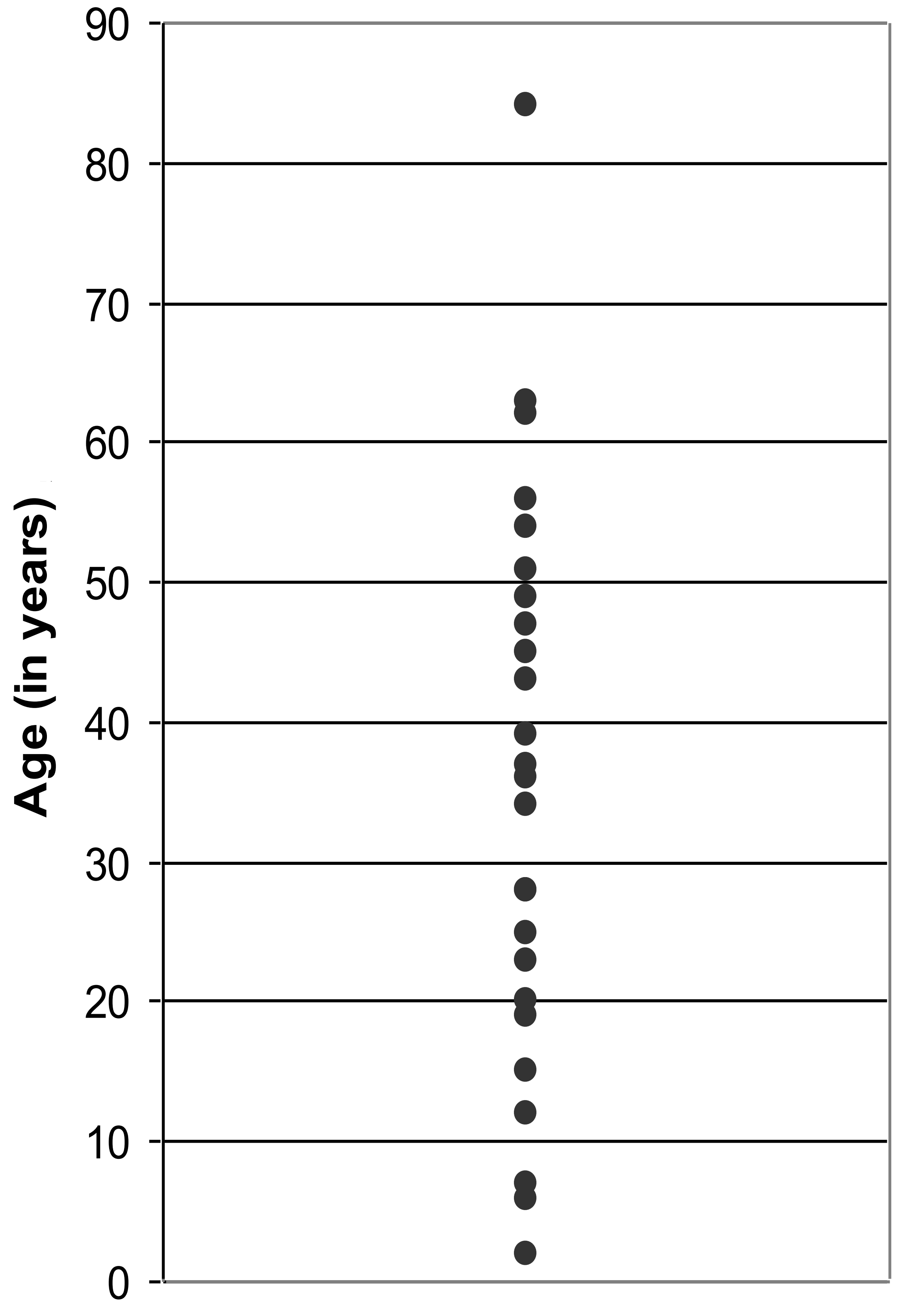
40	92	93	94	95	96	98	Mean = 87 Median = 94
40	50	55	94	100	100	100	Mean = 77 Median = 94

Standard Deviation



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

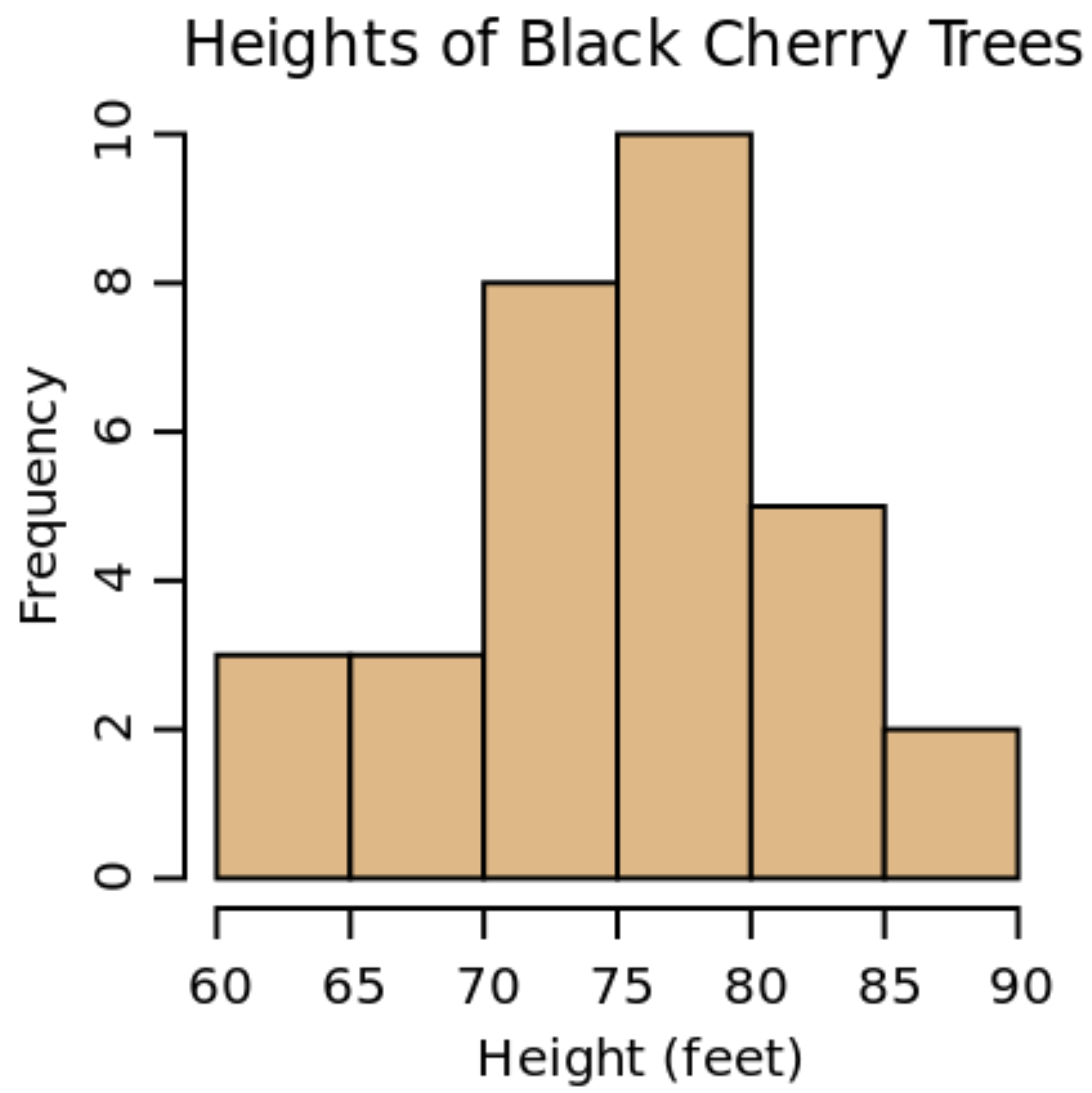
Second standardized statistical central moment

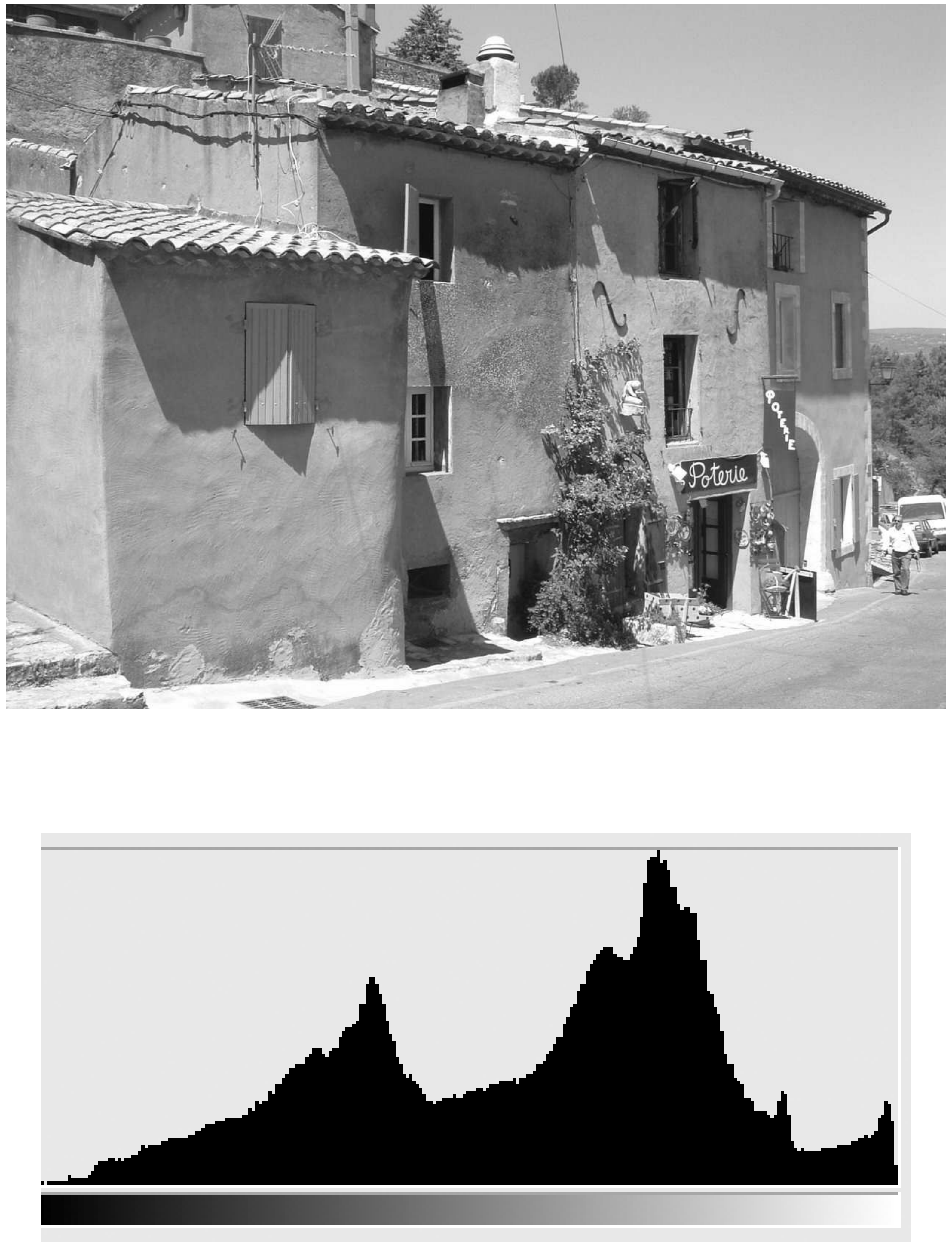


- Narrowly distributed age values (SD = 7.6)
- Widely distributed age values (SD = 20.4)

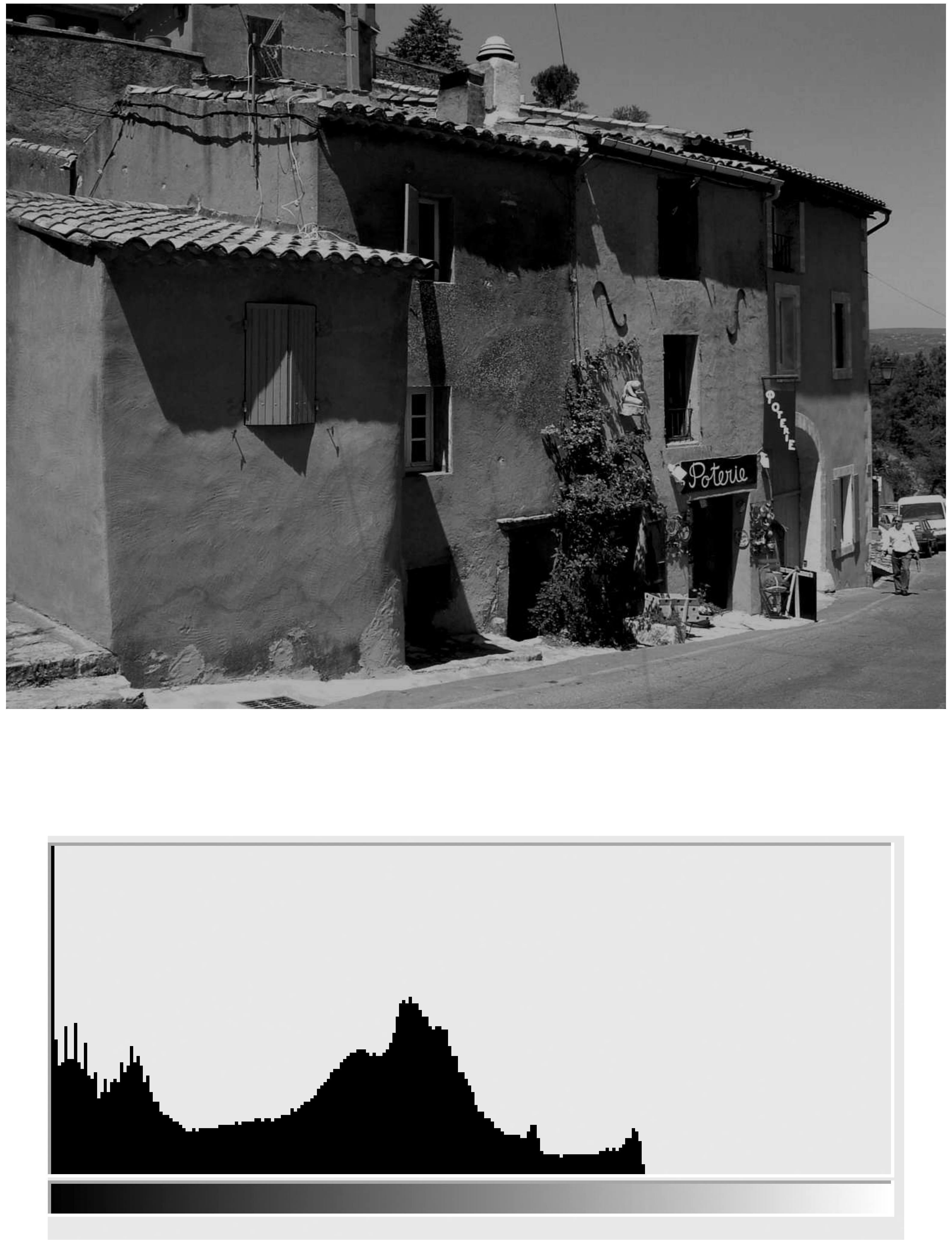
Histograms

- Graphical representation of the distribution of numerical data
- Estimate of the probability distribution of a continuous variable (quantitative variable)
- Construction:
  - First **bin** the range of values
    - divide the entire range of values into a series of intervals
    - bins are usually specified as consecutive, non-overlapping intervals of a variable
    - must be adjacent, and are often (but are not required to be) of equal size
  - Count how many values fall into each bin

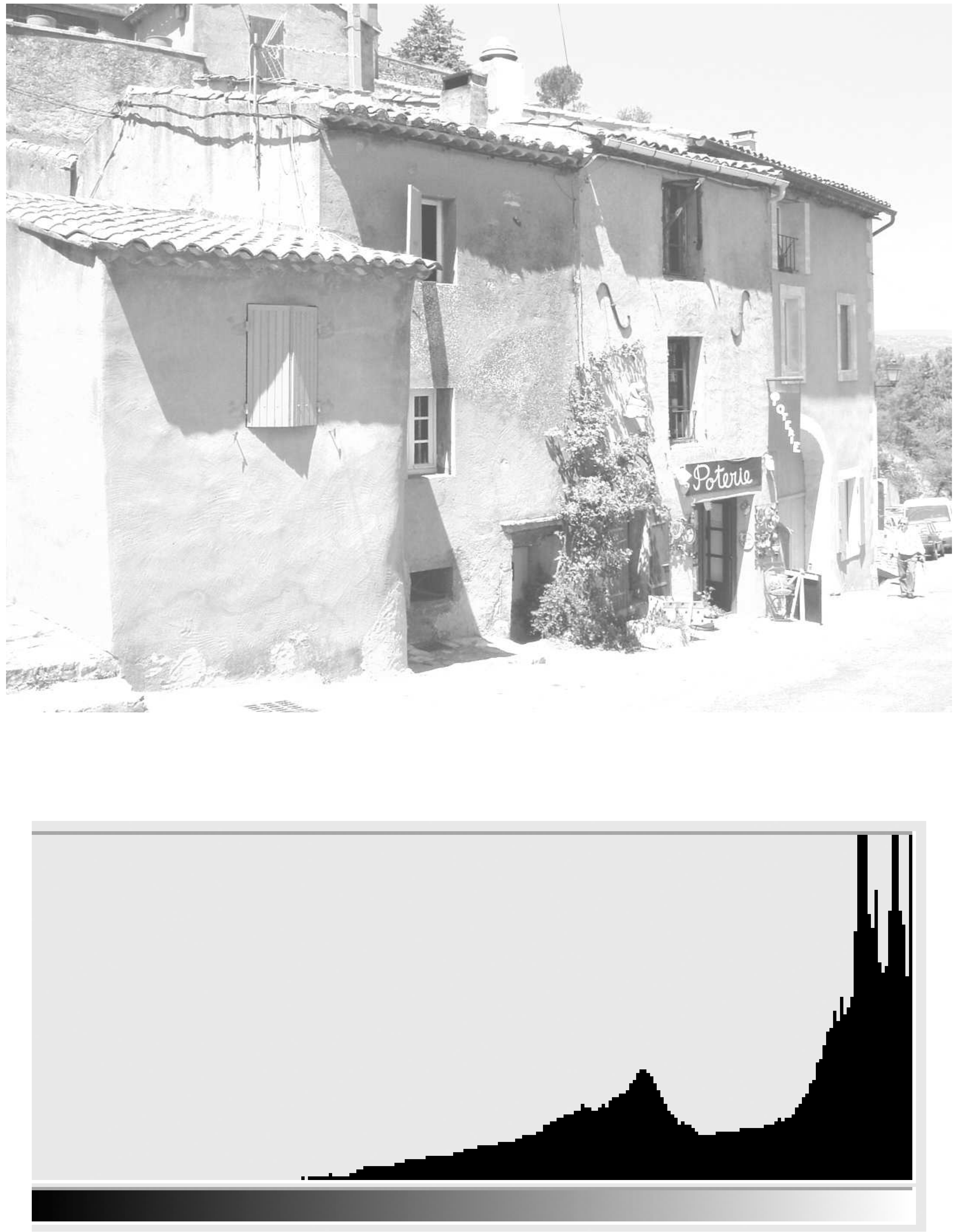




normal



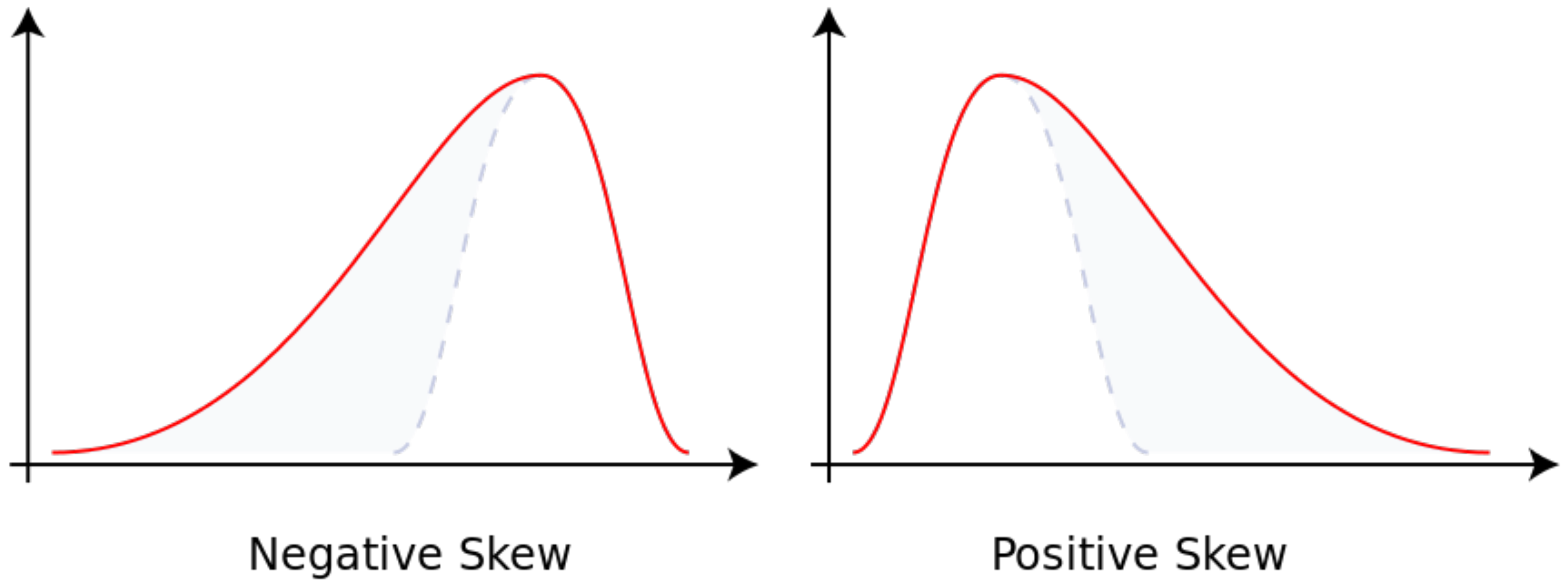
underexposed



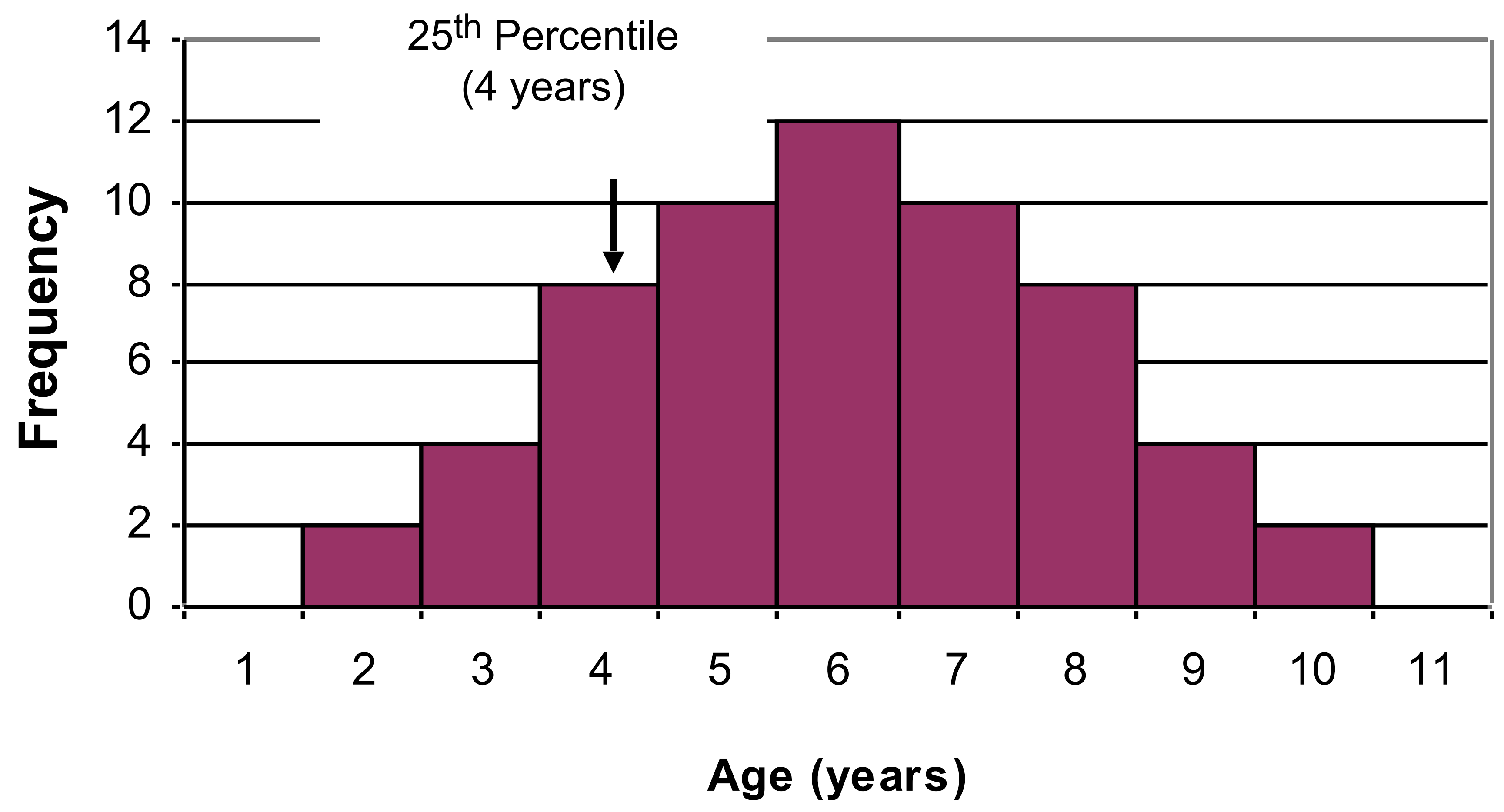
overexposed

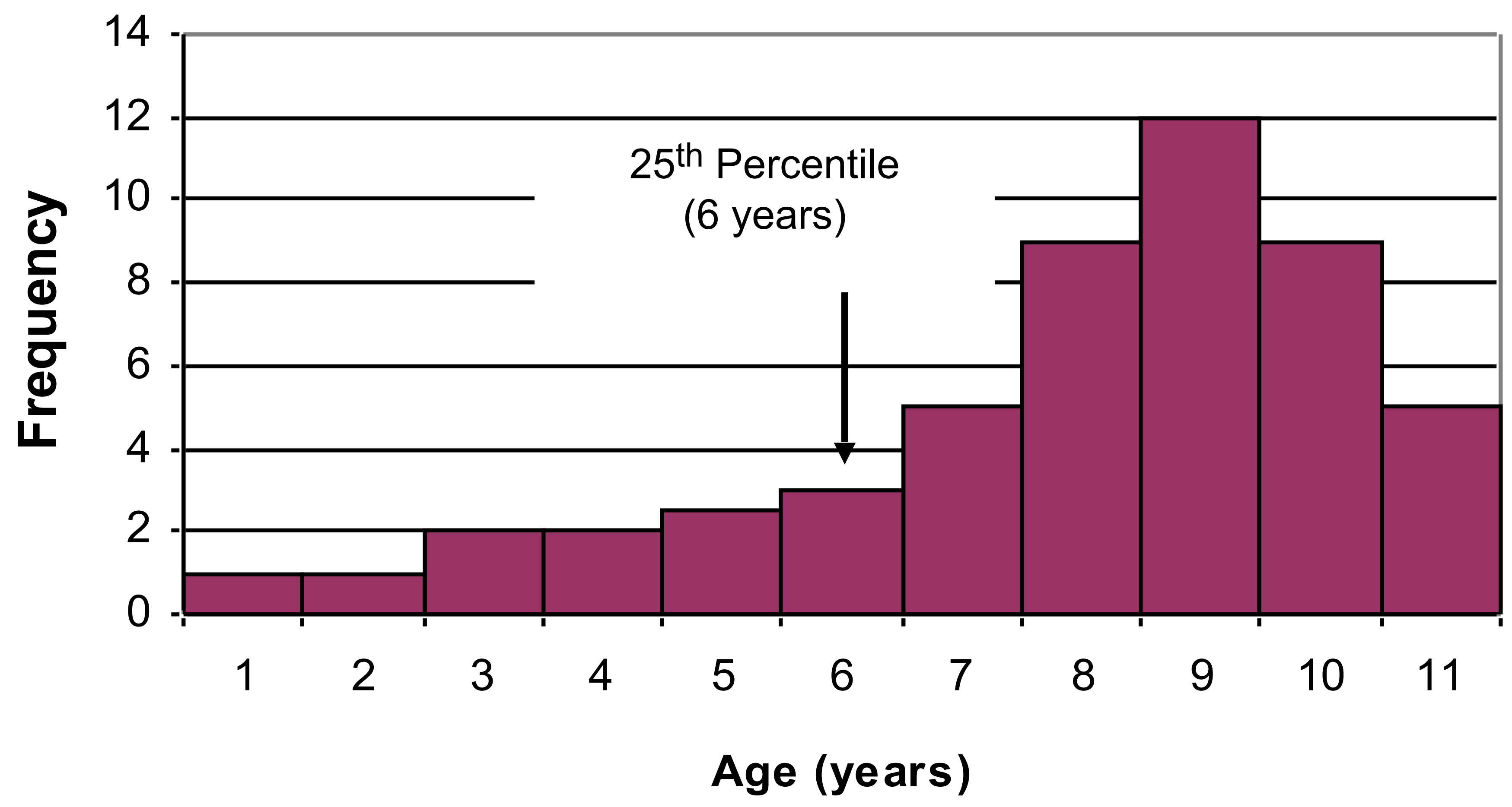
Percentiles and Skewness

- Percentiles – the percent of the distribution that is equal to or below a certain value
  - Median is the 50th percentile
- Skewness – whether most values occur low in the range, high in the range, or grouped in the middle
  - Third standardized statistical moment



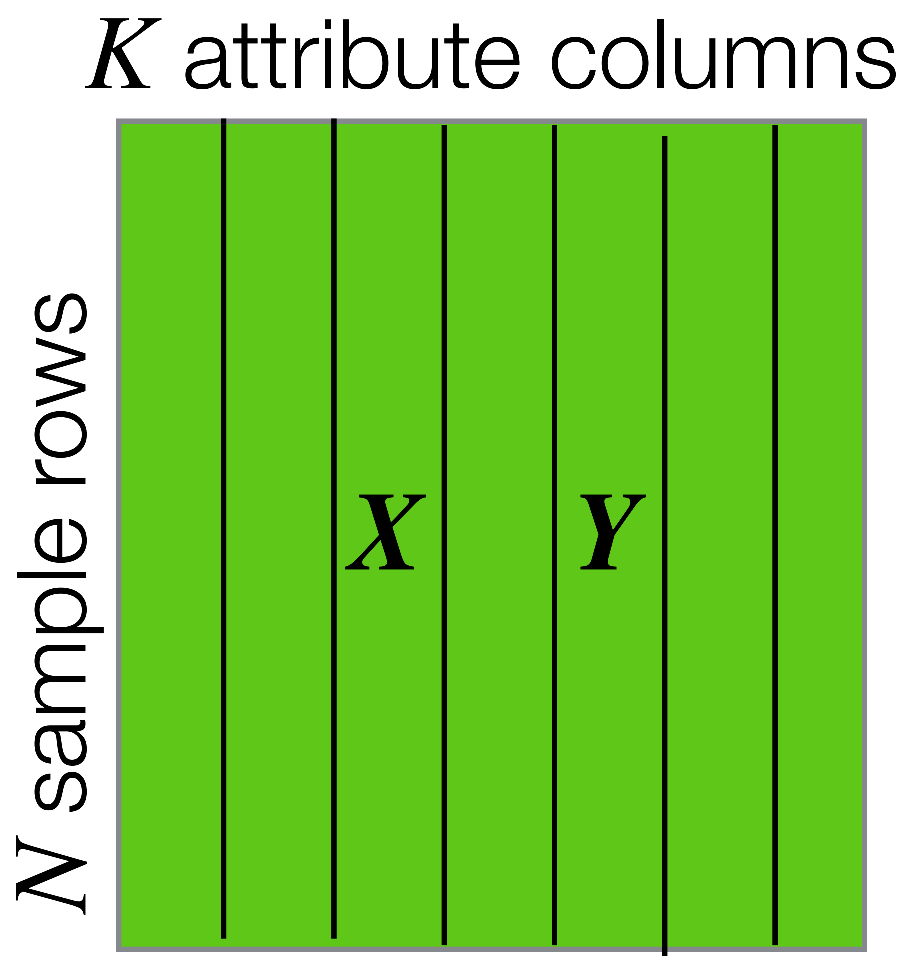
Distribution curves for variable AGE

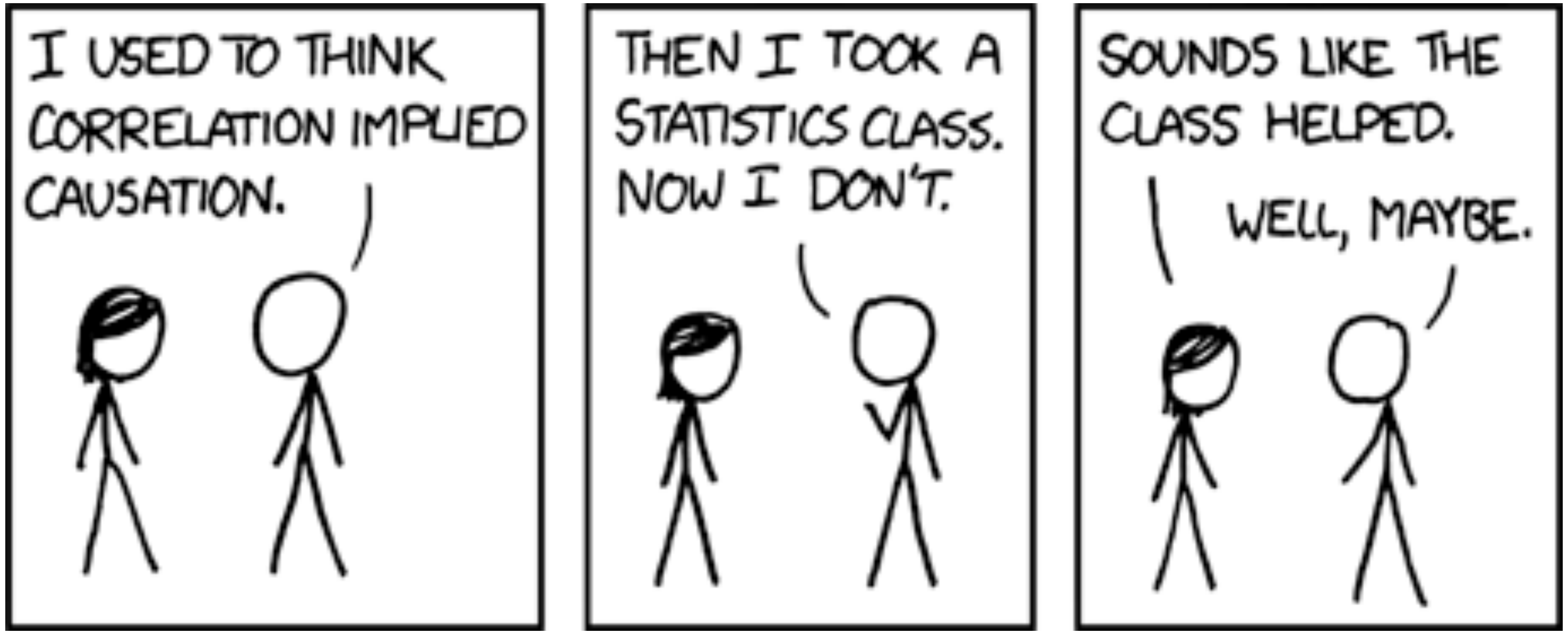




Correlation

- Measuring statistical association between two scalar variables $x_i \in \boldsymbol{X}$ and $y_i \in \boldsymbol{Y}$ (data columns)
  - Possibly, but not necessarily, involving causality





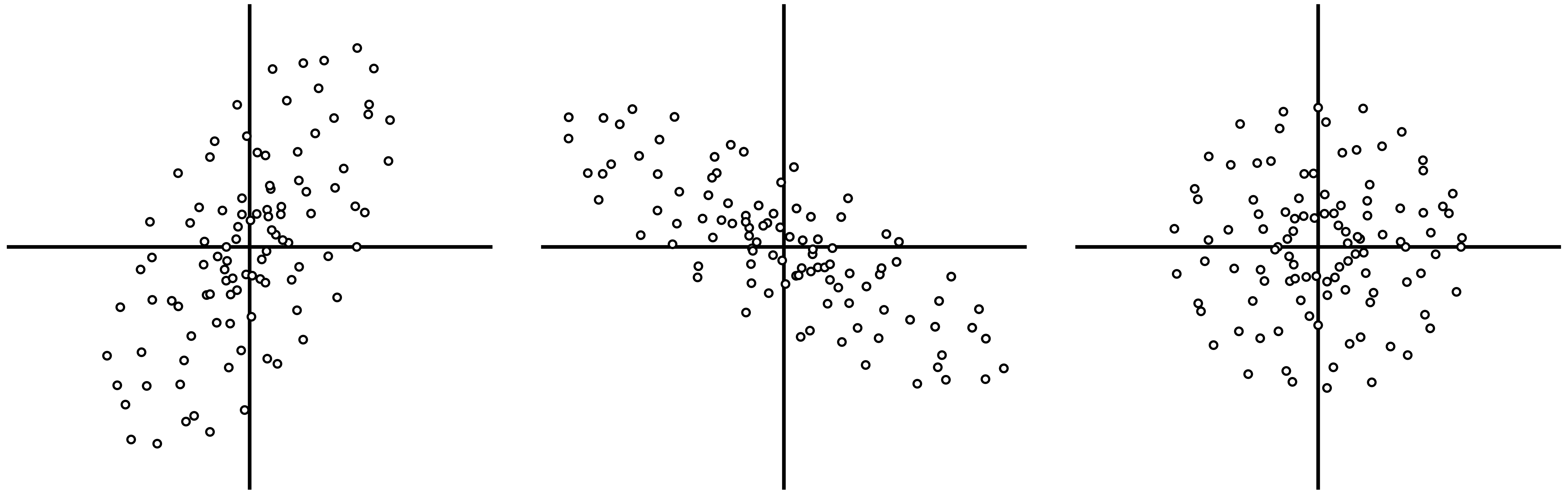
xkcd.com/552/

Review of Correlation

- Broad class of statistical relationships, often referring to how close variables are to a linear relationship
- Pearson correlation coefficient, most common but sensitive only to a linear relationship
  - Based on the covariance between sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ and their std-deviations $\boldsymbol{\sigma}$

$$\rho_{\boldsymbol{X},\boldsymbol{Y}} = corr(\boldsymbol{X},\boldsymbol{Y}) = \frac{cov(\boldsymbol{X},\boldsymbol{Y})}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}}$$

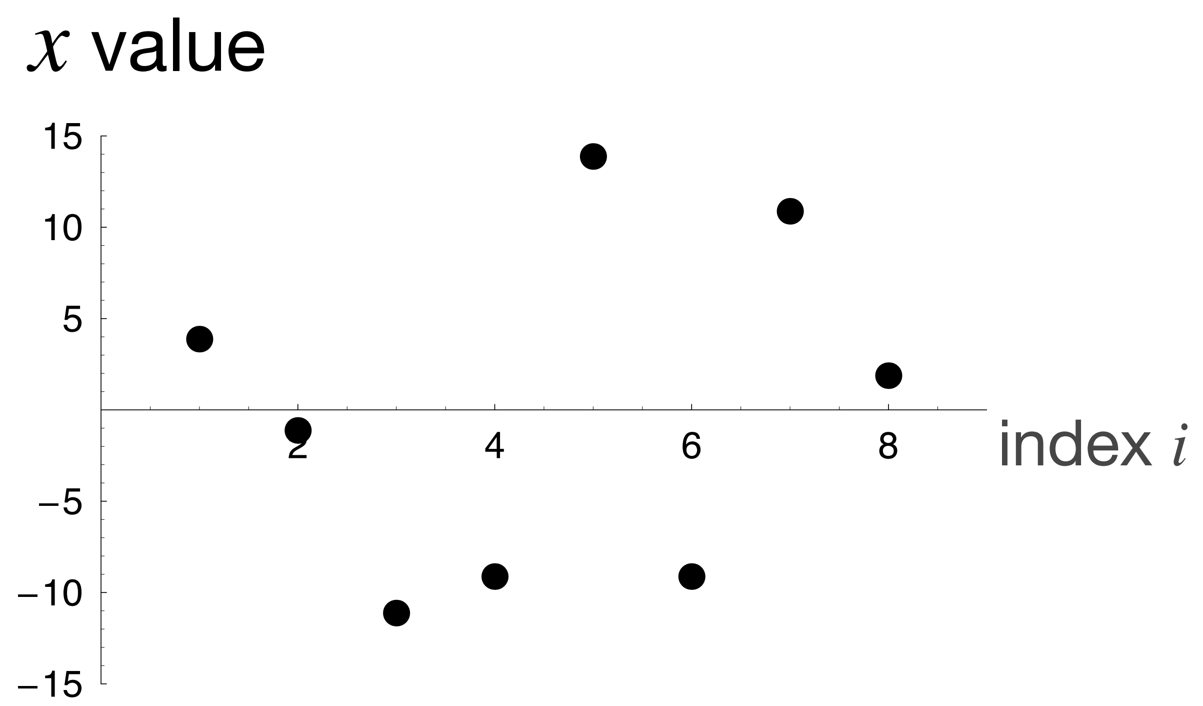
$$cov(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\boldsymbol{X}})(y_i - \mu_{\boldsymbol{Y}})$$

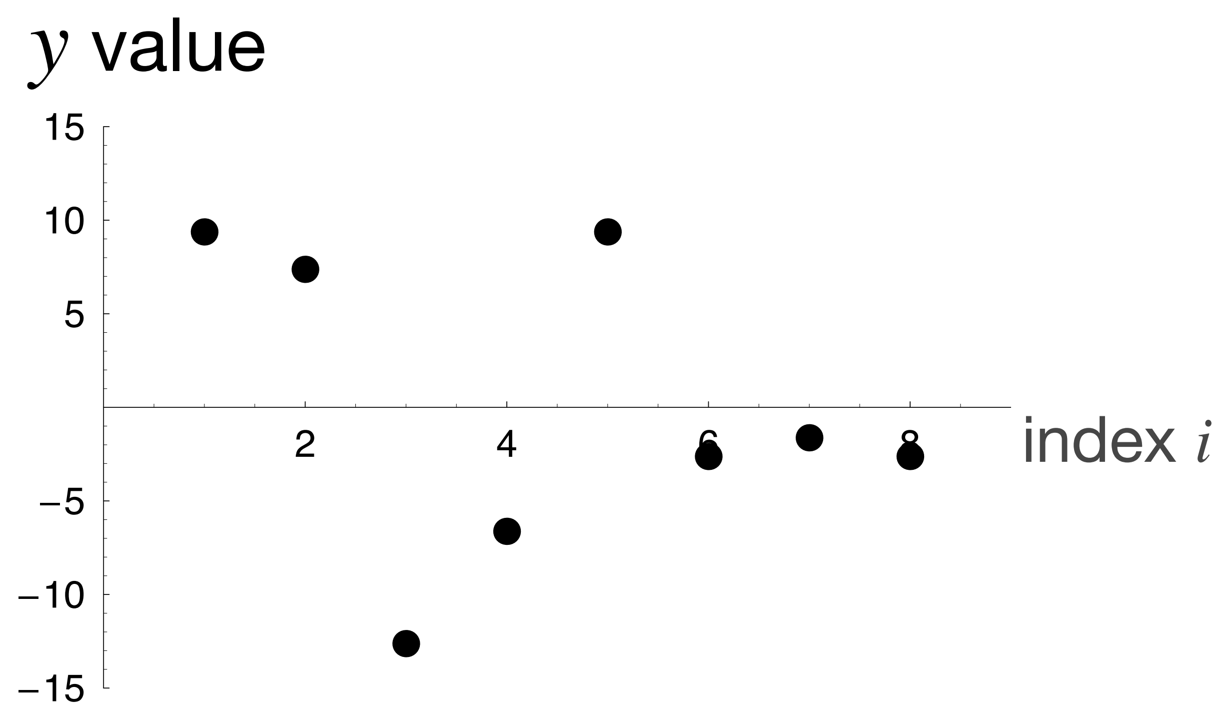


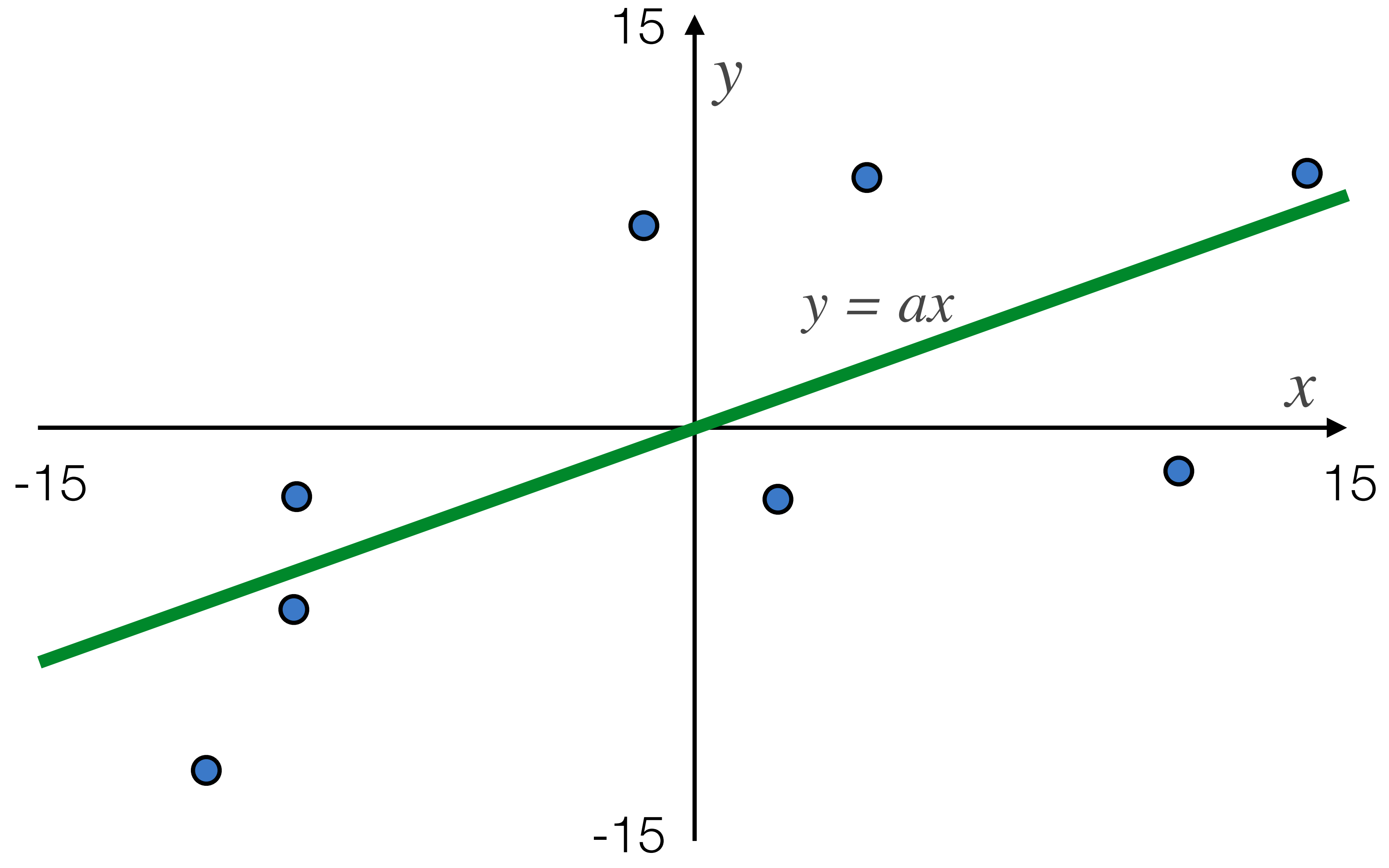
p = 0.7 p = -0.7 p = 0.0

Linear Regression

- Given two scalar variables $x_i \in \boldsymbol{X}$ and $y_i \in \boldsymbol{Y}$
  - E.g. zero mean *weights* and *heights*
    - with means $\mu_x = 61.125\mathrm{kg}$ and $\mu_y = 162.625\mathrm{cm}$
- Plot y_i against x_i
- Find (best) scalar factor a such that forall i: $y_i \simeq a\, x_i$
  - but this is overdetermined
- Matrix form $\boldsymbol{y} = a\, \boldsymbol{x}$ with solution $a = \dfrac{\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{x}}$
  - $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ and $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ are scalars

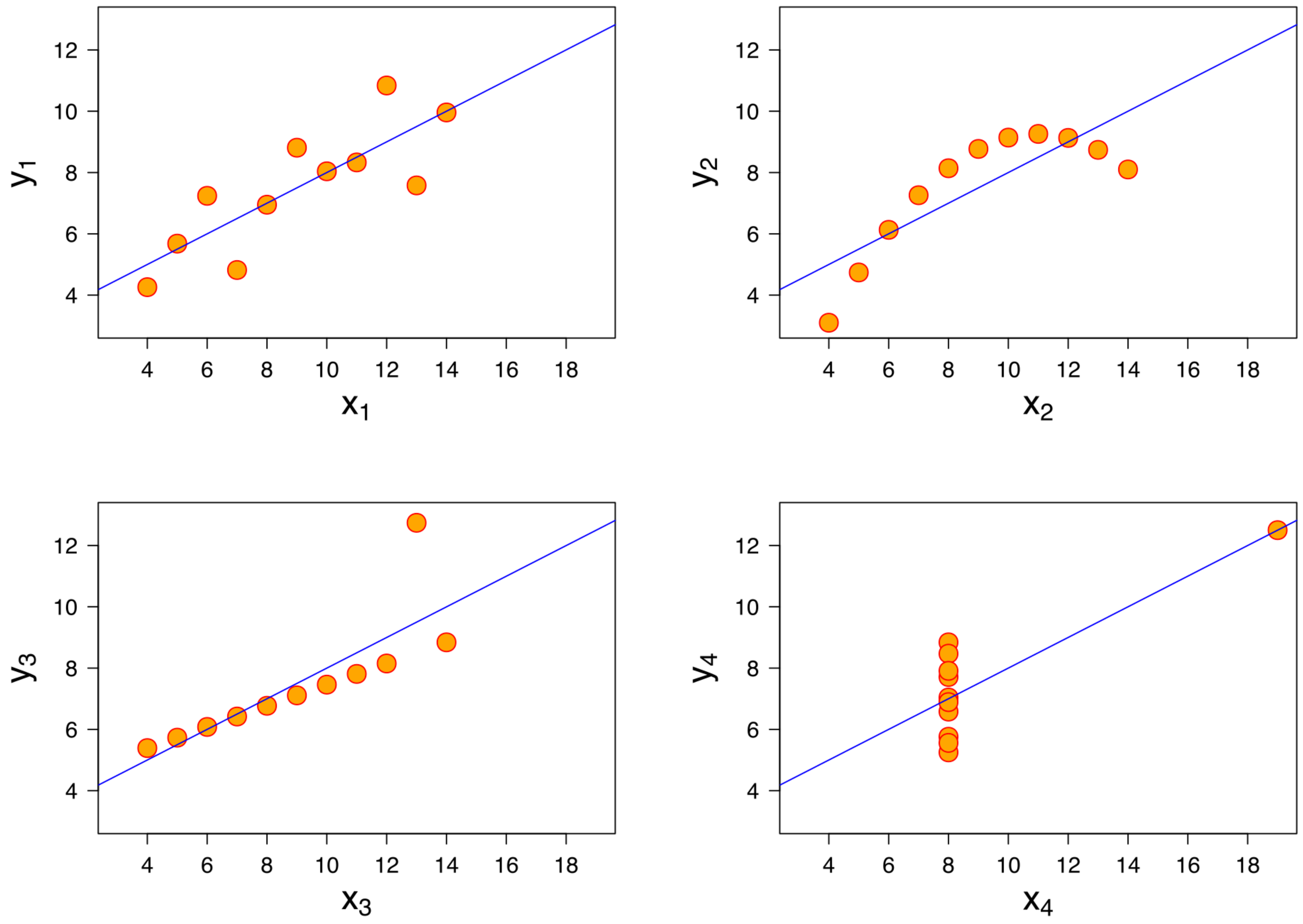






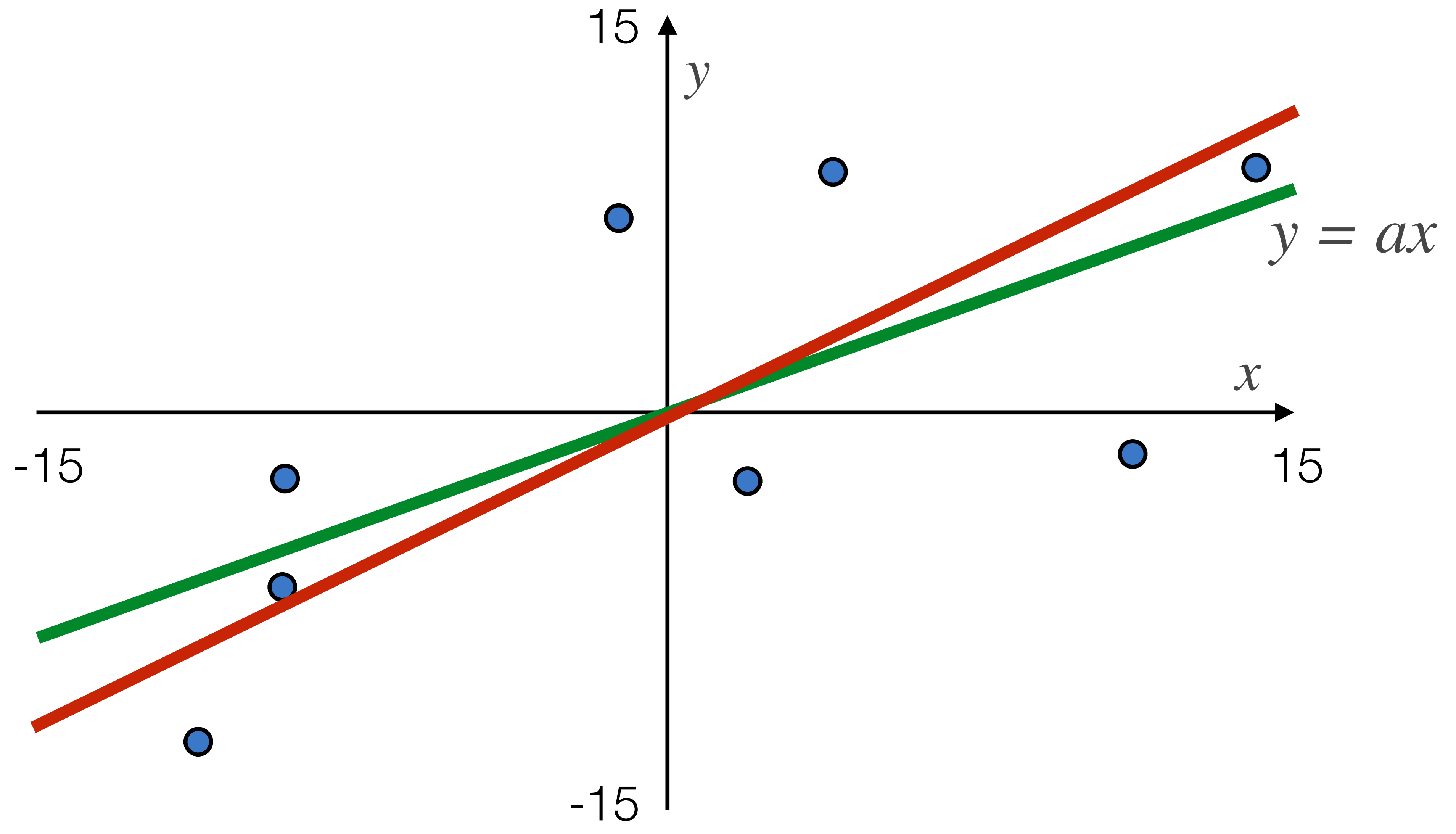
Anscombe's Quartet

- Linear regression may be the same for strongly varying sample data distribution
  - Even having identical means, standard deviation and correlations
- General problem of models fitted to discrete data



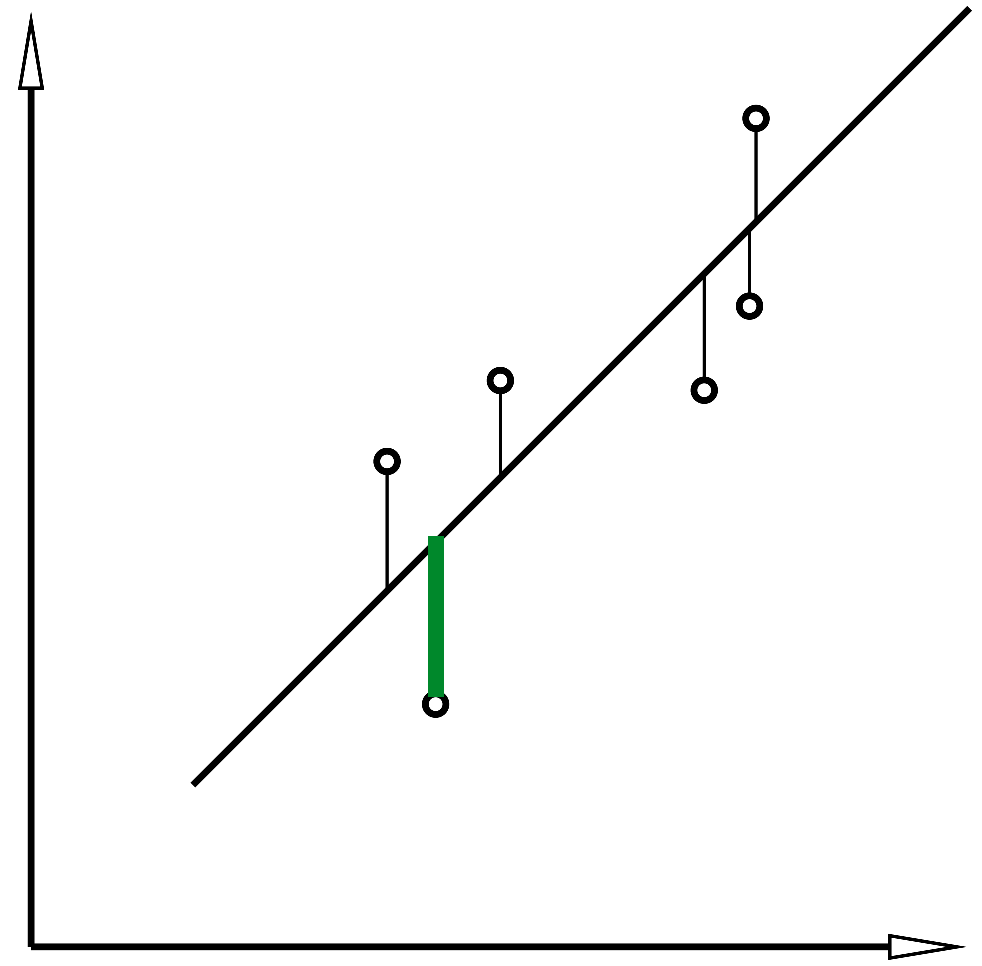
Dominant Line

- Given two scalar variables $x_i \in \boldsymbol{X}$ and $y_i \in \boldsymbol{Y}$ organized into data matrix $\mathbf{D} = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_N & y_N \end{pmatrix}$
  - N mean-centered, zero-mean data values
- Eigenvector of largest eigenvalue of $\mathbf{D}^T\mathbf{D}$ describes the *dominant line*
  - $\mathbf{D}^T\mathbf{D}$ covariance matrix
  - Different from the linear regression

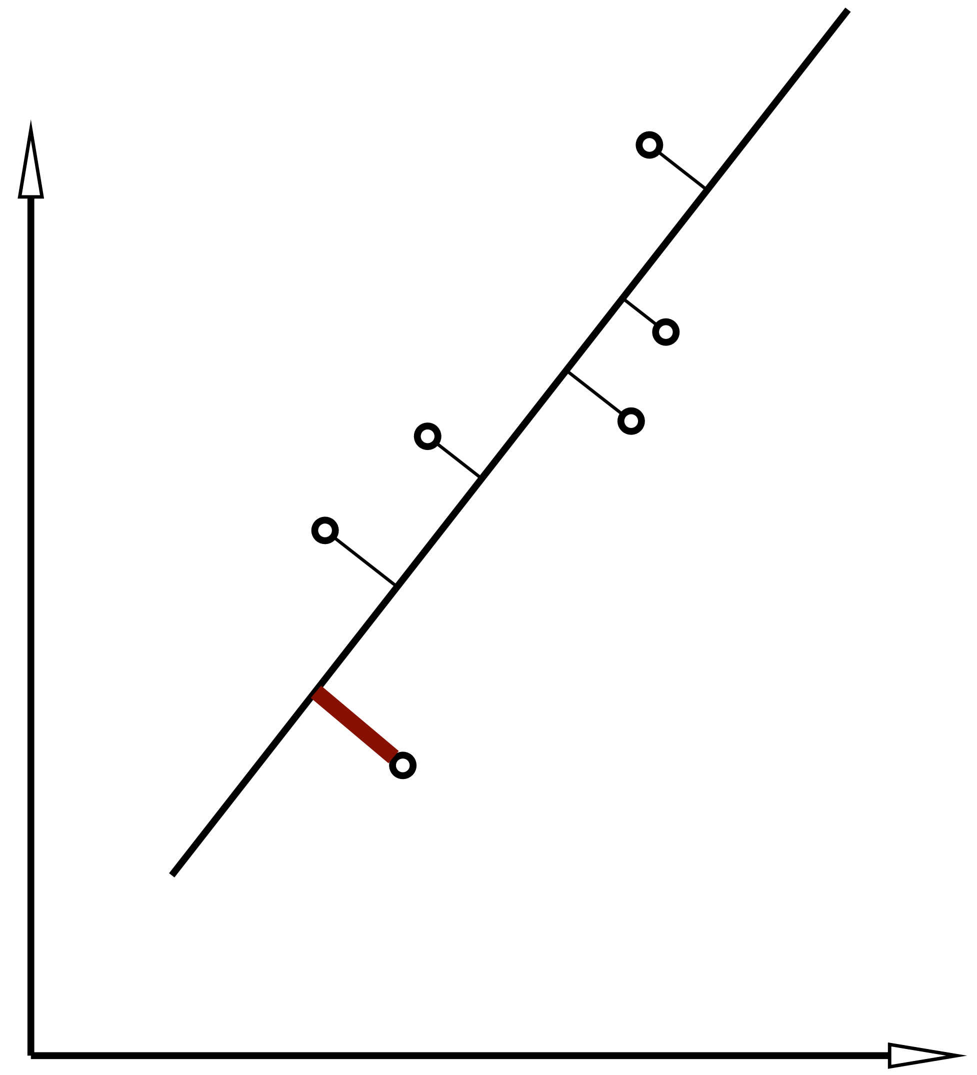


Regression and Dominant Lines

- Regression and dominant lines solve two different optimization problems
  - Regression finds line which minimizes the *vertical offsets* of data points
  - The dominant line minimizes the *perpendicular distances* of the data points



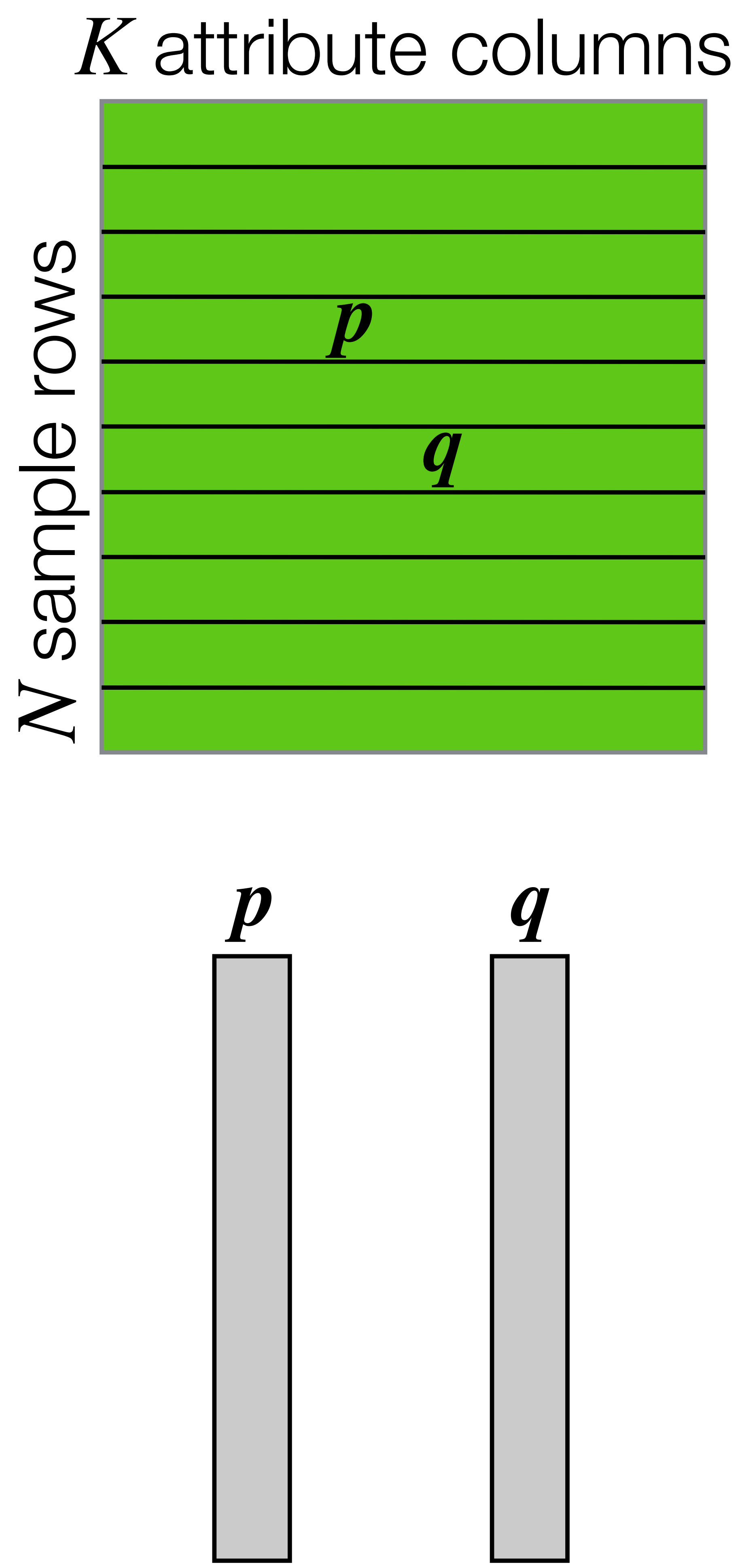
Linear regression



Dominant line

Measures of Similarity and Dissimilarity

- Defining how (dis)similar two data points $\boldsymbol{p}$ and $\boldsymbol{q}$ are is important in basic data analysis tasks
  - Two rows of data matrix, viewed as individual (K-dimensional) data vectors
- *Dissimilarity* measure: numerical measure of how different $\boldsymbol{p}$ and $\boldsymbol{q}$ are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- *Similarity* measure: numerical measure of how alike $\boldsymbol{p}$ and $\boldsymbol{q}$ are
  - Higher when objects are more alike
  - Often falls into normalized range [0,1]



Similarity Measures

- Measures proximity of data points $\boldsymbol{p}$ and $\boldsymbol{q} \in \mathbb{R}^K$ as function $s(\boldsymbol{p}, \boldsymbol{q})$
  - Satisfies properties:
    - $s(\boldsymbol{p}, \boldsymbol{q}) = s(\boldsymbol{q}, \boldsymbol{p})$
    - $s(\boldsymbol{p}, \boldsymbol{q}) \leq s(\boldsymbol{p}, \boldsymbol{p})$
  - Other typical properties:
    - $s(\boldsymbol{p}, \boldsymbol{q}) = 1 \quad \Leftrightarrow \quad \boldsymbol{p} = \boldsymbol{p}$
    - $s(\boldsymbol{p}, \boldsymbol{q}) \geq 0$
- Normalized similarity if $s(\boldsymbol{p}, \boldsymbol{p}) = 1$

- Cosine $\quad s(\boldsymbol{p}, \boldsymbol{q}) = \dfrac{\boldsymbol{p}^T\boldsymbol{q}}{\sqrt{\boldsymbol{p}^T\boldsymbol{p}} \cdot \sqrt{\boldsymbol{q}^T\boldsymbol{q}}}$
- Overlap $\quad s(\boldsymbol{p}, \boldsymbol{q}) = \dfrac{\boldsymbol{p}^T\boldsymbol{q}}{\min\left(\boldsymbol{p}^T\boldsymbol{p}, \boldsymbol{q}^T\boldsymbol{q}\right)}$
- Dice $\quad s(\boldsymbol{p}, \boldsymbol{q}) = \dfrac{2 \cdot \boldsymbol{p}^T\boldsymbol{q}}{\boldsymbol{p}^T\boldsymbol{p} + \boldsymbol{q}^T\boldsymbol{q}}$
- Pruducts $\boldsymbol{v}^T\boldsymbol{w}$ correspond to the scalar or dot-product between vectors $\boldsymbol{v}$ and $\boldsymbol{w}$

Similarity and Dissimilarity Measures

- Dissimilarity: numerical measure $d(\boldsymbol{p}, \boldsymbol{q})$ of how different two data objects $\boldsymbol{p}$ and $\boldsymbol{q}$ are
  - Common range from 0 (objects are alike) to ∞ (objects are different)
    - possibly within a restricted specific range $[d_{min}, d_{max}]$
  - May be defined also on non-numerical input data points
    - weighted sum of edge link connections between nodes in graph
    - edit distance between strings
- A similarity measure can be formed from a *dissimilarity* $d(\boldsymbol{p}, \boldsymbol{q})$:
  - $s(\boldsymbol{p}, \boldsymbol{q}) = 1 - d(\boldsymbol{p}, \boldsymbol{q})$ if d is in the range [0, 1]
  - $s(\boldsymbol{p}, \boldsymbol{q}) = 1 / (1 + d(\boldsymbol{p}, \boldsymbol{q}))$ if d is in the range [0, ∞]
  - $s(\boldsymbol{p}, \boldsymbol{q}) = e^{-d(\boldsymbol{p}, \boldsymbol{q})}$ if d is in the range [0, ∞]
  - $s(\boldsymbol{p}, \boldsymbol{q}) = 1 - (d(\boldsymbol{p}, \boldsymbol{q}) - d_{min}) / (d_{max} - d_{min})$ if d is in the range $[d_{min}, d_{max}]$

Distance Metric

- Distance metric $d(\boldsymbol{p}, \boldsymbol{q})$ between data points $\boldsymbol{p}$ and $\boldsymbol{q} \in \mathbb{R}^K$
  - Satisfies properties:
    - $d(\boldsymbol{p}, \boldsymbol{q}) = d(\boldsymbol{q}, \boldsymbol{p})$
    - $d(\boldsymbol{p}, \boldsymbol{q}) = 0 \quad \Leftrightarrow \quad \boldsymbol{p} = \boldsymbol{q}$
    - $d(\boldsymbol{p}, \boldsymbol{q}) \leq d(\boldsymbol{p}, \boldsymbol{r}) + d(\boldsymbol{r}, \boldsymbol{q})$
    - $d(\boldsymbol{p}, \boldsymbol{q}) \geq 0$

- Norm ||.|| as distance metric defined on $\boldsymbol{p}$ - $\boldsymbol{q}$, thus $d(\boldsymbol{p}, \boldsymbol{q}) = ||\boldsymbol{p} - \boldsymbol{q}||$
  - Satisfies properties:
    - $||\boldsymbol{p}|| = 0 \quad \Leftrightarrow \quad \boldsymbol{p} = (0, 0, \dots 0)$
    - $||a \cdot \boldsymbol{p}|| = |a| \cdot ||\boldsymbol{p}||$
    - $||\boldsymbol{p} + \boldsymbol{q}|| \leq ||\boldsymbol{p}|| + ||\boldsymbol{q}||$
    - $||\boldsymbol{p}|| \geq 0$

- Minkowski or p-norm on vector $\boldsymbol{v} \in \mathbb{R}^K$
  - $p = 2$ Euclidean norm $\quad ||\boldsymbol{v}||_p = \sqrt[p]{\sum_{i=1}^{K} |v_i|^p}$
  - $p = 1$ Manhattan norm $\quad ||\boldsymbol{v}||_1 = \sum_i |v_i|$
  - $p = \pm\infty$ infinity norms $\quad ||\boldsymbol{v}||_{\pm\infty} = \max/\min |v_i|$
    - supremum/infimum

Recap

- **Review of variable types:** nominal categorical, ordinal ranked or numerical
- **Univariate data analysis:** descriptive statistics; mean, median, mode, standard deviation, histograms, percentiles and skewness
- **Correlation:** Pearson correlation, linear regression, Anscombe's quartet, dominant line
- **Distance and similarity:** distance metric and norm; cosine, overlap and dice similarity measures; similarity from dissimilarity

- Required textbook Chapter(s): 12.1 - 12.3