Human-Computer Interaction

Statistics I: Descriptive Statistics

Professor Bilge Mutlu

Questions

To ask questions during class:

- Go to slido.com and use code **#2938904** or direct link or scan QR code
- Anonymous
- I will monitor during class

Today's Agenda

- Topic overview: *overview; descriptive statistics*
- Hands-on activity

Why do we need to use statistics?

Statistical methods enable us to analyze quantitative data, specifically (1) to inspect data quality and characteristics and (2) to discover relationships (e.g., causal) among experimental variables or to estimate population characteristics.

1 ↠ **Descriptive** statistics

2 ↠ **Inferential** statistics

*What is the difference between **descriptive** and **inferential** statistics?*

A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of collected data, while **descriptive statistics** is the process of using and analyzing those statistics.[1]

Inferential statistics, or statistical inference (or modeling), is the process making propositions about a population using data drawn from the population through sampling.[2]

Simply put, using descriptive statistics, we summarize a sample of data; using inferential statistics, we make propositions about the population.

[1] Wikipedia: Desciptive Statistics

[2] Wikipedia: Inferential Statistics

When do we use descriptive and inferential statistics?

Usually, descriptive and inferential statistics are used together.

Descriptive statistics:

- To assess data quality and structure
- To describe population characteristics
- To assess dependence among variables

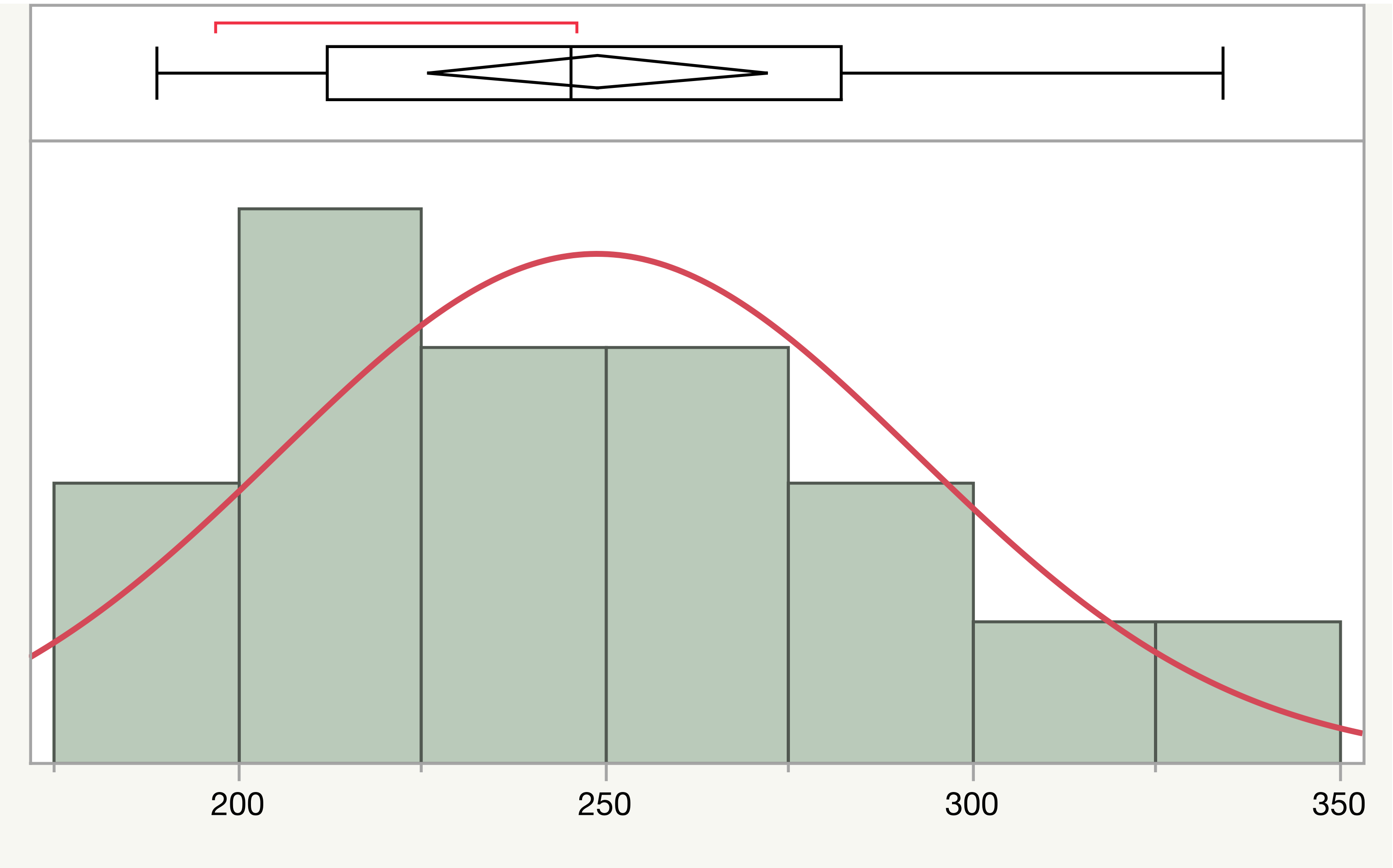
Inferential statistics:

- To test hypotheses
- To estimate parameters
- To perform clustering or classification

How do we perform descriptive statistics?

First, by preparing our data table and inspecting our data distribution.[3]

Group	Participants	Task Completion Time
No prediction	Participant 1	245
No prediction	Participant 2	236
No prediction	Participant 3	321
No prediction	Participant 4	212
No prediction	Participant 5	267
No prediction	Participant 6	334
No prediction	Participant 7	287
No prediction	Participant 8	259
With prediction	Participant 9	246
With prediction	Participant 10	213
With prediction	Participant 11	265
With prediction	Participant 12	189
With prediction	Participant 13	201
With prediction	Participant 14	197
With prediction	Participant 15	289
With prediction	Participant 16	224



[3] Lazar et al., 2017, Chapter 4

What are the types of analyses in descriptive statistics?

Univariate analysis involves describing the distribution of a single variable, including *type/form* of distribution, *central tendency*, and *dispersion.*

Bivariate or **multivariate analysis** involves describing the relationships between pairs of variables in terms of *correlation*, *covariance*, and *slope.*

What do we look at in univariate analysis?

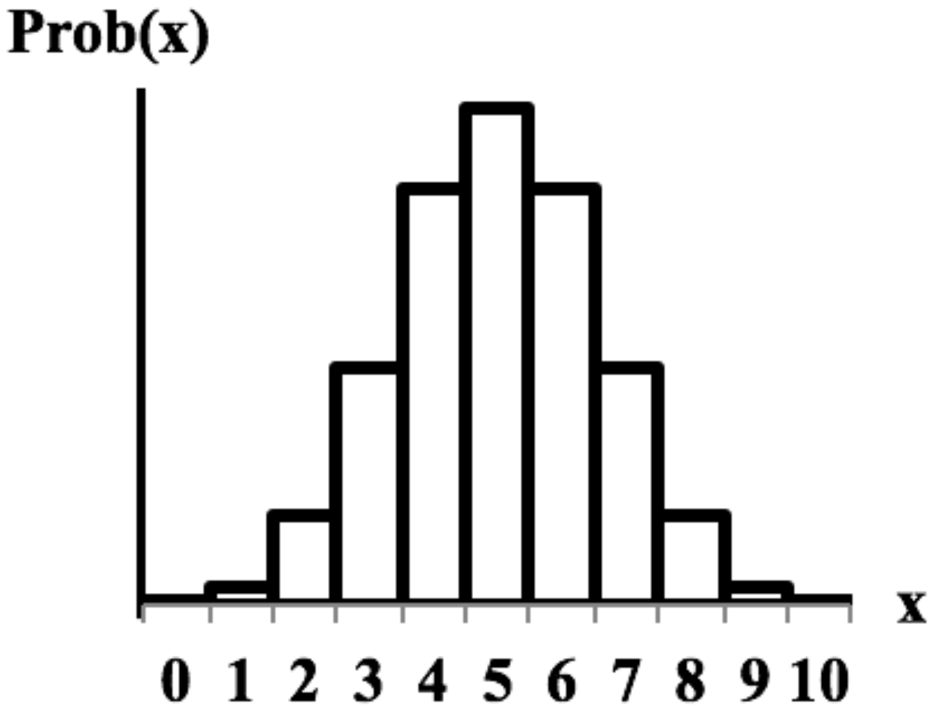
1. Distribution — what does our distribution look like?[4]
2. Central tendency — where is the majority of our data?[5]
3. Dispersion — how much does the deviate from the center?[5]

[4] For discrete, ordinal, or continuous data types

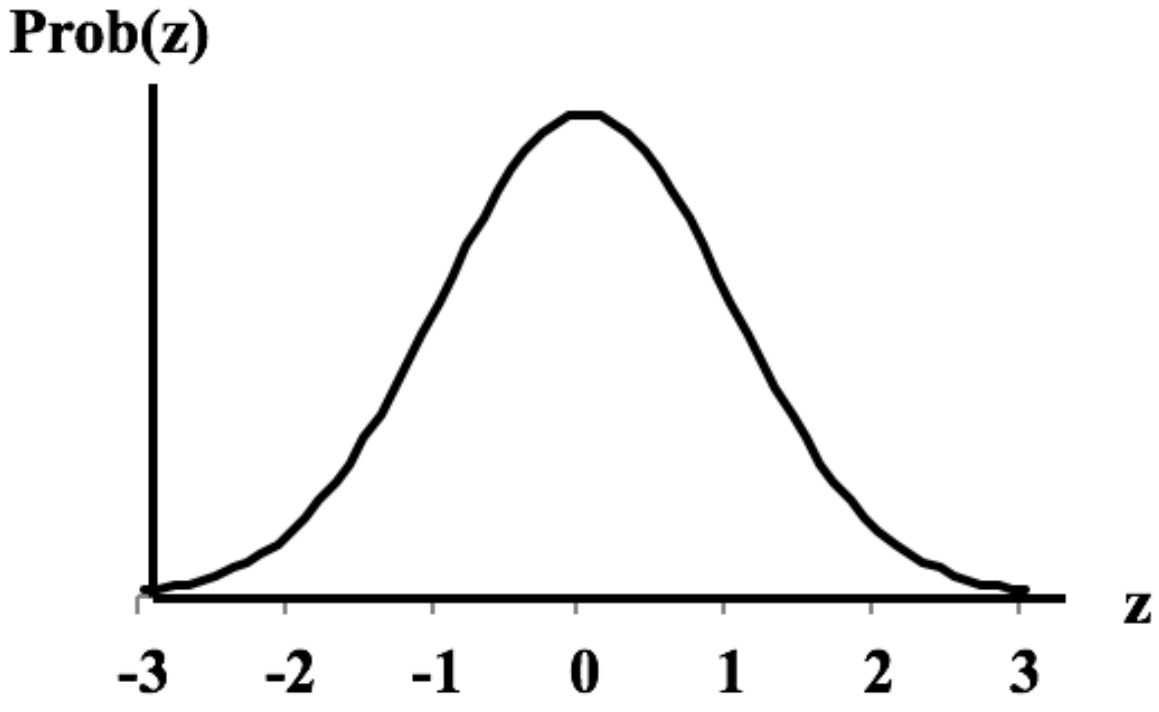
[5] For continuous data types only

Distribution[6]

Distributions can be **discrete** or **continuous**.



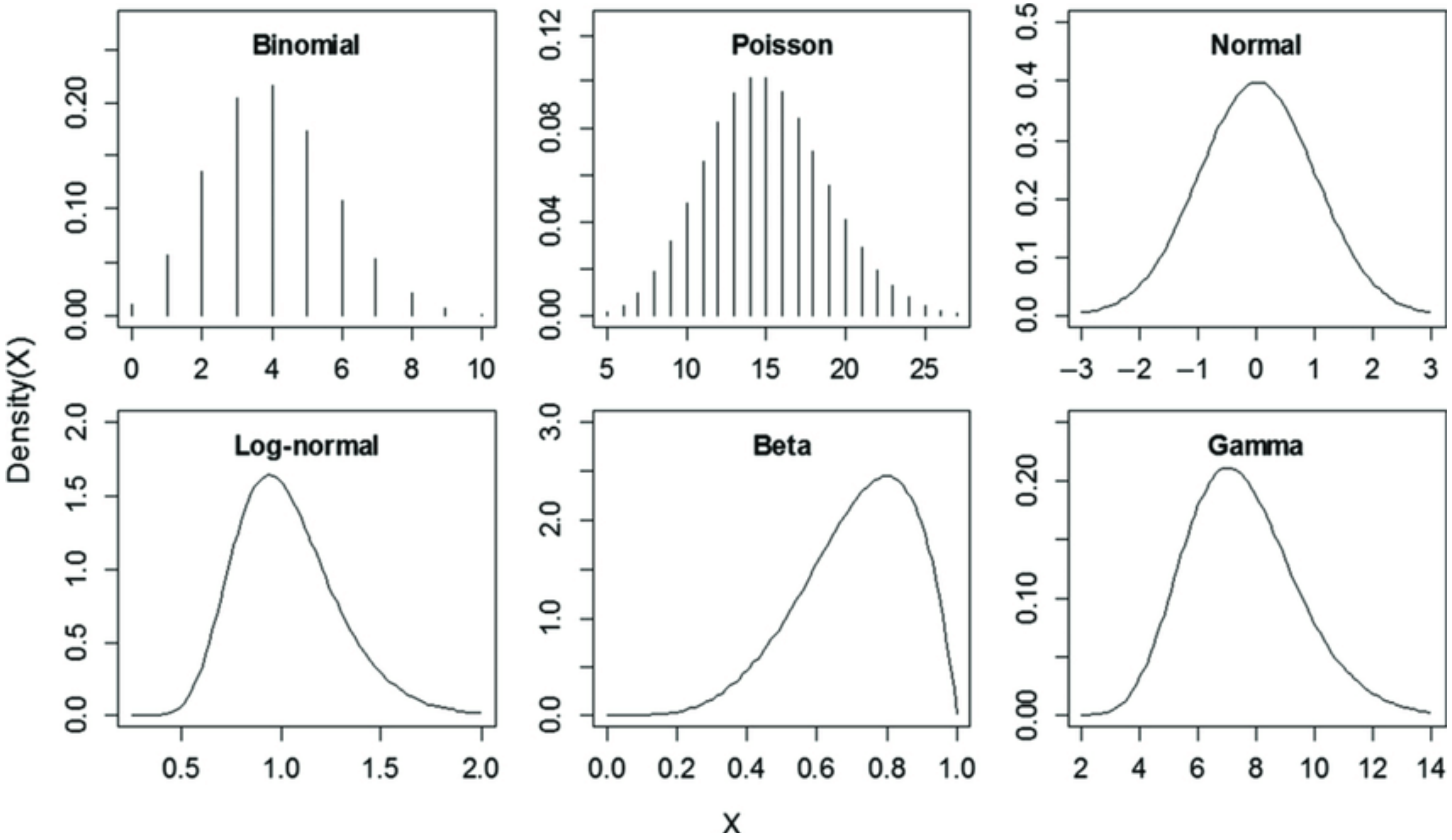
Binomial Distribution
Discrete Data & Discrete
Probability Curve



Standard Normal Distribution
Continuous Data and Continuous
Probability Curve

[6] Image source

Data from discrete or continuous variables can take different forms and follow different probability distributions.[7]



[7] Image source: Daniel Wolcott

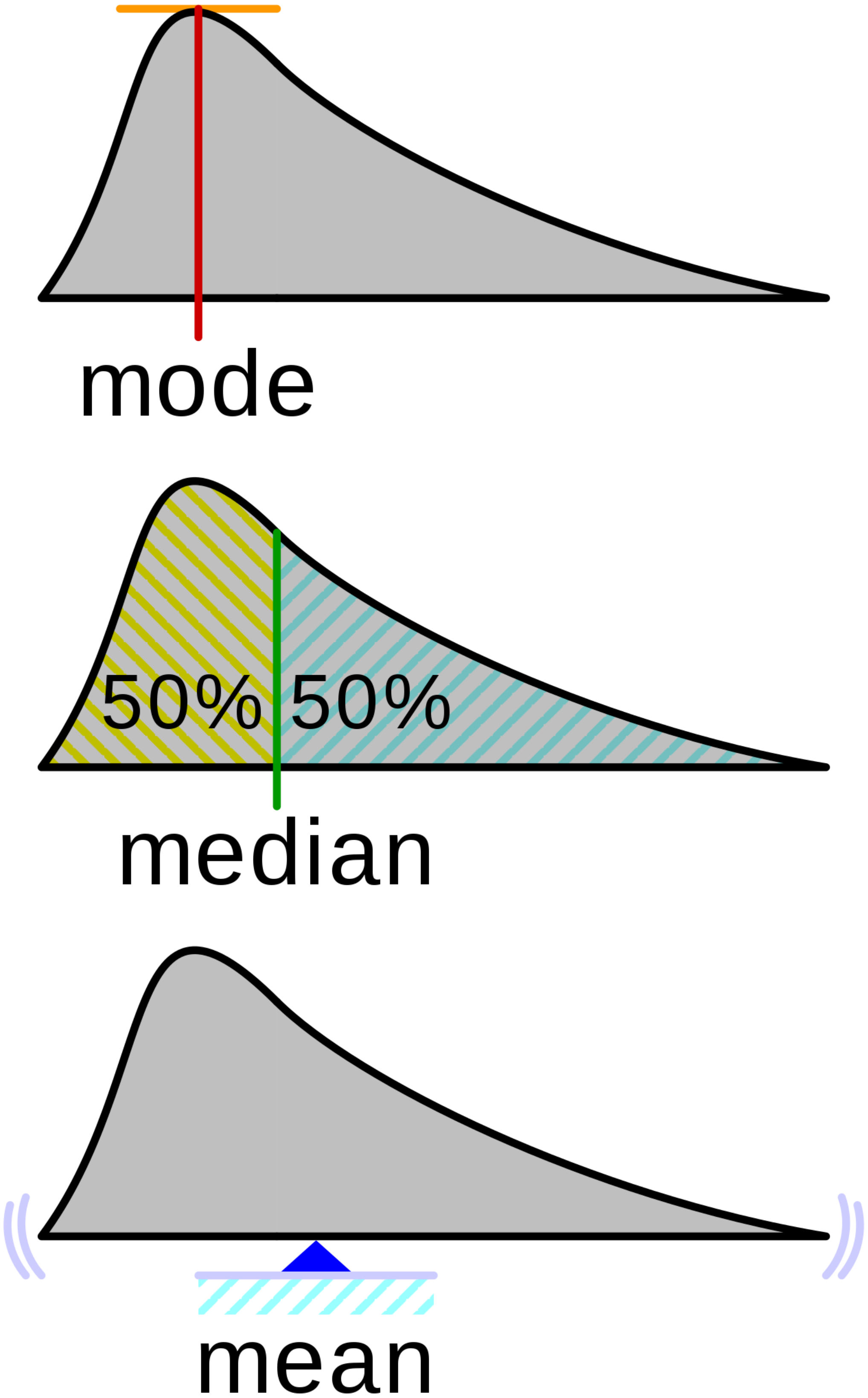
Central tendency[8]

Central tendency is the tendency for values of a variable to gather around the middle of the distribution.

Mean is the arithmetic average of all the values in the distribution. $\sum \frac{x}{n}$ where x is the values the variable can take and n is the set size.

Median is the middle value when all the values in the distribution are ordered.

Mode is the value that occurs most frequently in the data.



[8] By Cmglee - Own work, CC BY-SA 3.0

Dispersion[9]

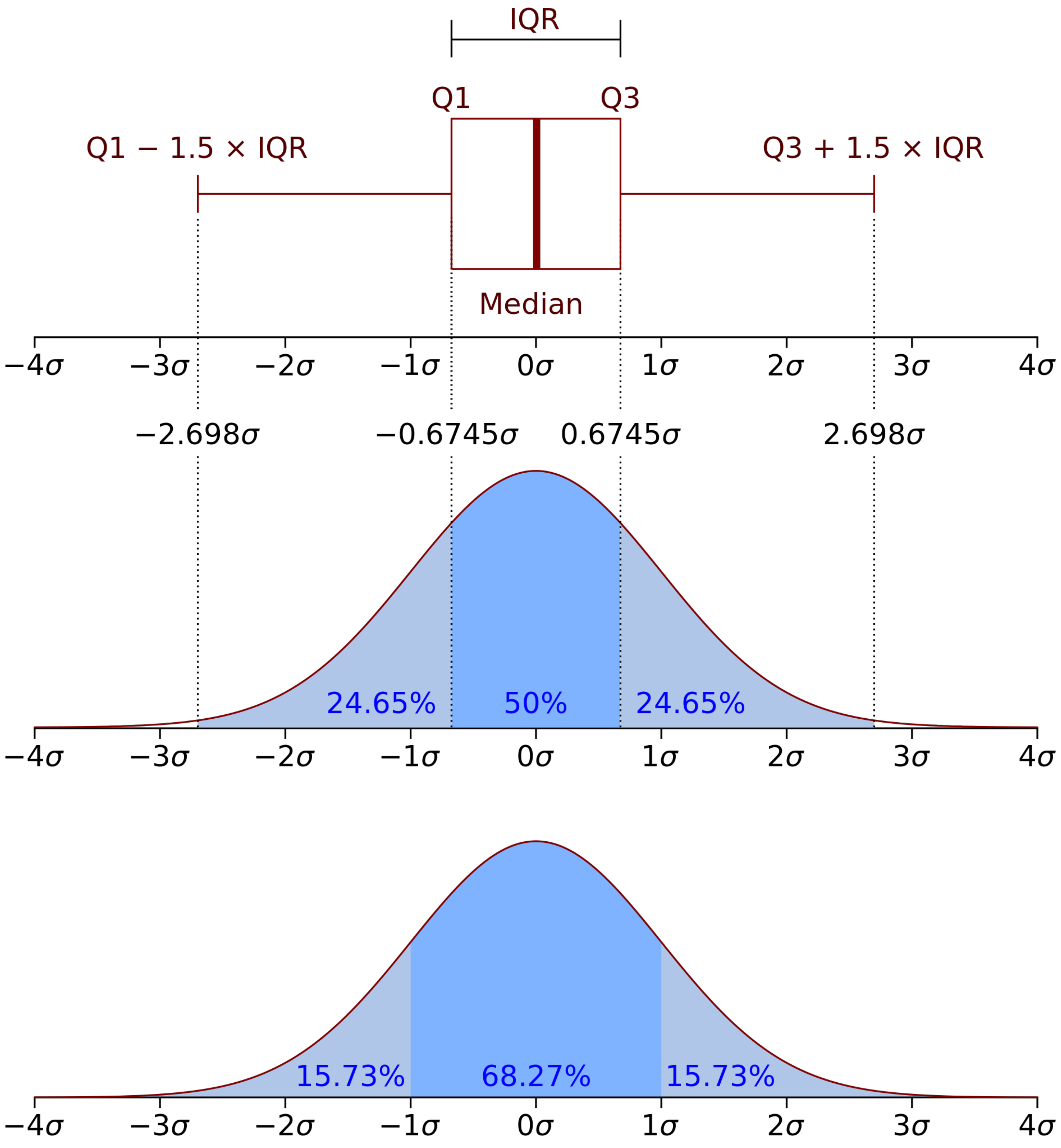
Dispersion captures the *spread* and *shape* of the data distribution.

Range is the difference between the smallest and the largest values.

Quartiles break the distribution to four equally sized parts.

Variance is the squared deviation of the variable from its mean.

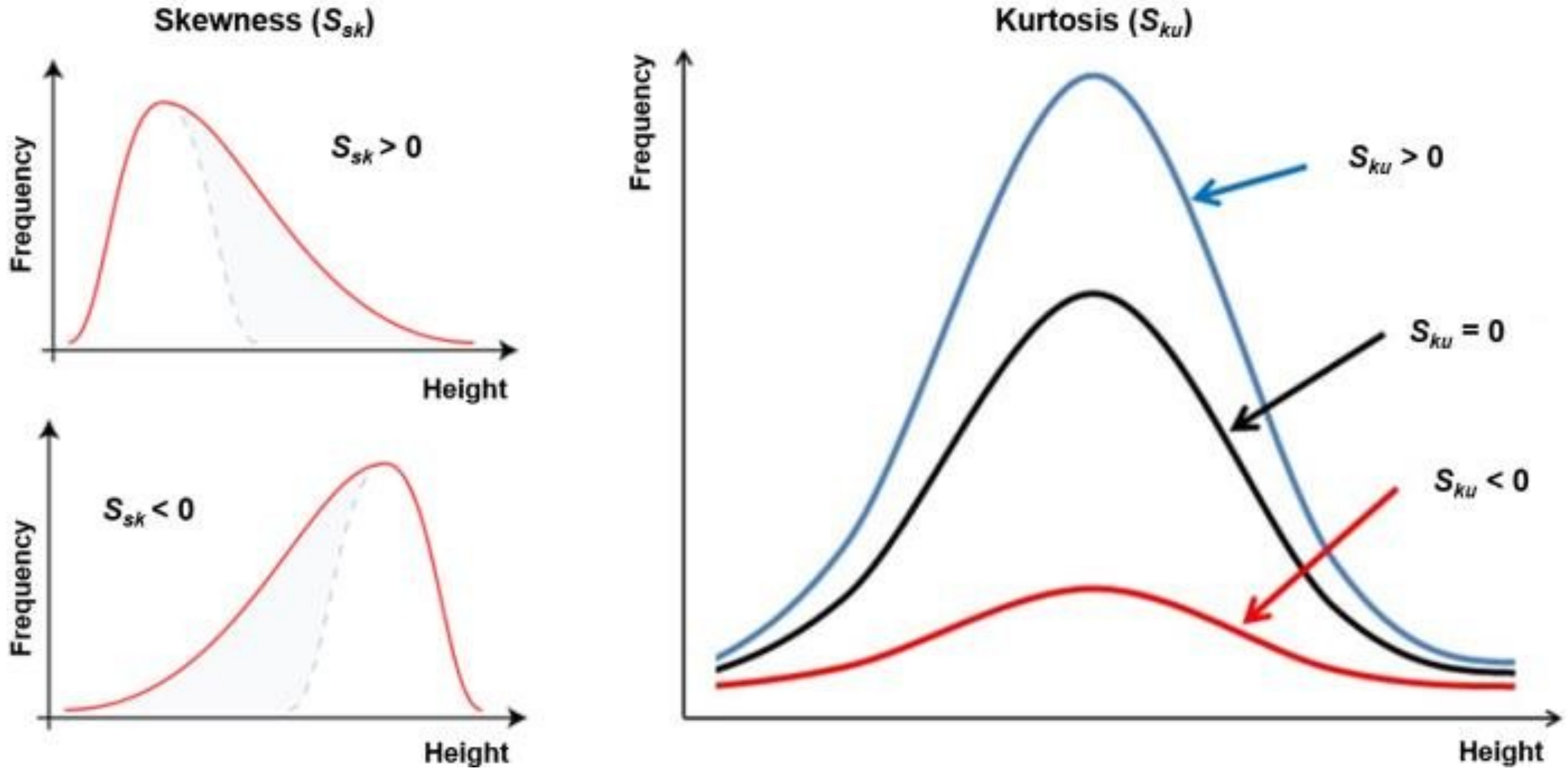
Standard deviation measures the amount of variation or dispersion in values.



[9] By Jhguch at en.wikipedia, CC BY-SA 2.5

Kurtosis measures how much the values gather in the peak or the tail of the distribution: *leptokurtic*, *mesokurtic*, *platykurtic*.

Skewness measures of asymmetry in the distribution: *positive*, *negative*.[10]

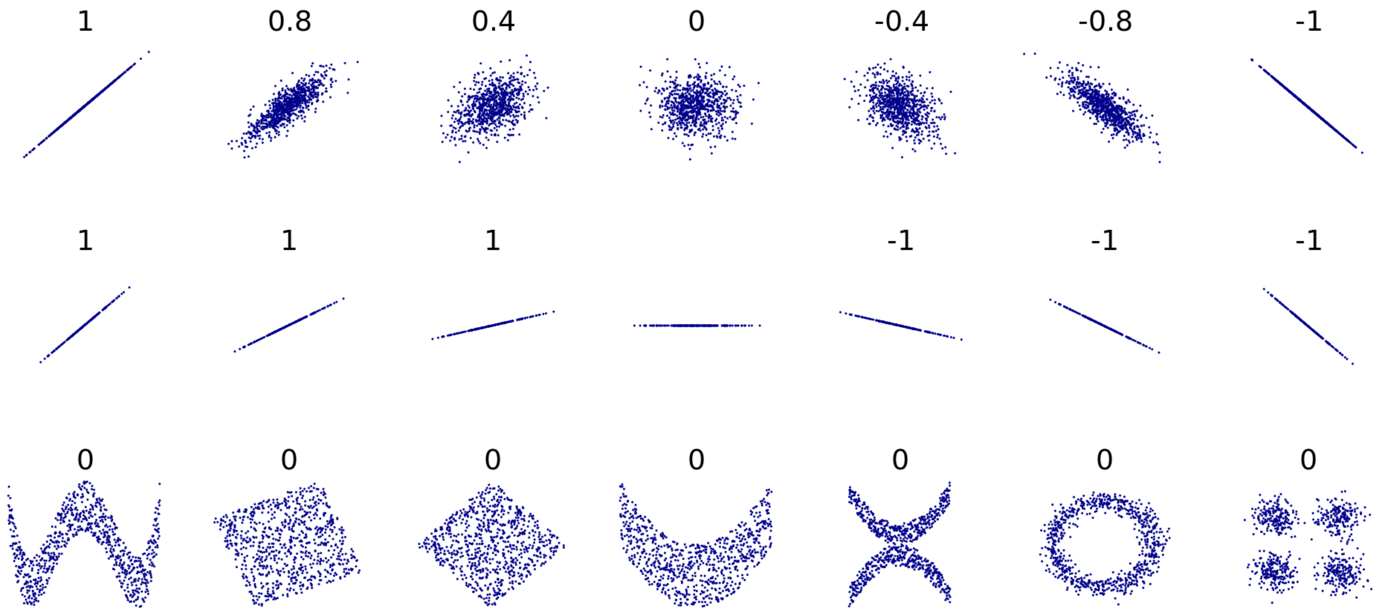


[10] Image source: Attila Bonyár

What do we look at in bivariate/multivariate analysis?[11]

Correlation and **covariance** measure the extent to which two variables are linearly related. Correlation is the normalized form of covariance.

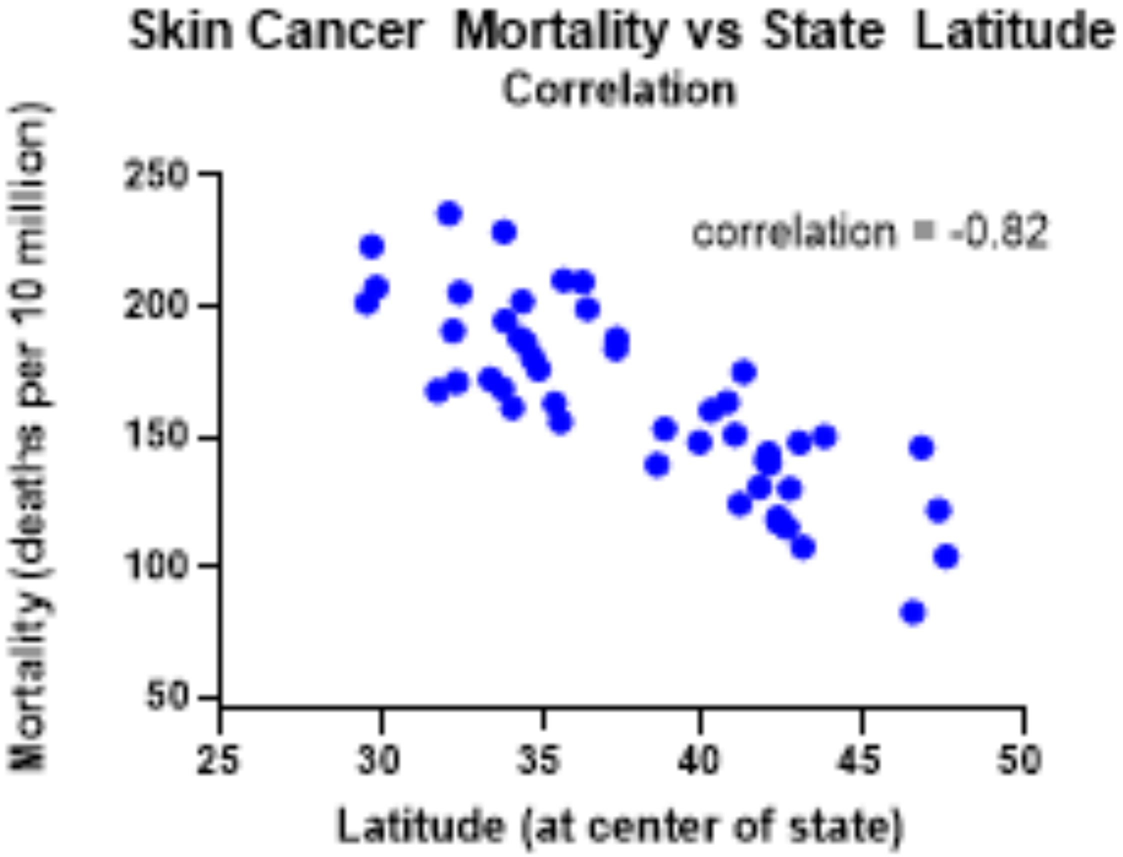
Pearson's r (when the variables are continuous) and *Spearman's* ρ (when one/both are discrete) measure correlation.

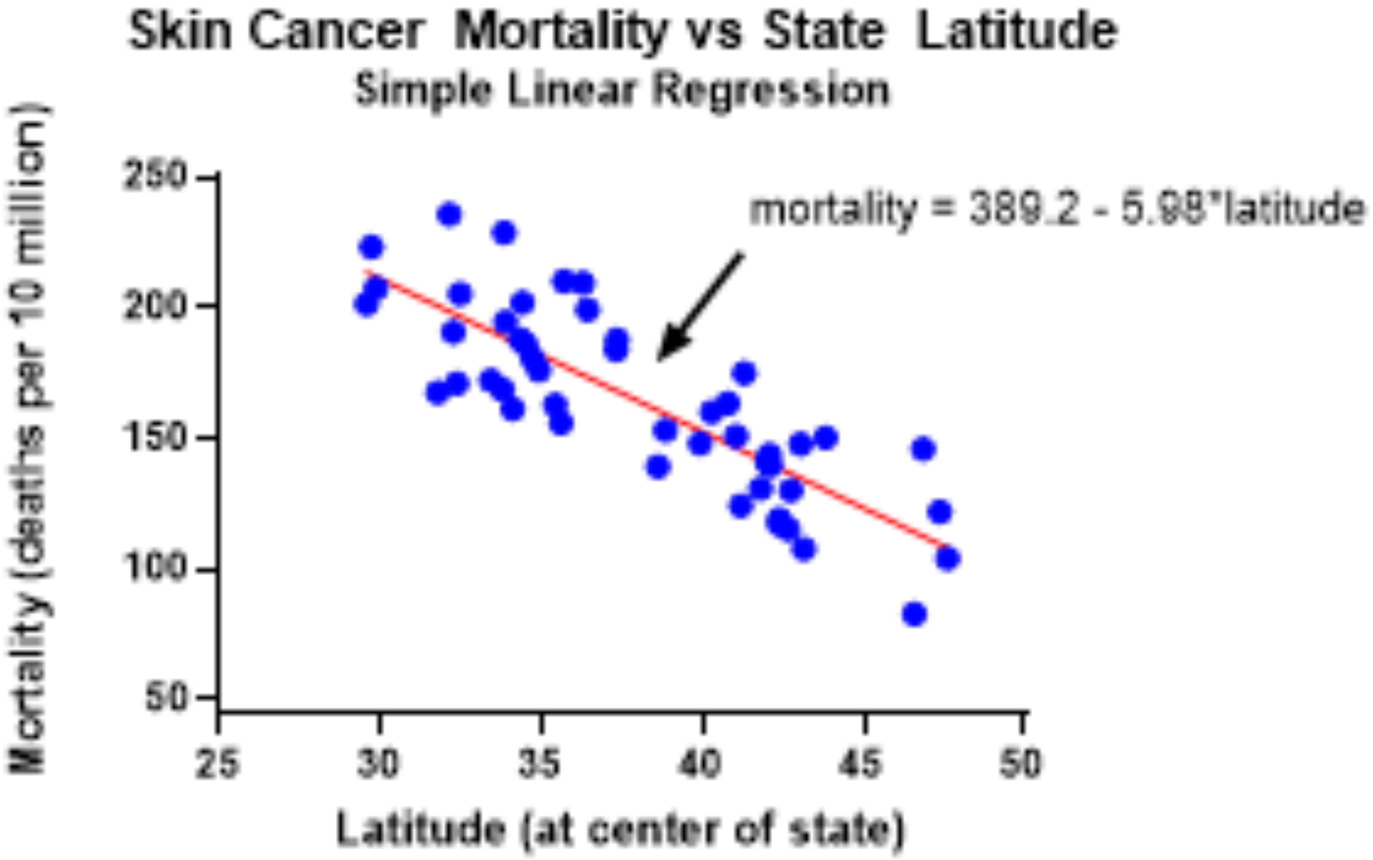


[11] By DenisBoigelot, Imagecreator, CC0

Is correlation descriptive or inferential?[12]

Can be used for *descriptive* or *inferential* statistics.





[12] Image source

How is correlation calculated?

We calculate what is called a **correlation coefficient**.

For a population:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

For a sample:

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

How do we interpret the correlation coefficient?

Correlation coefficient is a measure of relation between two variables that ranges -1 to 1.

» -1 represents a negative correlation

» 0 represents lack of correlation

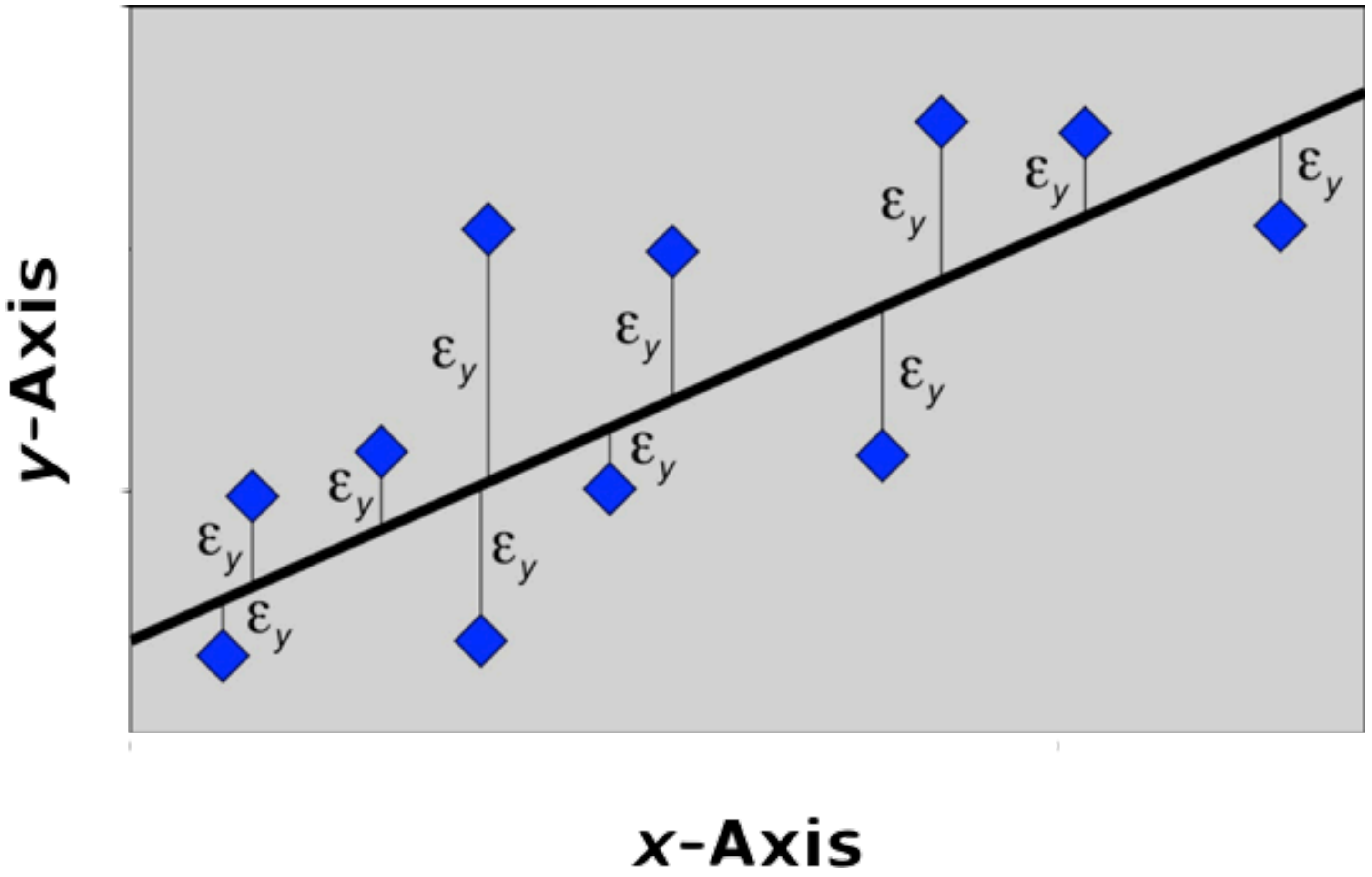
» 1 represents a positive correlation

Simple linear correlation: *Pearson's* r calculates the extent to which the variables are *proportional* or *linearly related* to each other.

r denotes the percent of variation in one variable that is related to the variation in the other. E.g., $r = .70$ ↠ 49% of the variance is related.

The proportion can be summarized by a simple line (*regression* or *least squares* line), determined such that the sum of the squared distances of all the data points from the line is the lowest possible.

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_1 X_i + \epsilon_i$$



How do we do descriptive statistics in R?

- `describe(var)` calculates all descriptive statistics
- `hist(var)` plots data histogram
- `plot(density(var))` plots the density plot
- `boxplot(var)` plots out a box plot
- `plot(var1, var2)` plots out a scatterplot
- `cov(vars)` calculates correlations among all vars