Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 1
Introduction and Descriptive Data Analysis

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Course Objectives

- introduce to the most widely used quantitative data analysis methods;
- explain the data analysis methods using real data, discuss challenges that may arise in real-life research;
- explain how to use data analysis tools in the most effective way to perform the research tasks;
- explain how to organize individual research project and present it’s results to the audience.

Main topics

- Introduction to data analysis;
- Descriptive data analysis;
- Investigating relationships;
- Regression analysis;
- Factor analysis;
- Cluster analysis;
- Panel data analysis;
- Time series analysis.

Grading System

$$\boxed{\begin{gathered}\boldsymbol{Final\ Grade = 0,15 * Control\ Work\ 1 + 0,15\ * Control\ Work\ 2 +} \\ \boldsymbol{0,3 * Exam + 0,2\ * Research\ Project + 0,2 * Regular\ Tasks}\end{gathered}}$$

If more than 50% of the control tasks are completed, the student may be exempted from taking the exam. In this case, the formula for calculating the final grade is as follows:

$$Final\ Grade = 0{,}2 * Control\ Work\ 1 + 0{,}2\ * Control\ Work\ 2 + 0{,}3\ * Research\ Project + 0{,}3 * Regular\ Tasks$$

Data Analysis Tools

Python 3

- Anaconda distribution contains a Python interpreter and a Jupyter Notebook instance. We will use Jupyter Notebook interactive development environment.
- Google Colab (interactive cloud environment for working with code)

Datasets Sources

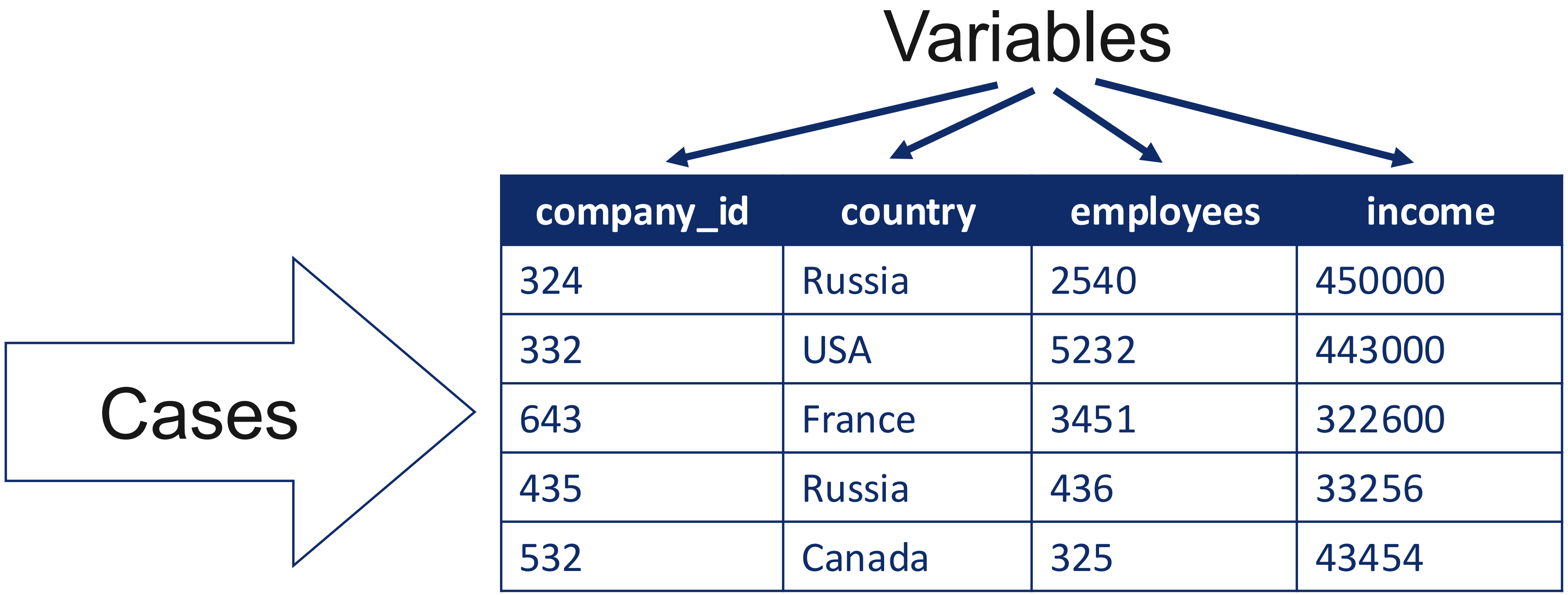
- Kaggle
- Google Dataset Search
- Harvard Dataverse
- Eurobarometer
- Росстат
- Портал открытых данных правительства Москвы

Data Structures

- Cross-sectional data
- Time series data
- Panel data

Cross-Sectional Data

Cross-sectional data, or a cross section of a study population is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time.



Variables

Cases

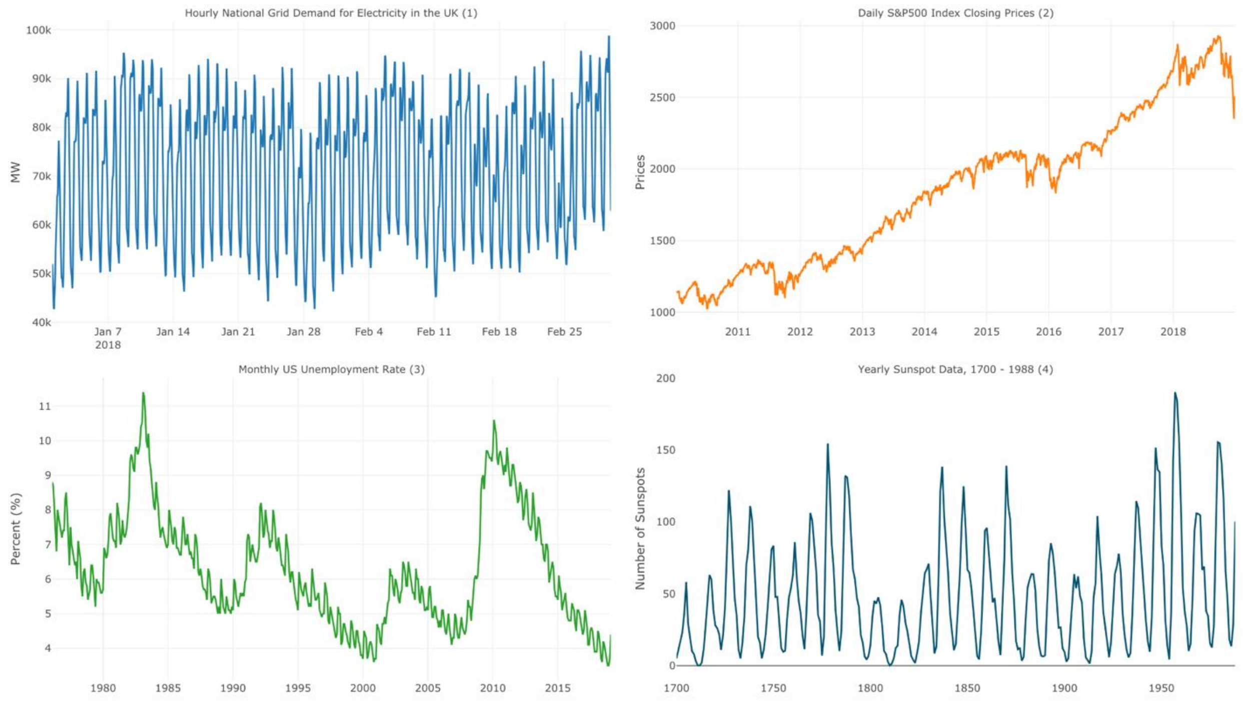
company_id	country	employees	income
324	Russia	2540	450000
332	USA	5232	443000
643	France	3451	322600
435	Russia	436	33256
532	Canada	325	43454

Cross Section: data structure

- **Rows** are **cases**. Each row represents a case or an observation.
- **Columns** are **variables**. Each column represents a variable or characteristic being measured.
- **Cells** contain **values**. Each cell contains a single value of a variable for a case.

Time Series Data

Time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time.



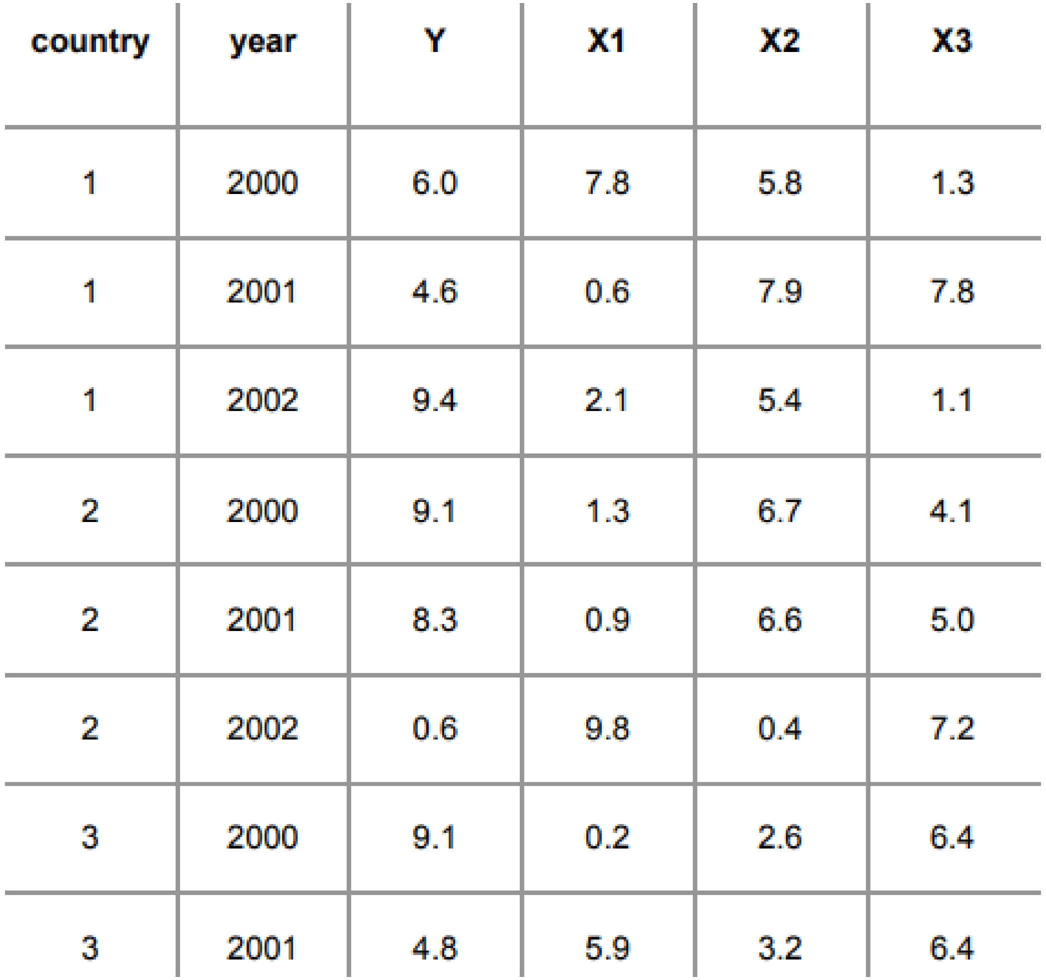
Panel Data

Panel Data or **Longitudinal Data** are multi-dimensional data involving measurements over time. Panel data contain observations of multiple phenomena obtained over multiple time periods for the same objects.

Panel data have three dimensions:

1. variables,
2. cases,
3. time.

Example of panel data



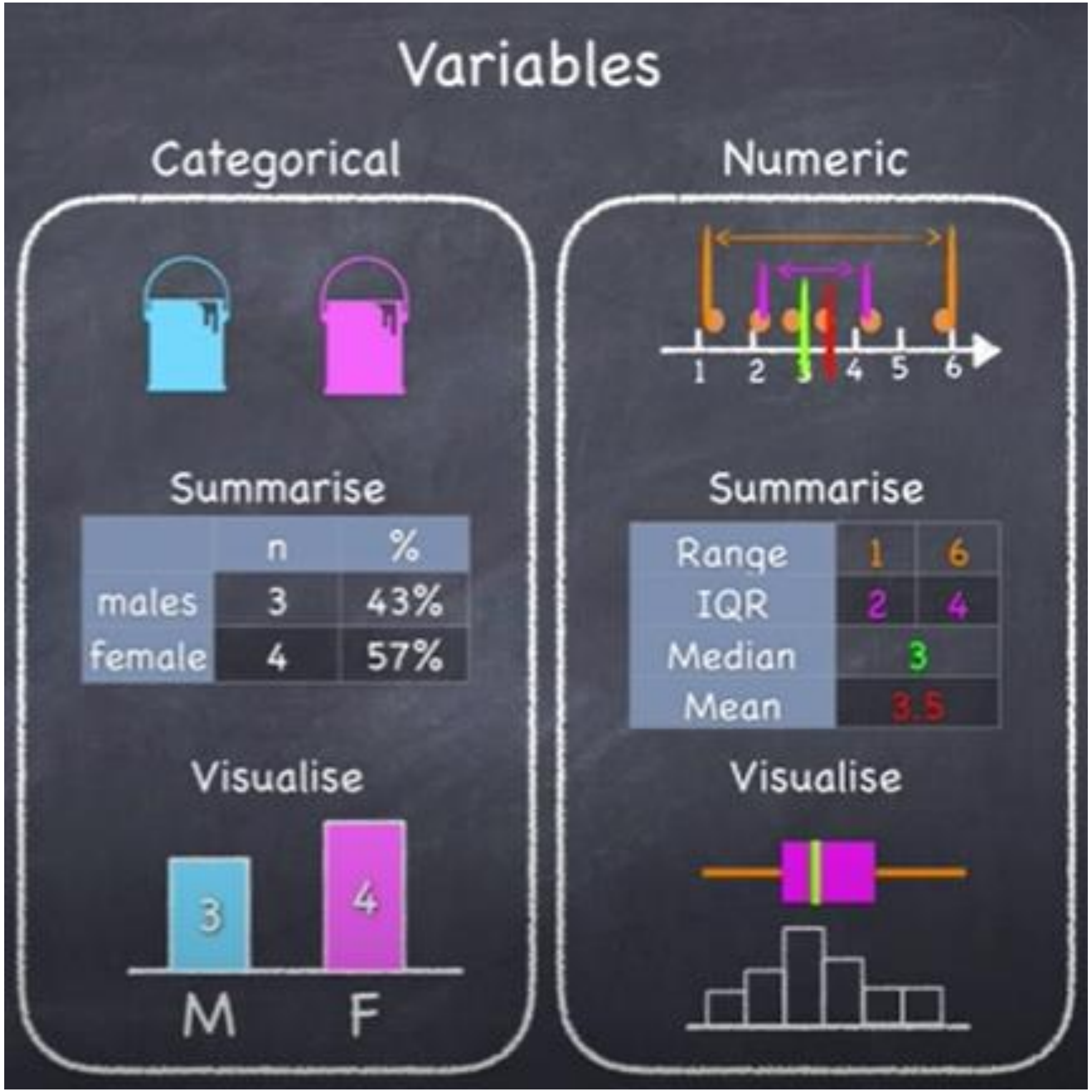
country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4

Continuous and Categorical variables

- **Continuous (or Scale) Variables** – data values are numeric values on an interval or ratio scale (e.g., age, income).

- **Categorical Variables** – variables that have values which fall into two or more discrete categories. E.g. male or female, employment category, country of origin

Two types of Categorical variables: Ordinal & Nominal



Variables

Categorical

Summarise

	n	%
males	3	43%
female	4	57%

Visualise

Numeric

Summarise

Range	1	6
IQR	2	4
Median	3	
Mean	3.5	

Visualise

Source: https://www.youtube.com/watch?v=I10q6fjPxJ0&t=441s

Nominal Scale

Data values represent categories with no intrinsic order, sequence of categories is arbitrary – ordering has no meaning in and of itself:

- country of origin: Wales, Scotland, Germany…
- car manufacturers: Kia, Tesla, Mazda
- job category: manager, it-specialist, cleaner
- company department: production, purchasing, HR, R&D

Nominal variables can be either string (alphanumeric) or numeric values that represent distinct categories (e.g., 1=Male, 2=Female).

Ordinal Scale

Values fall within discrete but ordered categories, the sequence of categories has meaning.

- Education categories: 1 = primary, 2 = secondary, 3 = college, 4 = university undergraduate, 5 = university postgraduate masters, 6 = university postgraduate PhD
- Wealth level: 1 = very poor, 2 = poor, 3 = good, 4 = very good
- Satisfaction level (Likert scale): 1 = very satisfied,
  2 = somewhat satisfied, 3 = neither satisfied nor dissatisfied,
  4 = somewhat dissatisfied, 5 = very dissatisfied

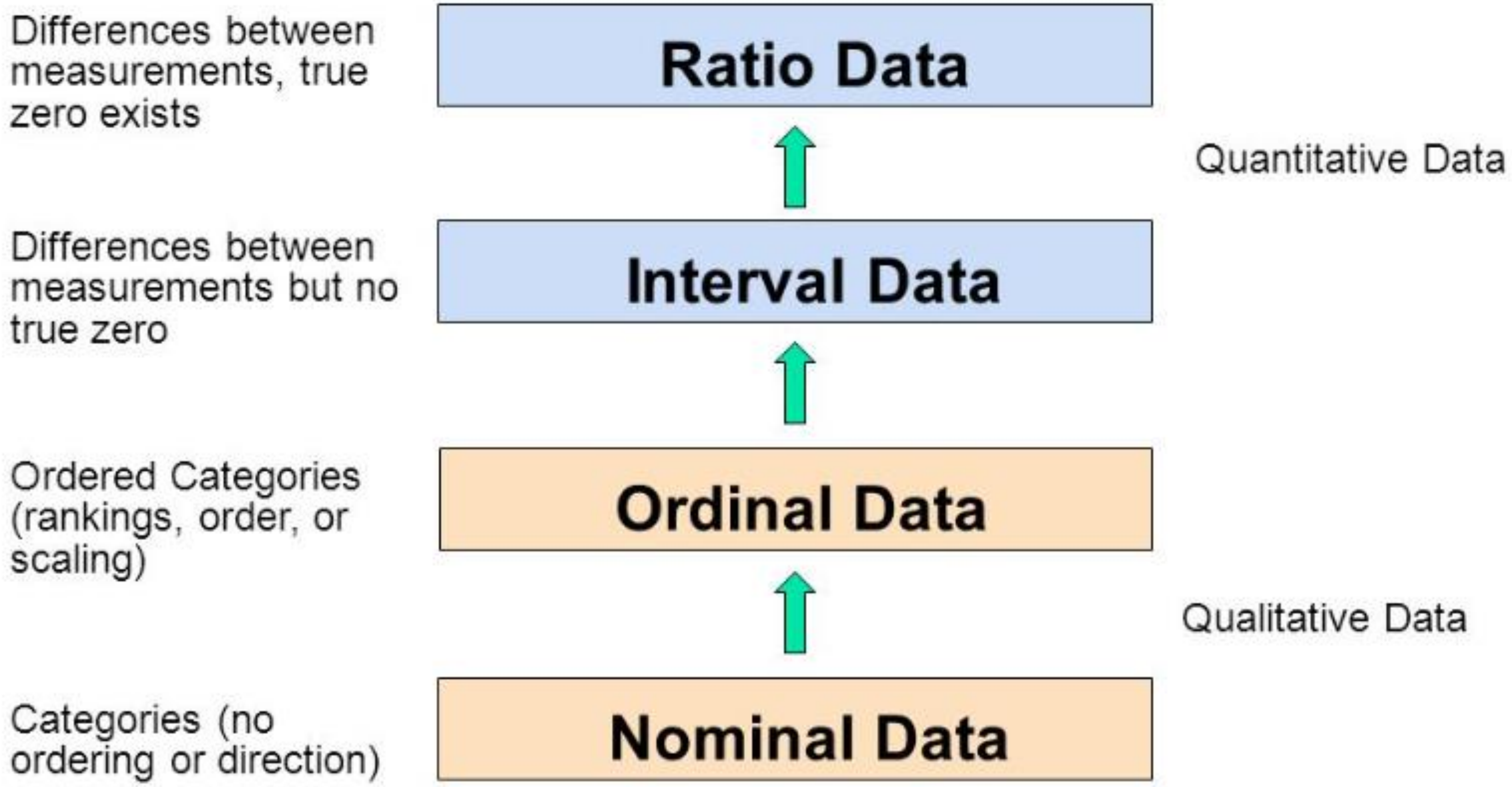
Interval Scale

Interval Scale is defined as a numerical scale where we can express and compare numerically the differences between the values. Interval scales hold no true zero and can represent values below zero. For example, you can measure temperature below 0 degrees Celsius, such as -10 degrees. For example, sea water temperature in Celsius in the morning is 18 degrees, in the evening is 24 degrees, i.e. in the evening it’s 5 degrees higher, but it cannot be said that it is 1.33 times higher.

Ratio Scale

Ratio variables never fall below zero. Height and weight measure from 0 and above, but never fall below it. A ratio scale has the same properties as interval scales. You can use it to add, subtract, or count measurements. Ratio scales differ by having a character of origin, which is the starting or zero-point of the scale.

Measurement Scales



Differences between measurements, true zero exists — **Ratio Data**

Differences between measurements but no true zero — **Interval Data**

Ordered Categories (rankings, order, or scaling) — **Ordinal Data**

Categories (no ordering or direction) — **Nominal Data**

Quantitative Data

Qualitative Data

Dichotomous coding

Select the languages you know:
1) English
2) Russian
3) French
4) Italian
5) German

Answers:
Dave: English, German
Ann: Russian
Peter: English, French, Italian, German
Oliver: Russian, German

Name	Lan_en	Lan_ru	Lan_fr	Lan_it	Lan_ge
Dave	1	0	0	0	1
Ann	0	1	0	0	0
Peter	1	0	1	1	1
Oliver	0	1	0	0	1

Categorical coding

Select the languages you know:
(you can choose any number of answers
/ select up to 4 options)

1) English
2) Russian
3) French
4) Italian
5) German

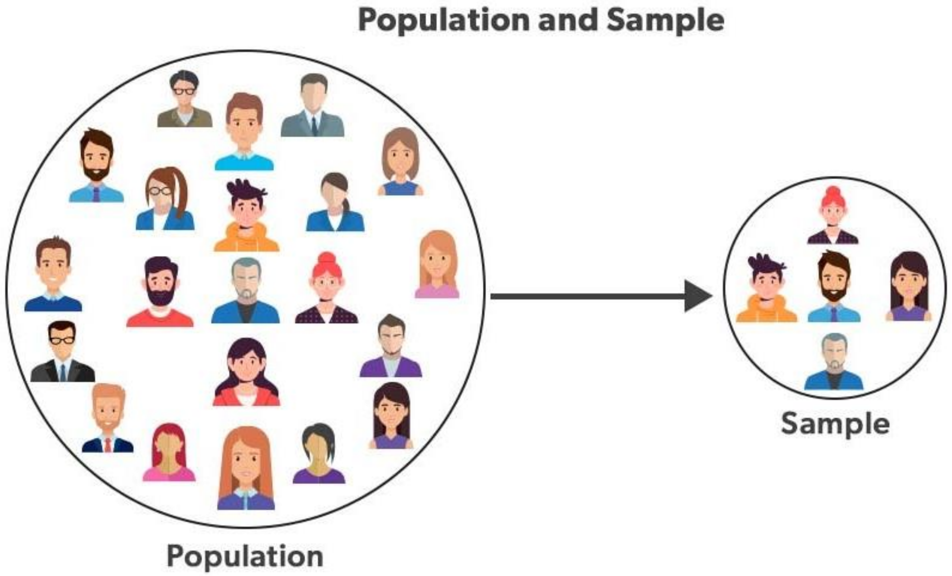
Answers:
Dave: English, German
Ann: Russian
Peter: English, French, Italian, German
Oliver: Russian, German

Name	Lan_1	Lan_2	Lan_3	Lan_4
Dave	1	5		
Ann	2			
Peter	1	3	4	5
Oliver	2	5		

Population and Sample

Population is the entire group that the researcher wants to draw conclusions about.

Sample is the specific group that the researcher will collect data from. The size of the sample is always less than the total size of the population.



Reasons for Sampling

- It’s sometimes impossible to study the whole population due to its size or inaccessibility.
- It’s easier and more efficient to collect data from a sample.
- Less resources are required to realize the research.

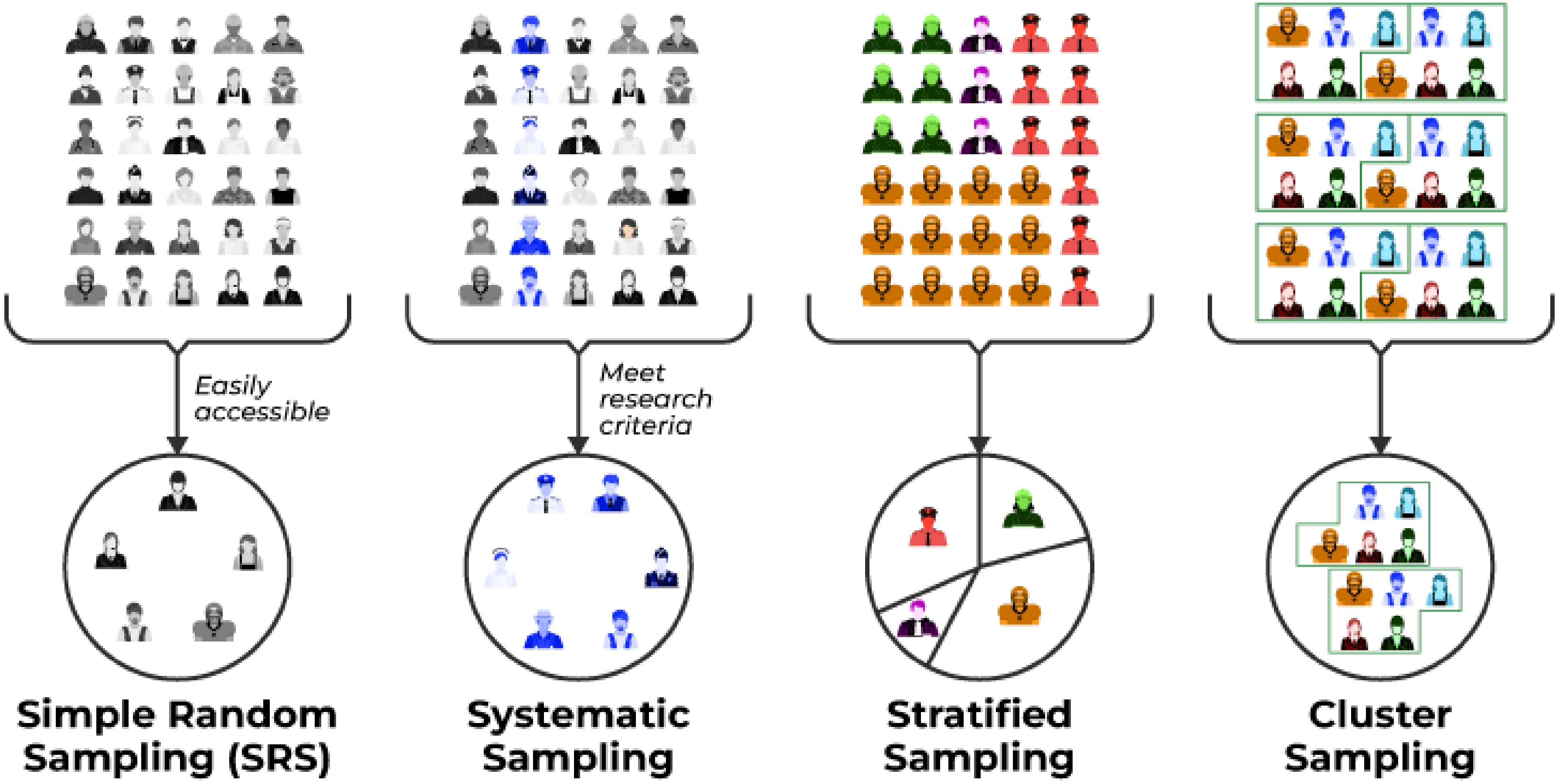
Selecting a Sample

Sample should be randomly selected and representative of the population. Probability sampling assures that every person or component of a population has an equal chance to be included in a sample that is randomly selected from the entire population. Probability sampling reduces the risk of sampling bias and enhances both internal and external validity.

Non-probability sampling could be used if the research is less concerned with generalizability. Non-probability samples are chosen for specific criteria, they may be more convenient or cheaper to access.

Unrepresentative Sample
Unrepresentative Sample
Representative Sample

Types of Probability Sampling



Types of Probability Sampling

1. Simple Random Sampling

Every item has an equal opportunity of being picked without their being any impact on others following them.

2. Systematic Sampling

Involves selecting every nth item from a population, for example, if you have a list of students and you select every 10th student.

3. Stratified Sampling

It divides the population into subgroups or strata based on certain characteristics (e.g., age, gender), and then samples are randomly selected from each stratum. This method ensures representation from each subgroup.

4. Cluster Sampling

It divides the population into clusters or groups, often based on geographical regions. A random sample of clusters is selected, and then all items within the chosen clusters are included in the sample.

Population parameter and sample statistic

Parameter is a measure that describes the whole population. **Statistic** is a measure that describes the sample.

We can use estimation or hypothesis testing to estimate how likely it is that a sample statistic differs from the population parameter.

Descriptive Analysis

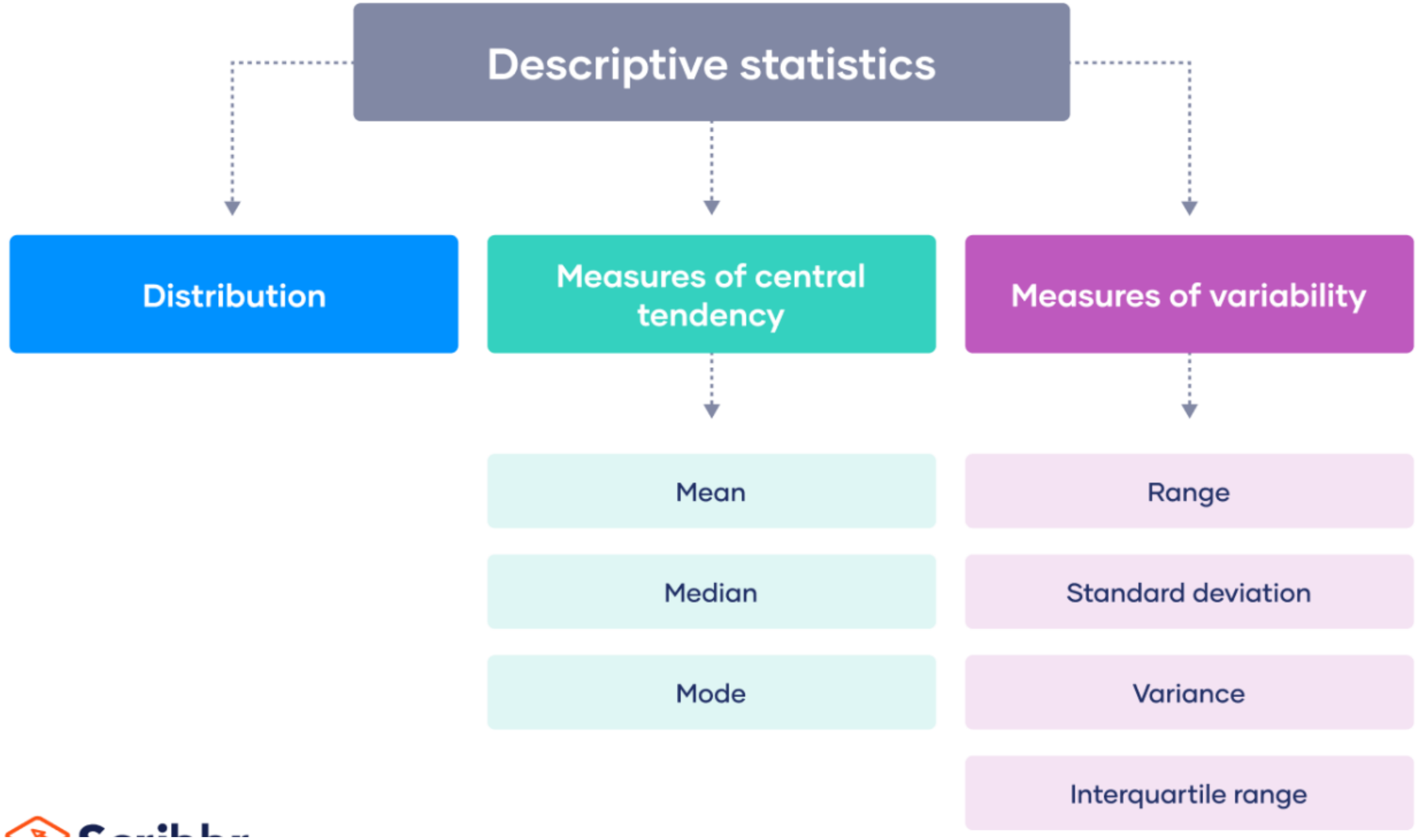
Descriptive data analysis quantitatively describes the main features of the collected data, provides simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of quantitative analysis of data. It also can be used for data screening, especially when a considerable number of cases are being studied.

Univariate analysis

Univariate analysis is aimed at the examination across cases of a single variable, focusing on three characteristics:

- distribution;
- central tendency;
- variability.

It is common to compute all three for each study variable.



Descriptive statistics

Distribution

Measures of central tendency

Measures of variability

Mean

Median

Mode

Range

Standard deviation

Variance

Interquartile range

Distribution

The distribution is a summary of the frequency of individual or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of cases who had that value.

For instance, computing the distribution of gender in the study population means computing the percentages that are male and female. The gender variable has only two values making it possible and meaningful to list each one.

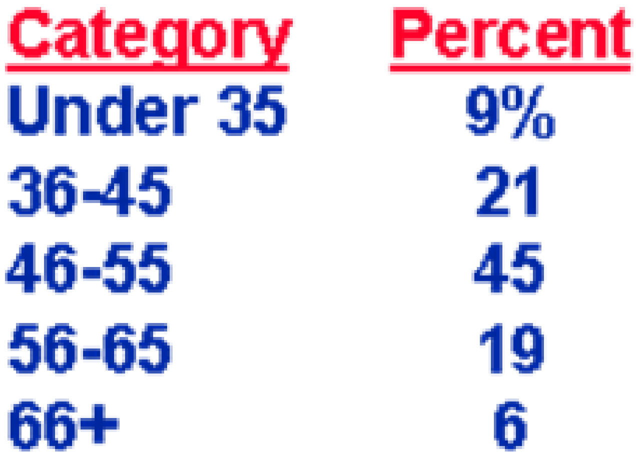
Male – 35 – 60,3%

Female – 23 – 39,7%

If it’s a scale variable, for example, income that has many possible values it’s advised to group the scores using ranges of values in order to reduce the number of categories and create ordinal variable. For instance, we might group incomes into ranges of 0-10 000, 10 001-30 000, etc.

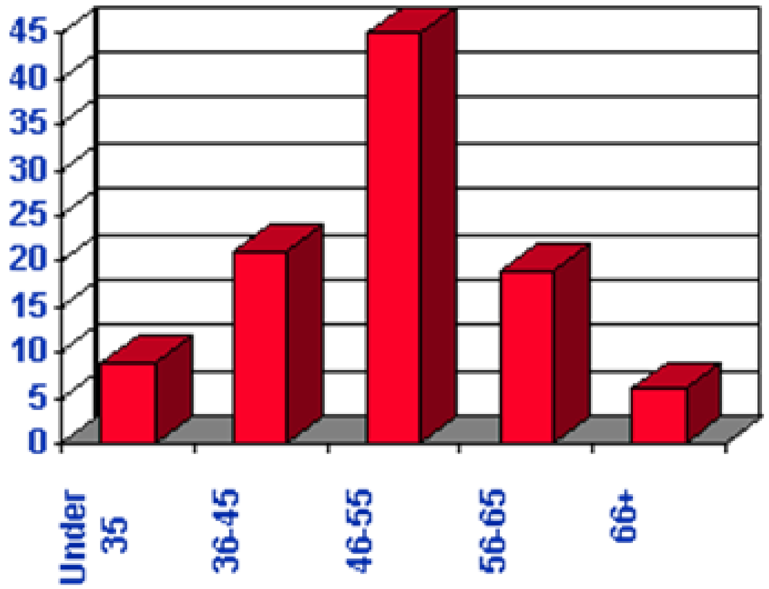
Frequency Analysis

Is most often used in the analysis of the values of categorical variables.



Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Frequency distribution table



Frequency distribution bar chart

Central Tendency

The central tendency of a distribution locates the "center" of a distribution of values. Three major types of estimates of central tendency: mean, median, mode.

Measurement Scale	Appropriate central tendency estimates
Nominal	mode
Ordinal	mode, median
Scale	mode, median, mean

Mode

The mode is the most frequently occurring value in the set. In some distributions there are multiple modal values. These are called multi-modal distributions.

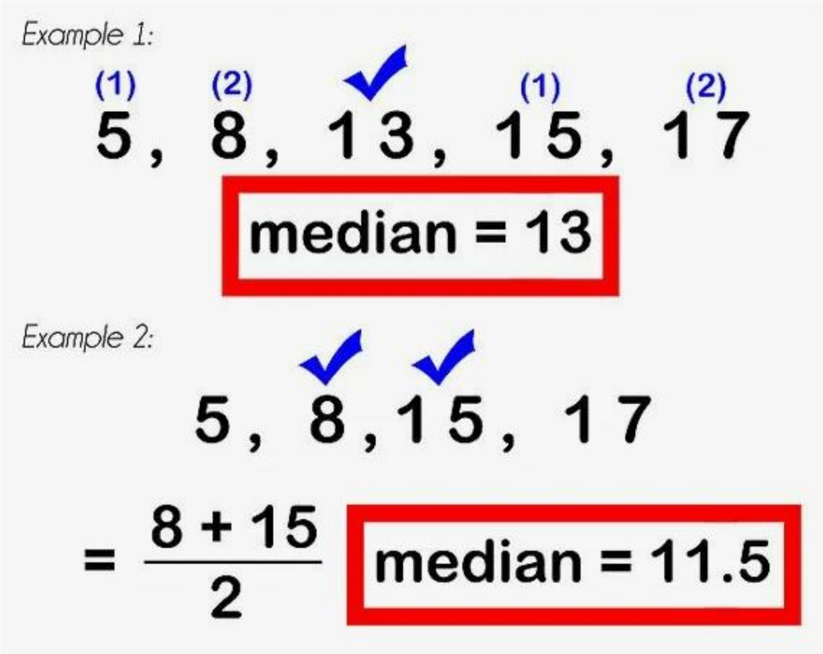
Mean

The Mean is the most commonly used method of describing central tendency. To compute the mean, take the sum of the values and divide by the count. For example, the mean grade in the course is determined by summing all the scores and dividing by the number of students taking the exam.

$$\bar{X} = \frac{\sum x_i}{n}$$

Median (second quartile / 50th percentile)

The median is a number that falls in the middle of a group. This is done by ordering the numbers from smallest to largest and locating the one that falls in the middle.



Example 1:

5, 8, 13, 15, 17

$\boxed{\text{median} = 13}$

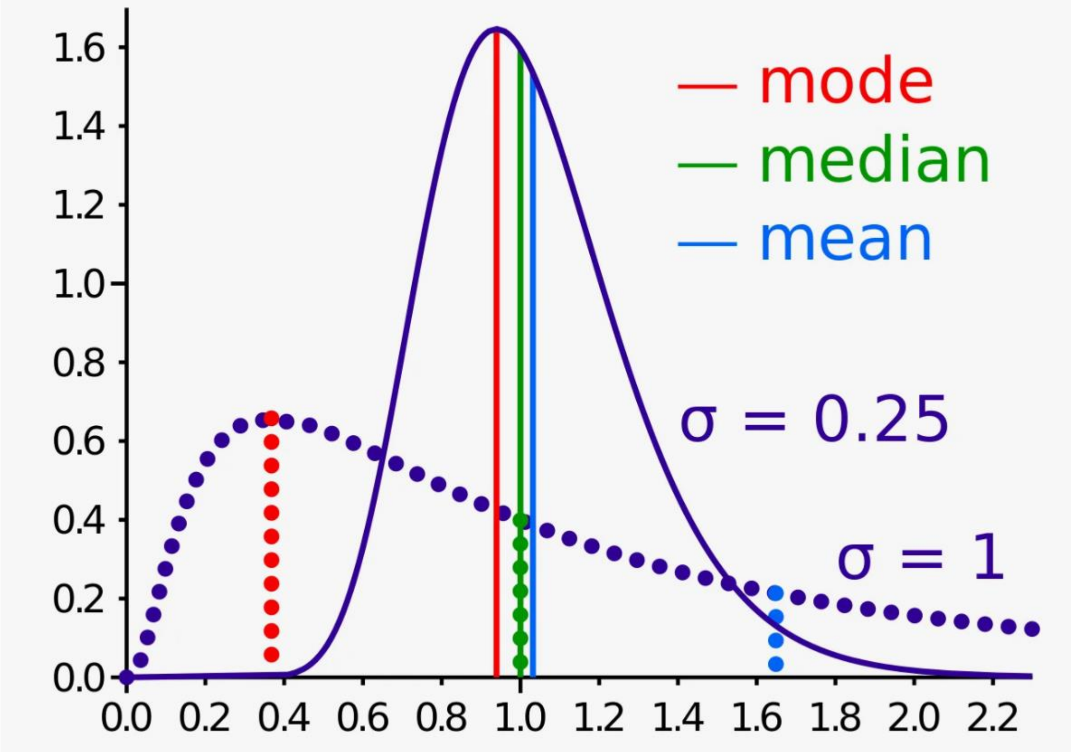
Example 2:

5, 8, 15, 17

$$= \frac{8+15}{2} \quad \boxed{\text{median} = 11.5}$$

Mean or Median

When the mean and the median have values that are very close, it’s recommended to use the mean to describe the central tendency. When they are different, it’s better to use median.



Dispersion

Dispersion is the spread of values around the central tendency. There are two common measures of dispersion, the **range** and the **standard deviation**. The range is calculated as the difference between the highest and the lowest value of the variable. The standard deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (the single outlier value could stands apart from the rest of the values).

Variance

Variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

$$\sigma^2 = \frac{\sum(x - \bar{X})^2}{n}$$

Standard Deviation

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is expressed in the same unit of measurement as the data, which isn't the case with the variance.

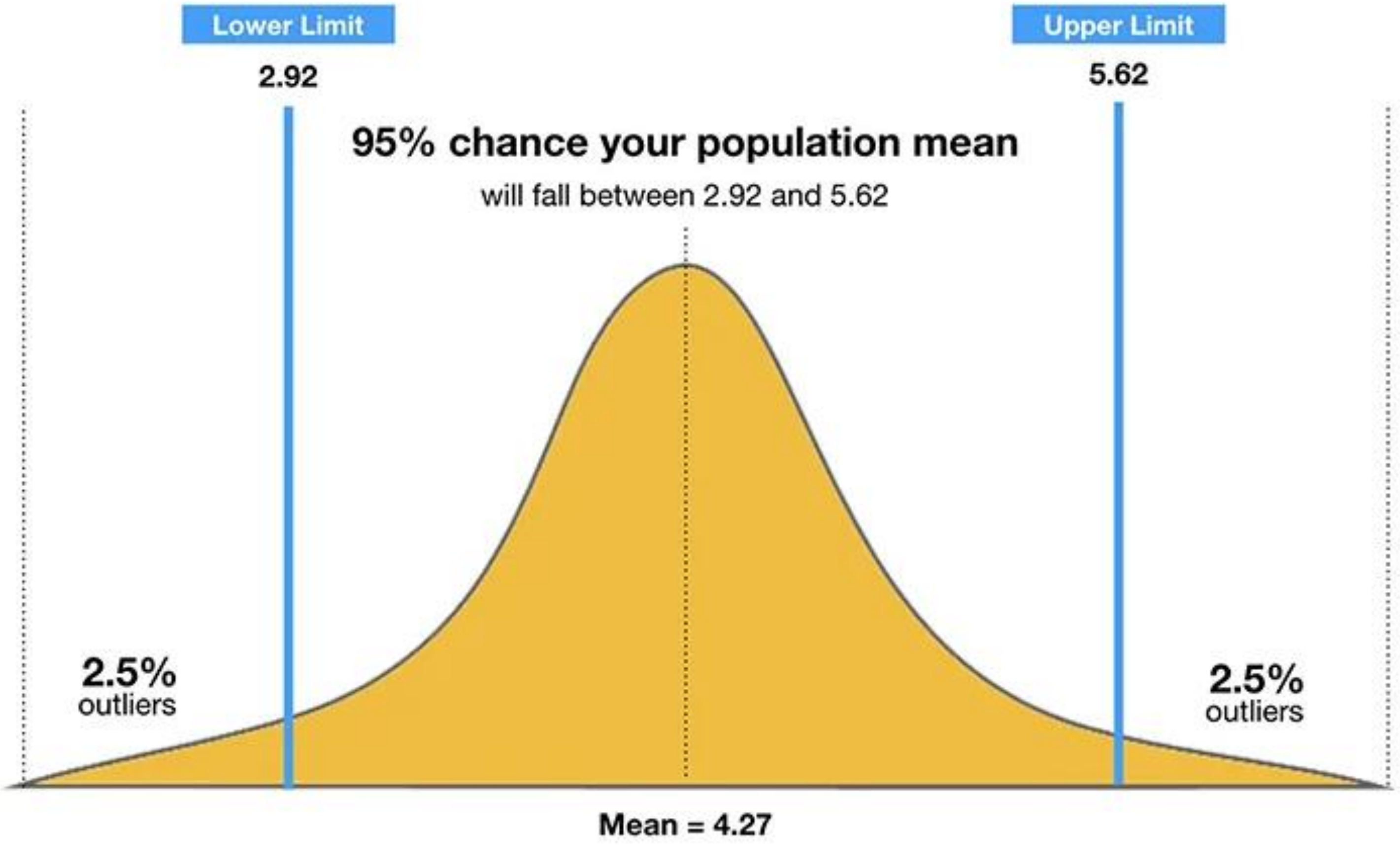
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x - \bar{X})^2}{n}}$$

Standard Error of the Mean

A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. Could be interpreted as a standard deviation of the means in different samples in relation to the mean of the population. If it's high, we can conclude that mean of our sample is not an ideal representation of reality.

$$m = \frac{\sigma}{\sqrt{n}}$$

Confidence Interval



Confidence Interval

The confidence interval (CI) is a range of values that's likely to include a population value with a certain degree of confidence. It is often expressed as a % whereby a population mean lies between an upper and lower interval. Therefore, a confidence interval is simply a way to measure how well your sample represents the population you are studying. The 95% confidence interval is a range of values that you can be 95% confident contains the true mean of the population.

- There is a 95% probability that the mean of the general population will fall into the interval from $\bar{X} - (1{,}96 * m)$ to $\bar{X} + (1{,}96 * m)$
- There is a 99% probability that the mean of the general population will fall into the interval from $\bar{X} - (2{,}58 * m)$ to $\bar{X} + (2{,}58 * m)$
- There is a 99,9% probability that the mean of the general population will fall into the interval from $\bar{X} - (3{,}29 * m)$ to $\bar{X} + (3{,}29 * m)$

Percentile Values

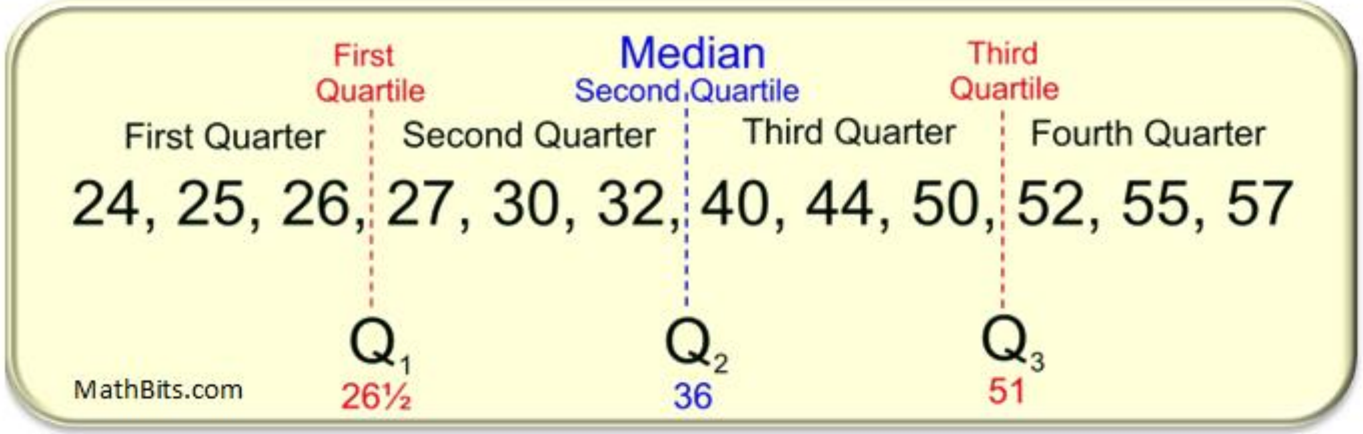
Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below. Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size.

You can also specify individual percentiles (for example, the 95th percentile, the value below which 95% of the observations fall).

Quartile

Quartile divides the number of data points into four parts, or quarters, of more-or-less equal size. The data must be ordered from smallest to largest to compute quartiles. The three main quartiles are as follows:

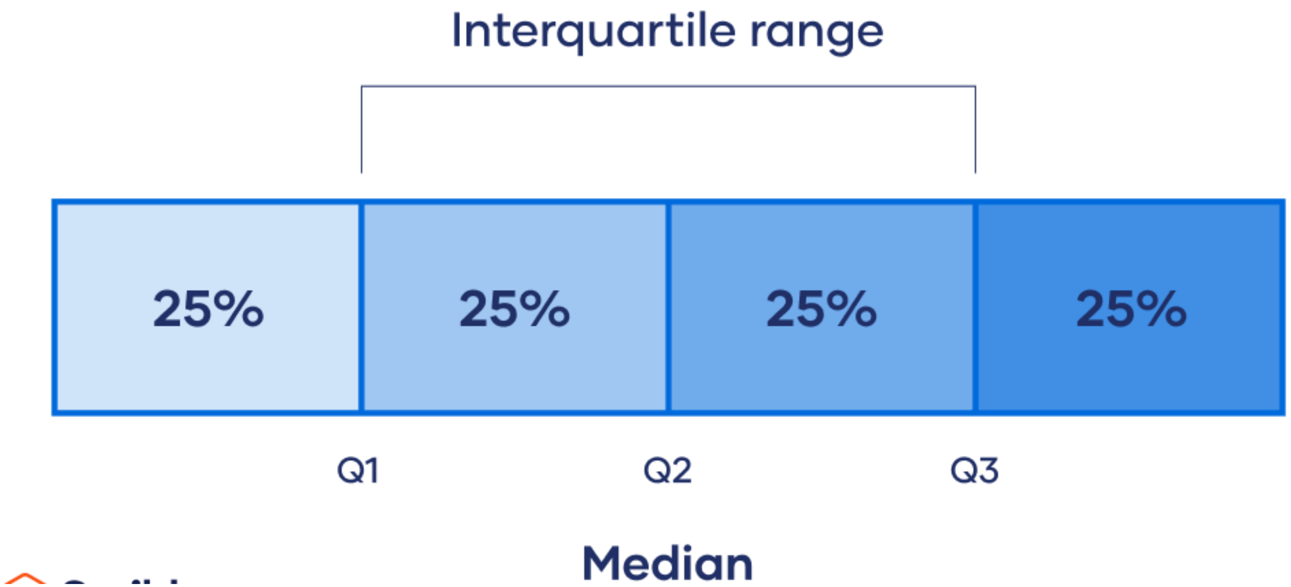
- The first quartile (Q1) is the 25th empirical percentile, as 25% of the data is below this point.
- The second quartile (Q2) is the median of a data set; thus 50% of the data lies below this point.
- The third quartile (Q3) is the 75th empirical percentile, as 75% of the data lies below this point.



Percentile-based data spread measures

- **Interquartile range** is the difference between the third and first quartiles.

$$\mathbf{IQR = Q3 - Q1}$$



Percentile-based data spread measures

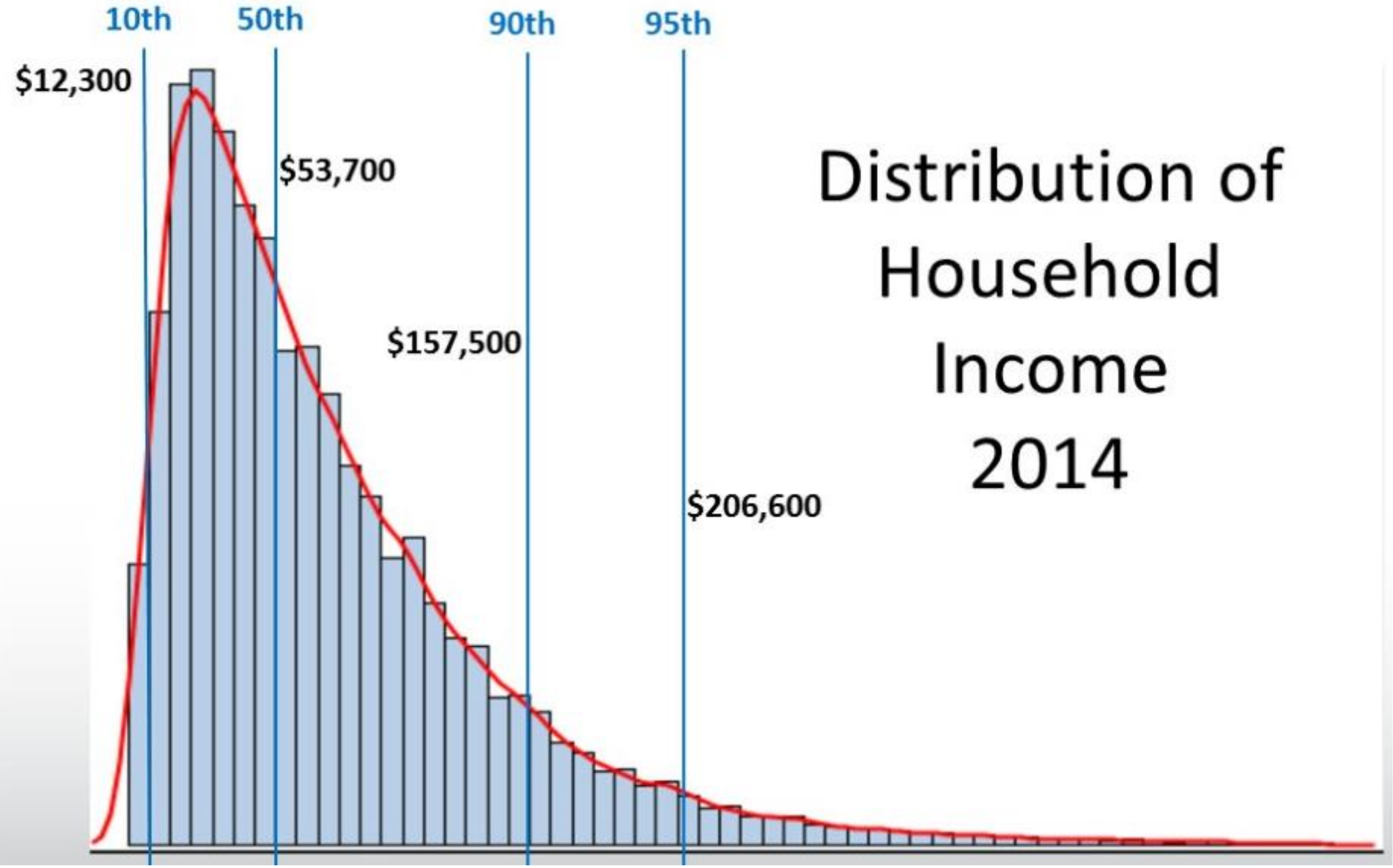
- **Quartile deviation** is half the difference between the upper and lower quartiles in a distribution (**IQR/2**). It is a measure of the spread through the middle half of a distribution. It can be useful because it is not influenced by extremely high or extremely low scores. Quartile Deviation is an ordinal statistic and is most often used in conjunction with the median.

$$\text{Quartile Deviation Formula} = \frac{Q_3 - Q_1}{2}$$

Percentile-based data spread measures

- **Decile ratio** is calculated as a ratio of the 10th decile to 1st decile. Income inter-decile ratios are used to evidence the disparities (or differences) between the richest and the poorest.

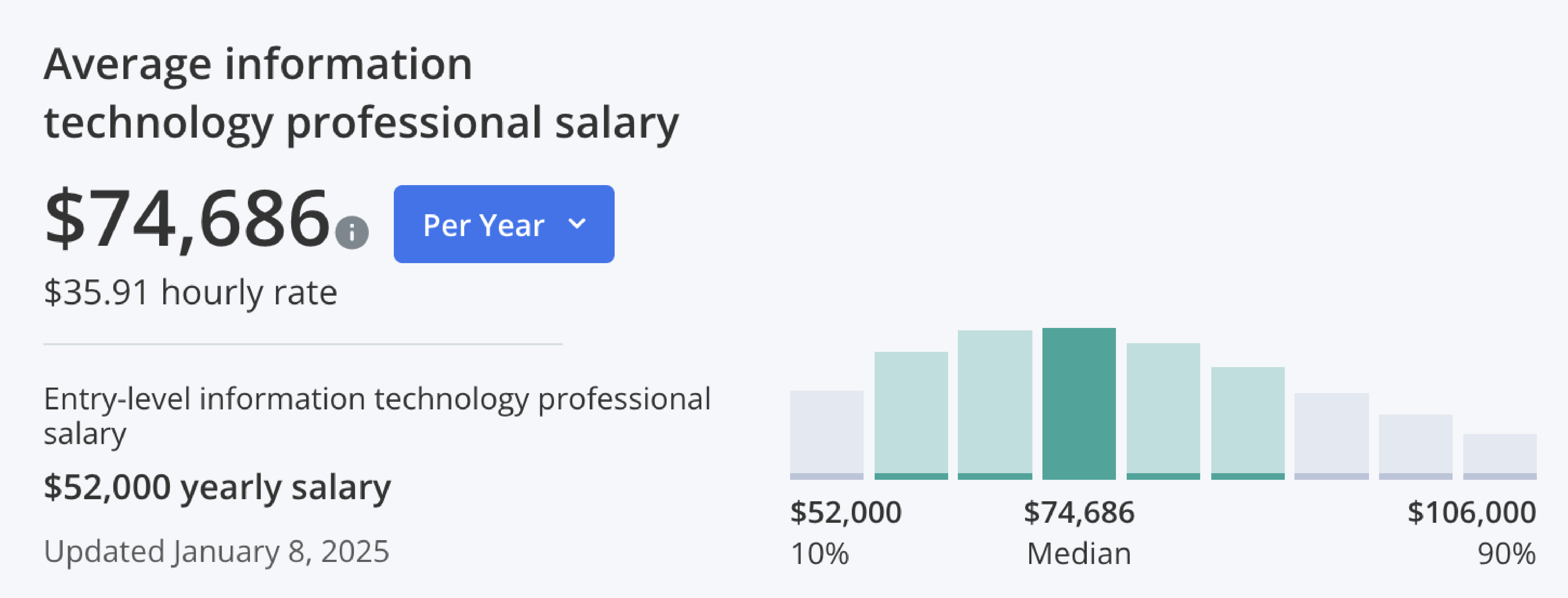
Decile Ratio: measuring inequality



Select Data Spread Measure

- If the **mean** is used to describe the central tendency, then the **standard deviation or the variance** should be used to describe the spread.

- If the **median** is used to describe the central tendency, then the **interquartile range or quartile deviation** should be used to describe the spread.



Average information technology professional salary

$74,686 ⓘ

Per Year ⌄

$35.91 hourly rate

Entry-level information technology professional salary

$52,000 yearly salary

Updated January 8, 2025

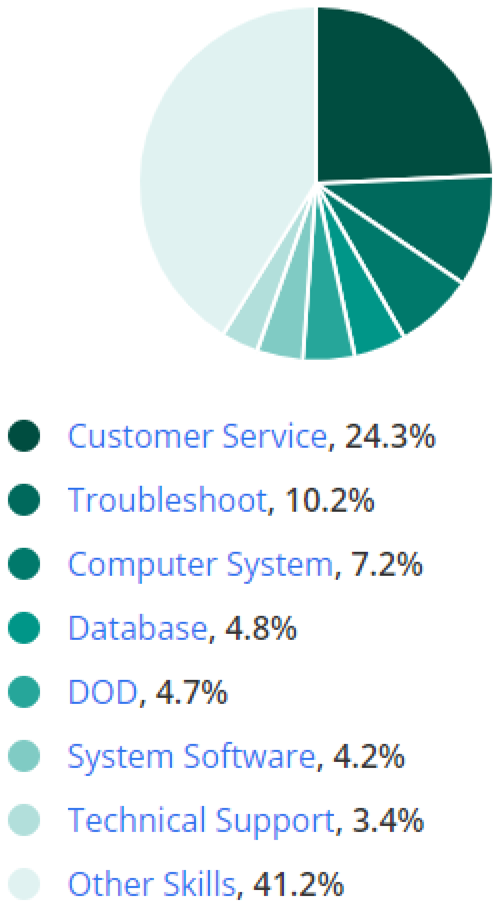
$52,000 10%

$74,686 Median

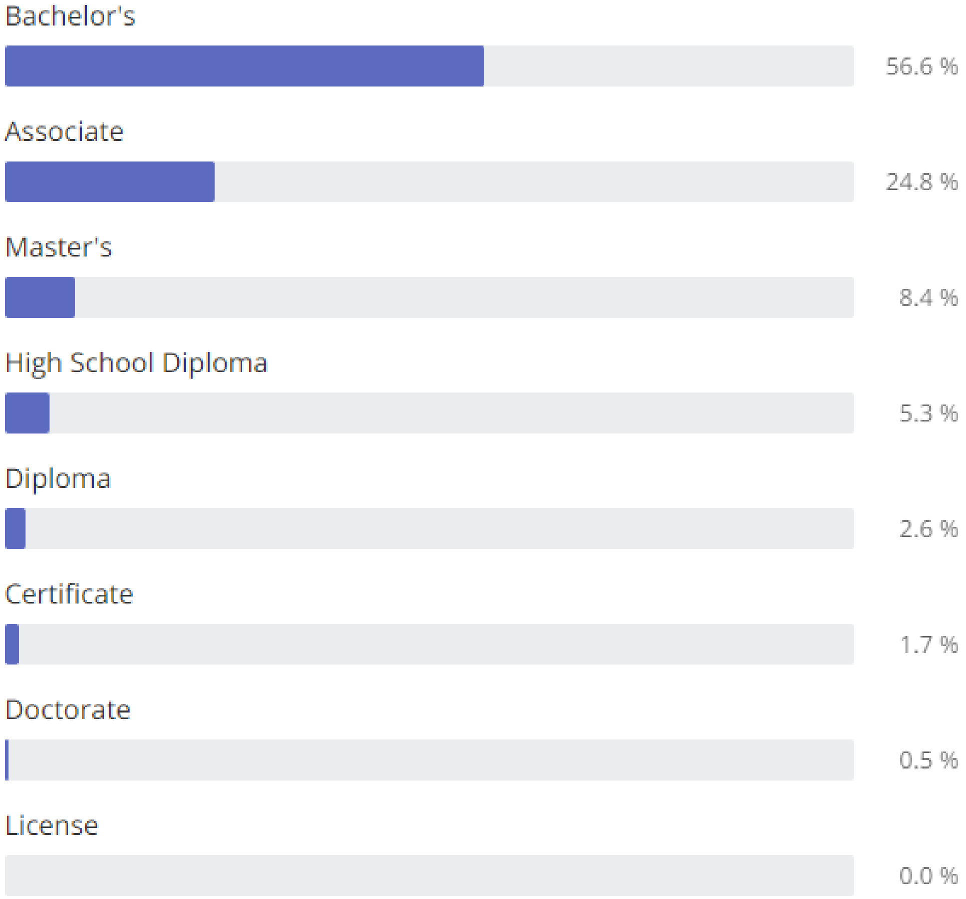
$106,000 90%

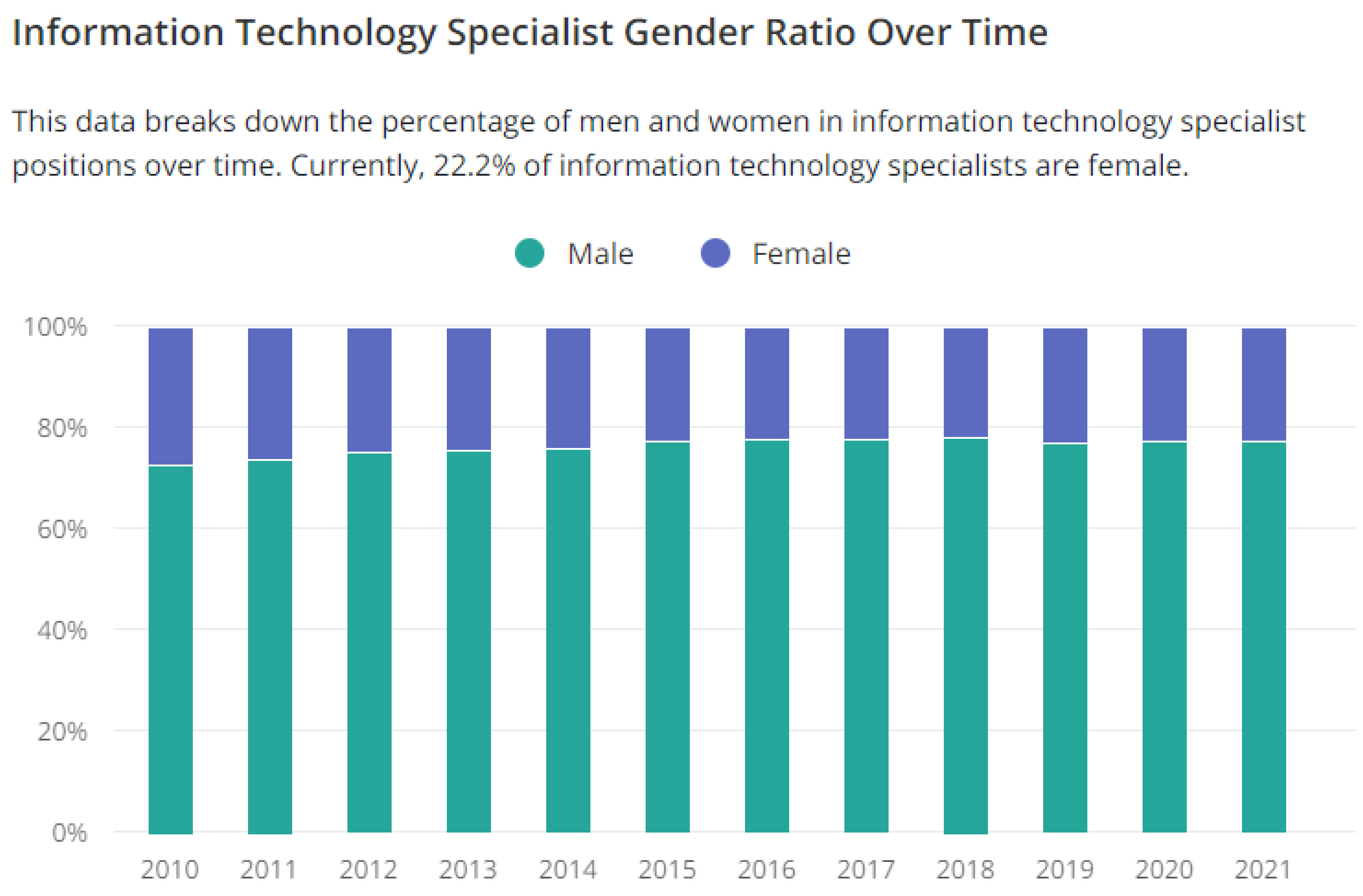
Source: www.zippia.com

The eight most common skills for information technology specialists in 2024 based on resume usage.



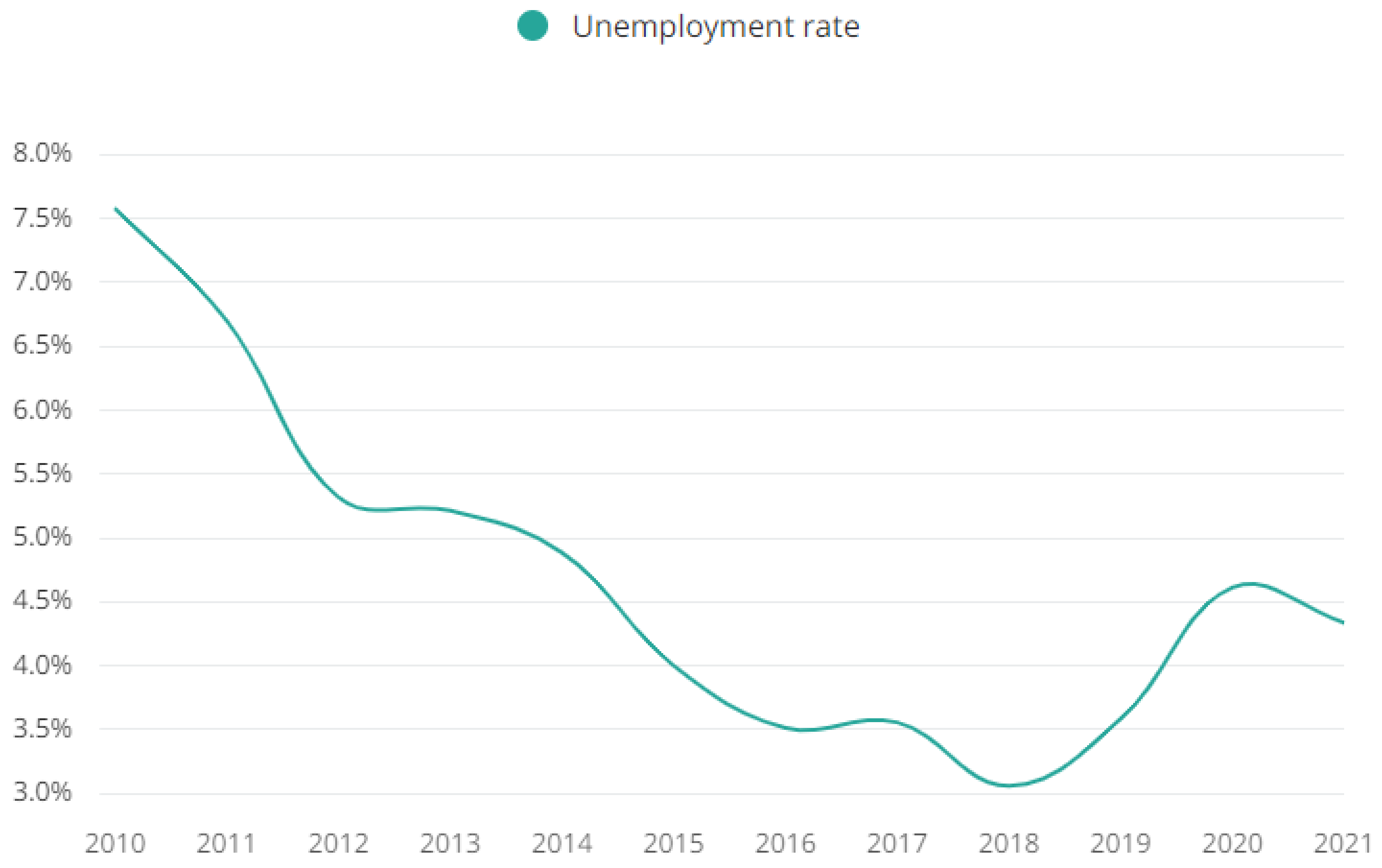
Most Common Information Technology Specialist Degrees



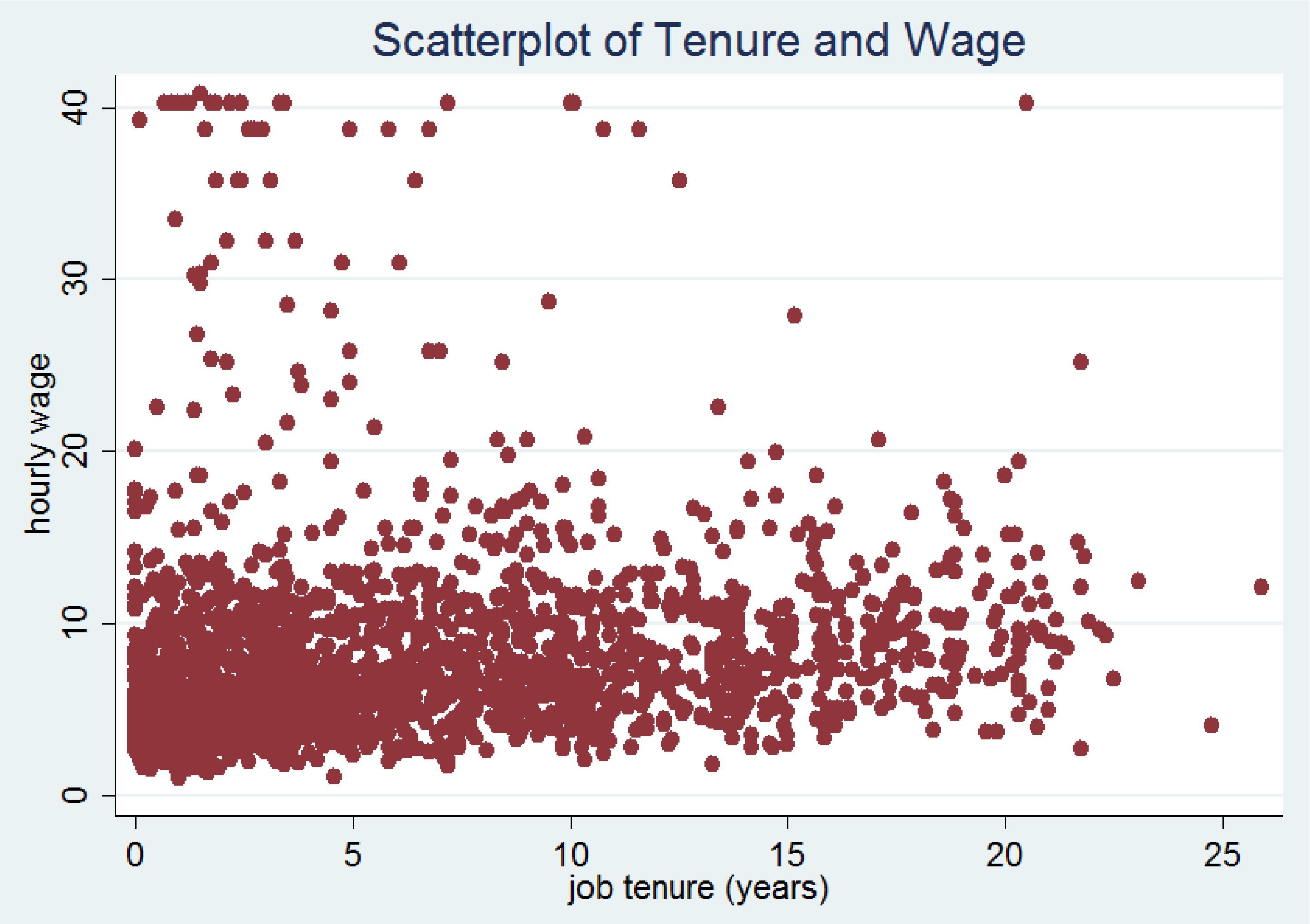


Information Technology Specialist Gender Ratio Over Time

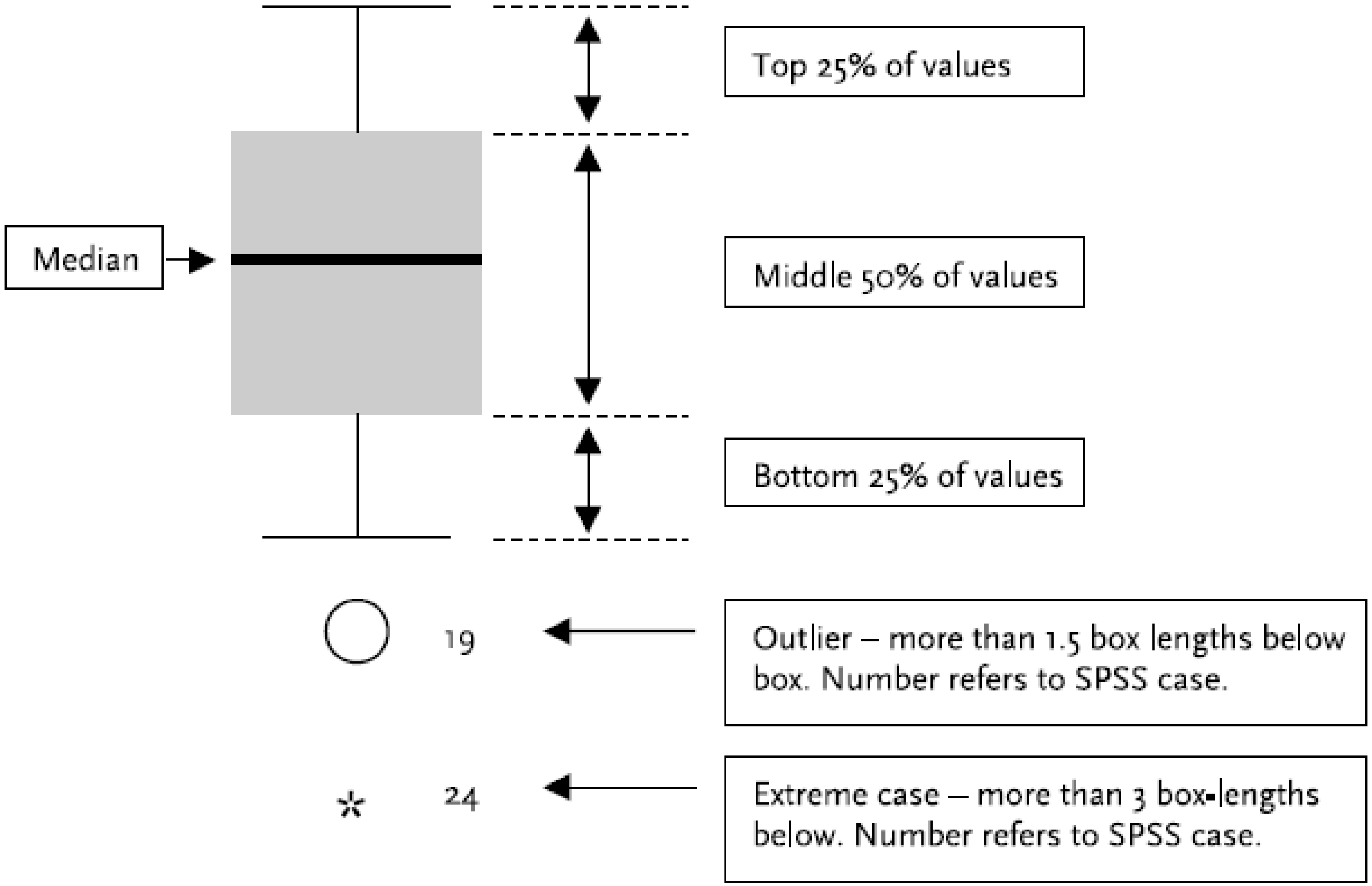
This data breaks down the percentage of men and women in information technology specialist positions over time. Currently, 22.2% of information technology specialists are female.


● Unemployment rate
8.0%
7.5%
7.0%
6.5%
6.0%
5.5%
5.0%
4.5%
4.0%
3.5%
3.0%
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Scatterplot



Boxplot

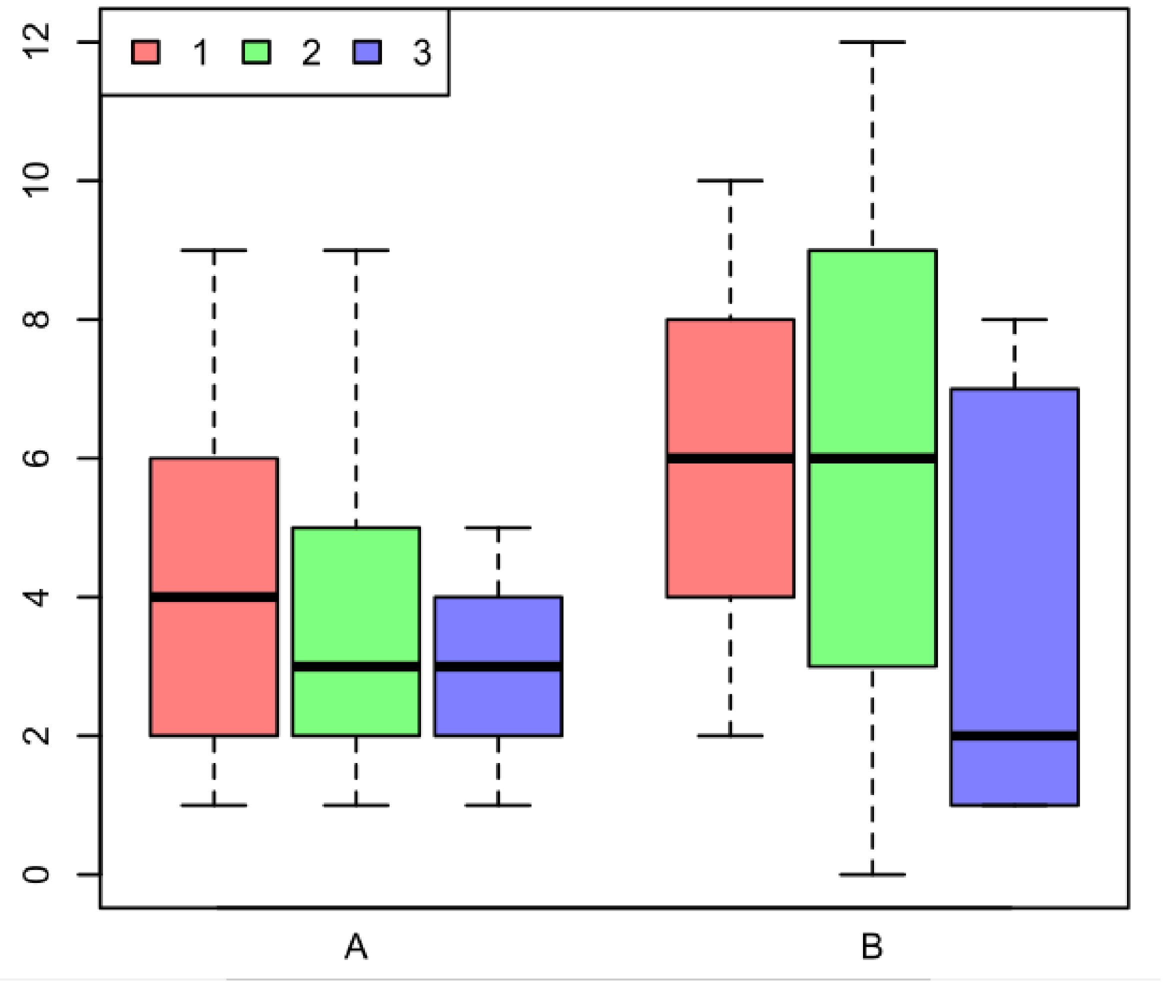

Top 25% of values
Median
Middle 50% of values
Bottom 25% of values
19
Outlier – more than 1.5 box lengths below box. Number refers to SPSS case.
24
Extreme case – more than 3 box-lengths below. Number refers to SPSS case.

Boxplot

Boxplot illustrates the spread of the data:
- the shaded box contains the middle 50 per cent of values;
- the line inside the box depicts the median value;
- the T-bar lines above and below the box reach to the highest and lowest values.

Notice the added inclusion of ‘outliers’ and ‘extreme cases’ which often occur in large datasets. These cases deserve special attention since they may skew any measures of central tendency. For example, a couple of extremely high ages would have increased the mean value, leaving the median the same. Outliers and extreme cases may also, of course, simply indicate an error in data entry which sometimes occurs in large datasets.

Clustered Boxplot



More graphs

https://python-graph-gallery.com/

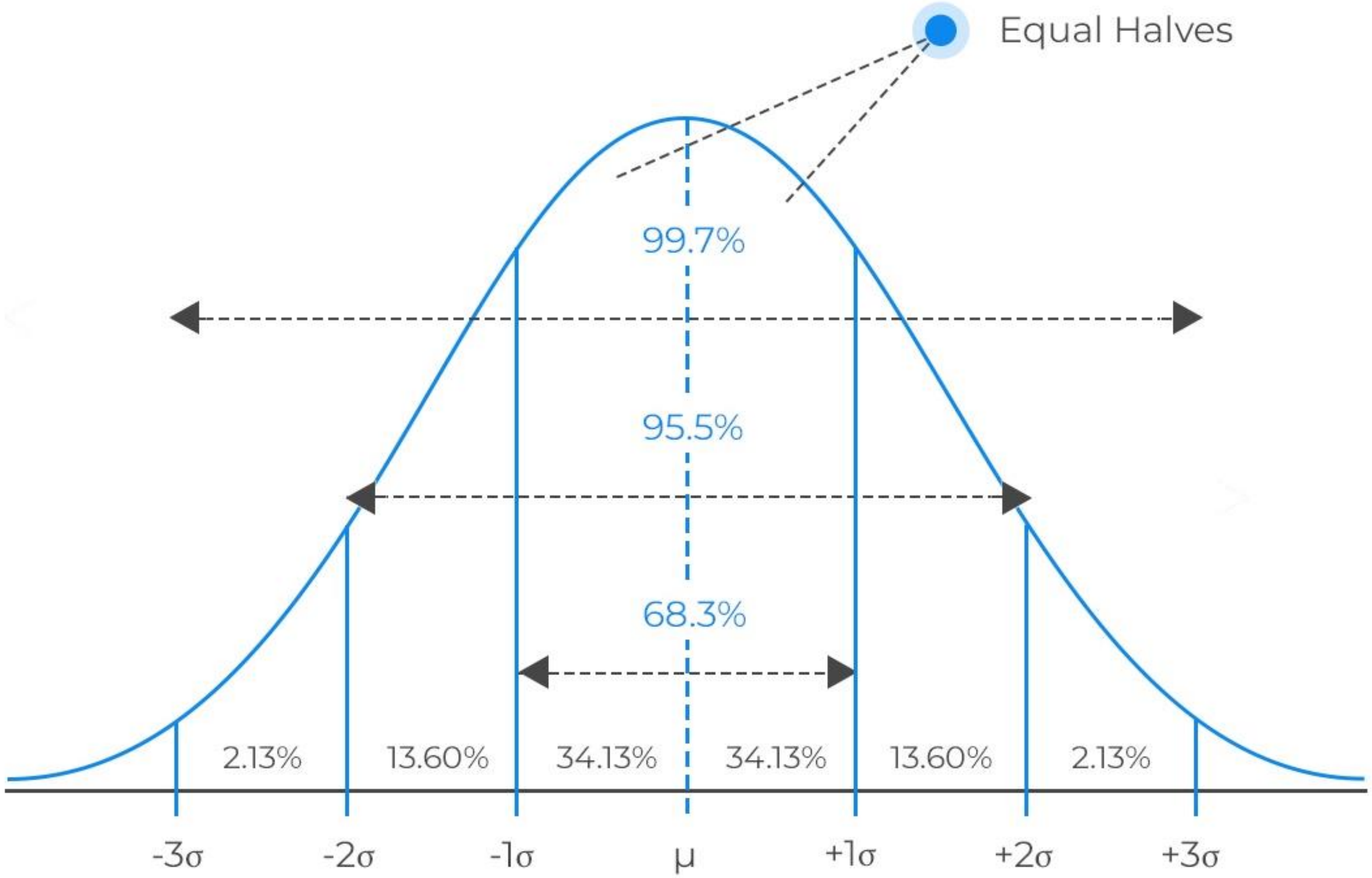
https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

Normal Distribution

Frequency distributions come in many different shapes and sizes. In an ideal world the data would be distributed symmetrically around the center of all scores. This is known as a normal distribution and is characterized by the bell-shaped curve.

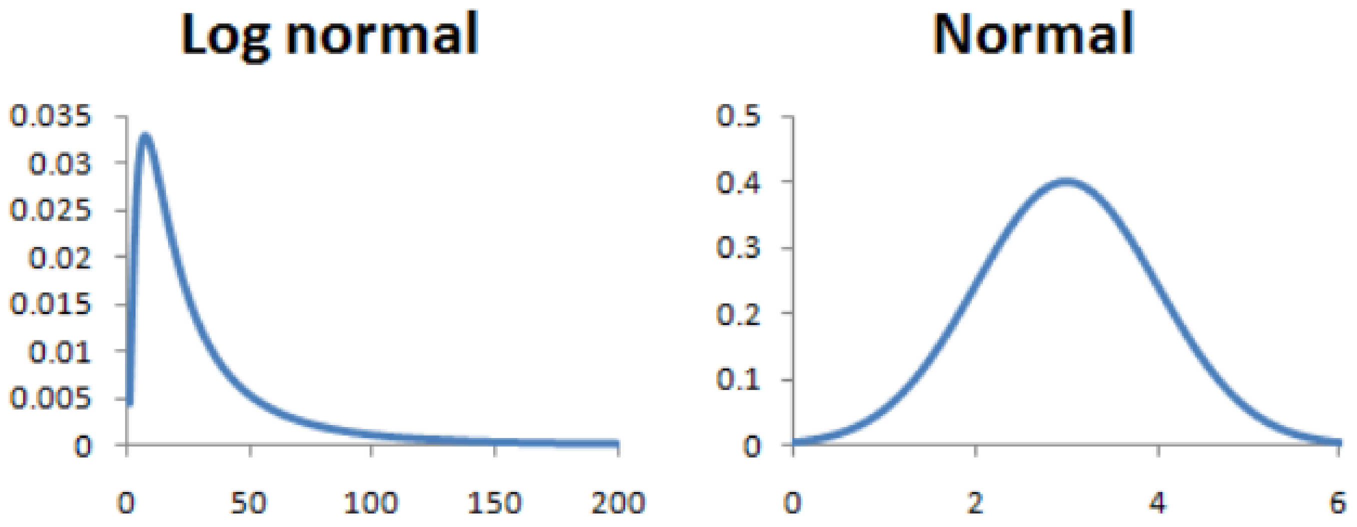
Note that Mean, Median and Mode are equal if the distribution is perfectly "normal" (i.e., bell-shaped).

Normal Distribution

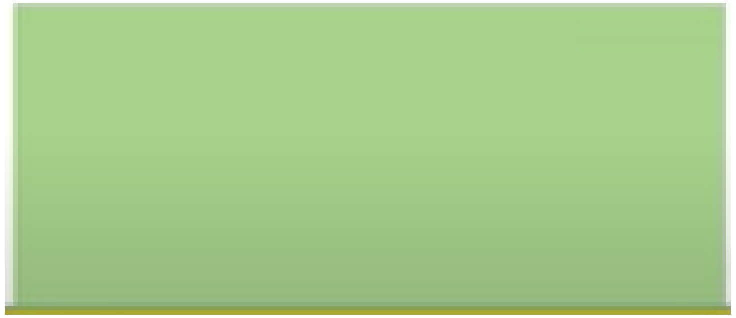


Log-Normal Distribution

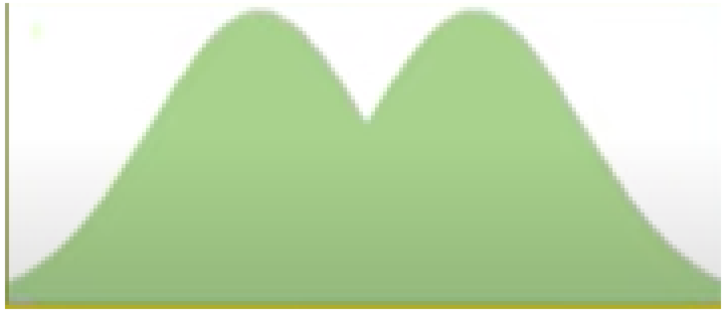
Log-Normal distribution is the discrete and ongoing distribution of a random variable, the logarithm of which is normally distributed. Log-Normal distribution follows the concept that instead of having the original raw data normally distributed, the logarithms of this raw data that are computed are also normally distributed.



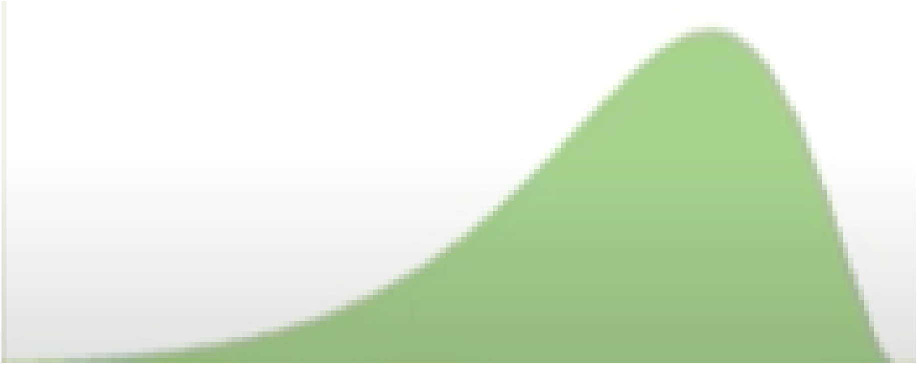
Other Distributions



Uniform distribution



Bimodal distribution



Skewed distribution

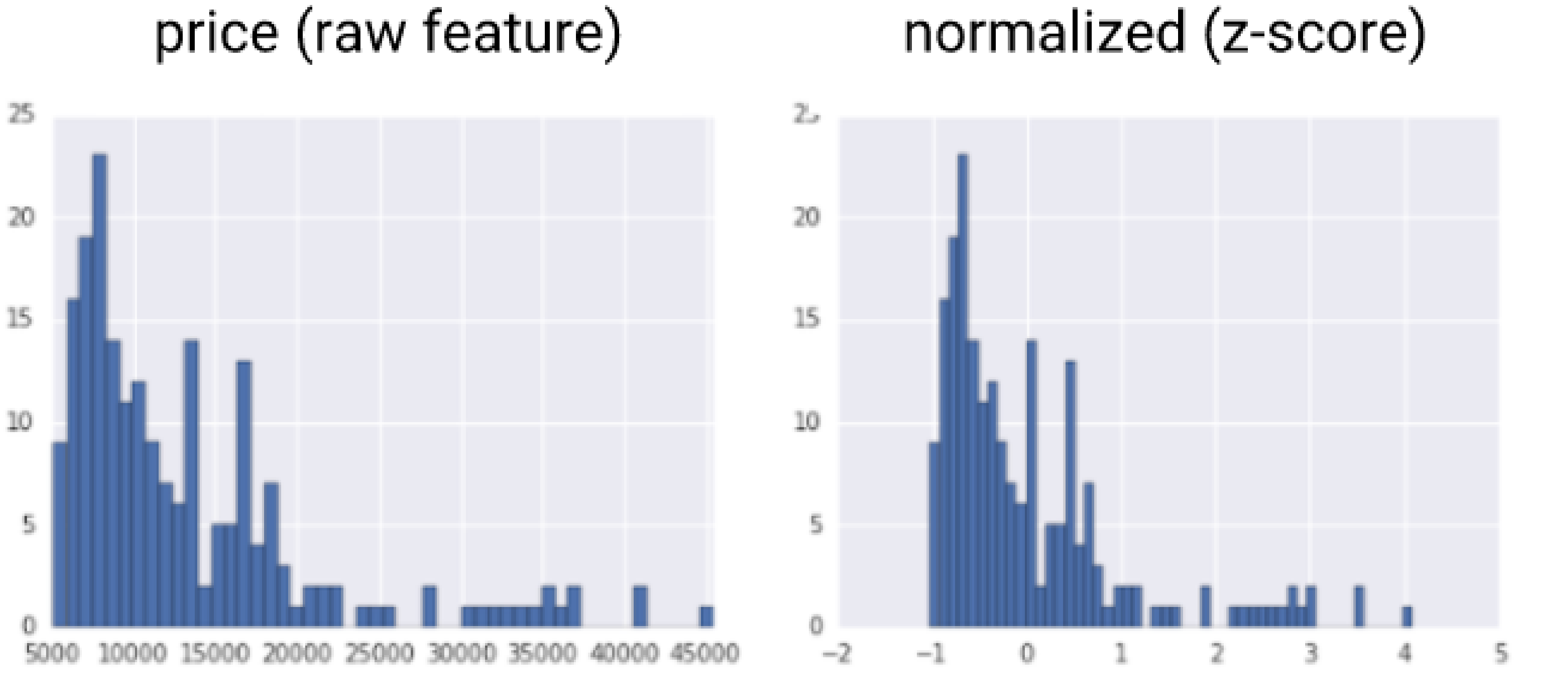
Z-Score Normalization

It’s possible to convert measures on very different scales, such as height and weight, into values that can be compared.

$$z_i = \frac{x_i - \overline{X}}{\sigma}$$

The mean of the standardized variable is 0 and the standard deviation is 1. Z-value is positive for observations that have a variable’s value above the mean, and negative if the values are below the mean.

Z-Score Normalization



Skewness and Kurtosis

There are two main ways in which a distribution of a variable can deviate from normal:

- Lack of simmetry (Skewness);
- Pointyness (Kurtosis).

In the case of a normal distribution, the value of Skewness is 0, and the Kurtosis is 3.

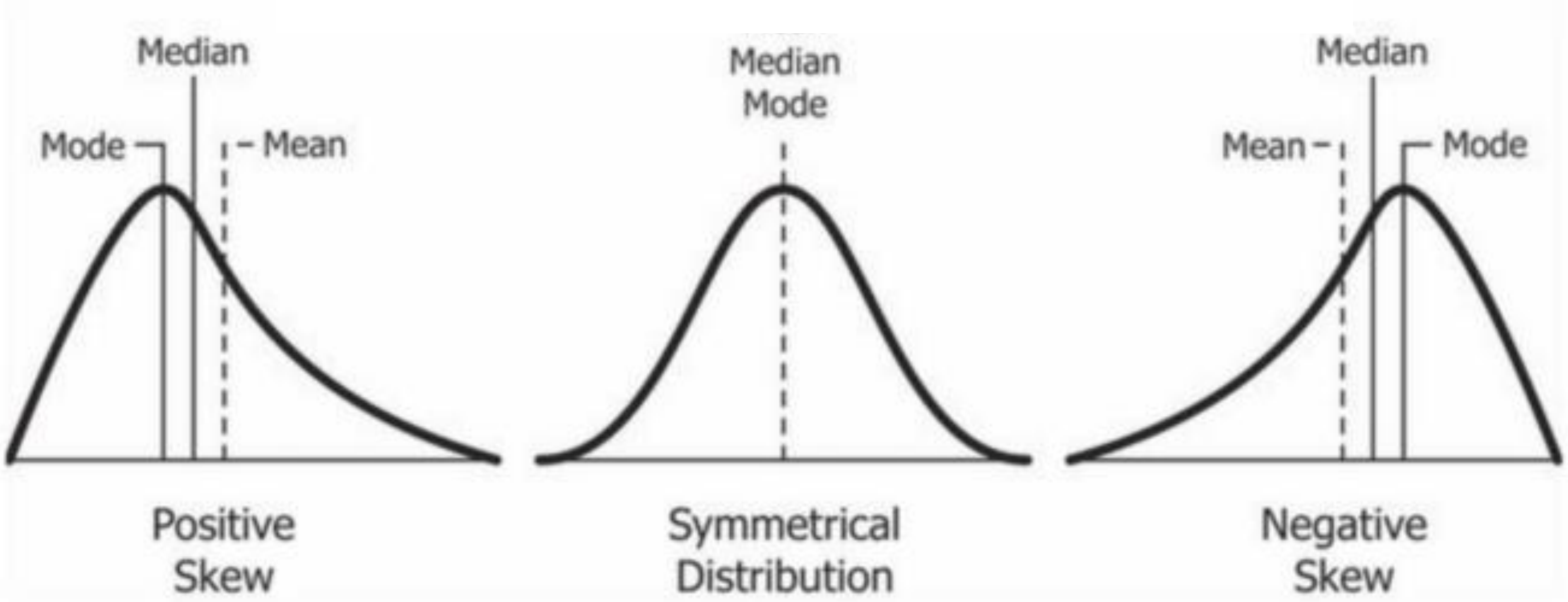
Skewness

Skewed distributions are not symmetrical and instead the most frequent scores (the tall bars of the graph) are clustered at one end of the scale. These distributions could be either:

- **positively skewed** (the frequent scores are clustered at the lower end and the tail points towards the higher or more positive scores) or
- **negatively skewed** (the frequent scores are clustered at the higher end and the tail points towards the lower more negative scores).

Skewness

$$skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$



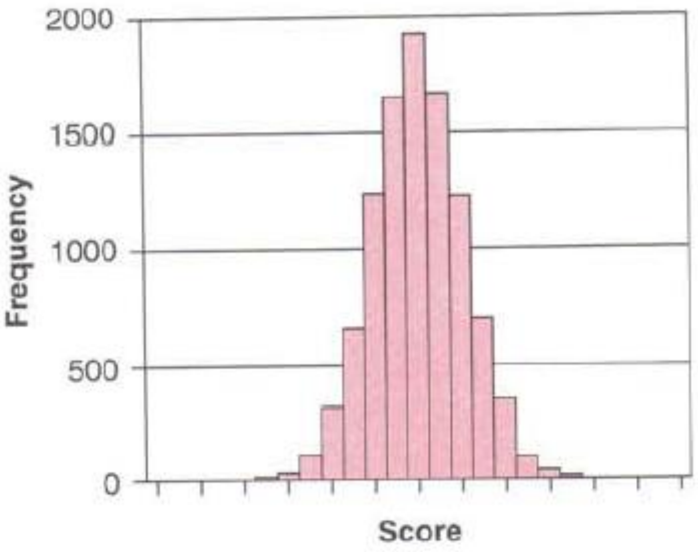
Kurtosis

Distributions may also vary in their **pointyness**, or kurtosis which refers to the degree to which scores cluster in the tails of the distribution.

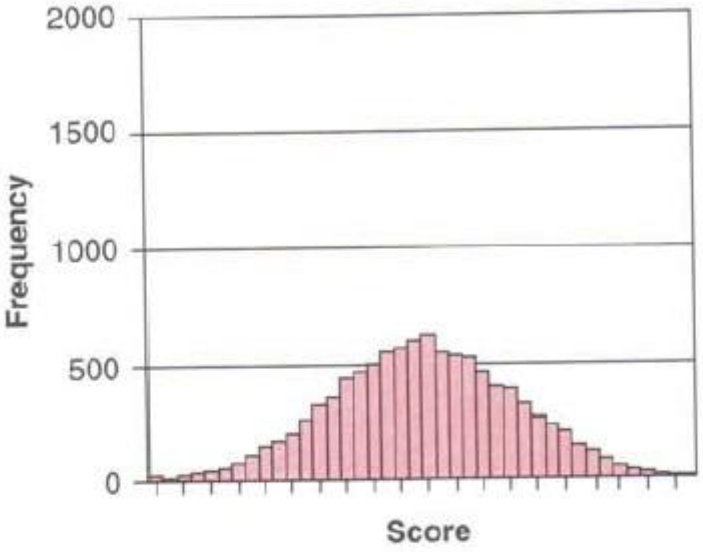
- The **leptokurtic** distribution is relatively thin in the tails and so looks quite pointy (value of kurtosis is >3).
- The **platykurtic** distribution is one that has many scores in the tails and so is quite flat.

Kurtosis

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

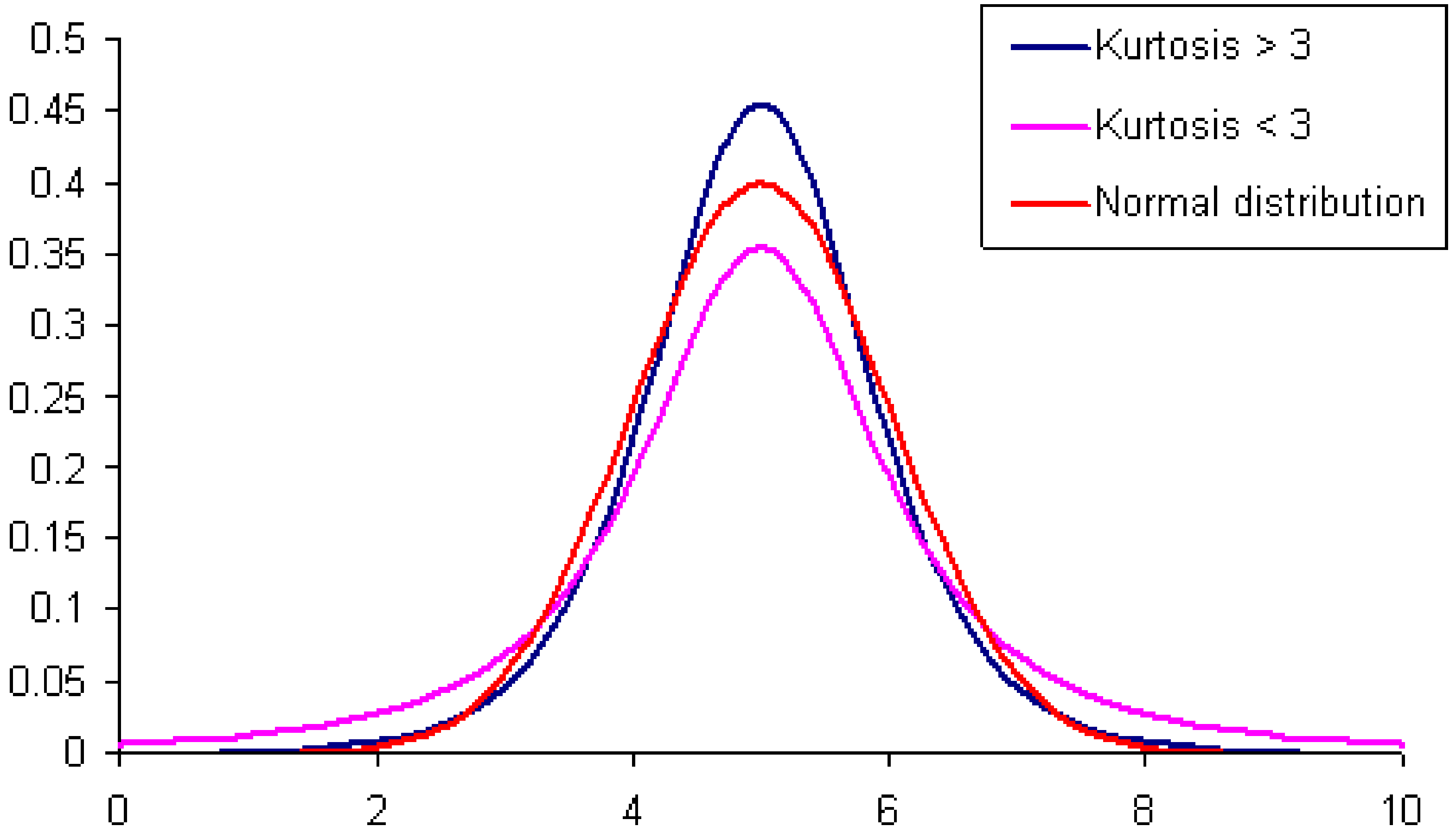


Leptokurtic distribution
(kurtosis > 3)



Platykurtic distribution
(kurtosis < 3)

Kurtosis


Kurtosis > 3
Kurtosis < 3
Normal distribution
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
2
4
6
8
10

Thank you for your attention!