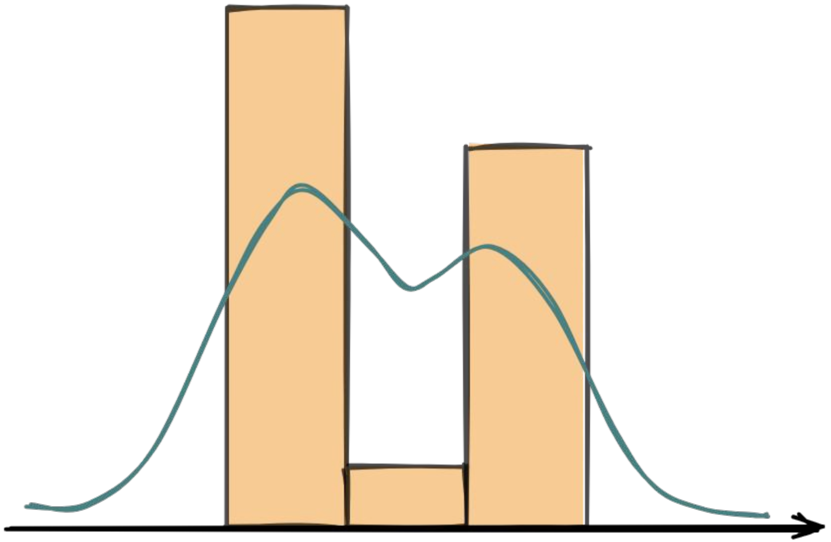


Statistics

Week 4

Middlesex University Dubai; CST4050: Winter21
Instructor: Dr. Ivan Reznikov

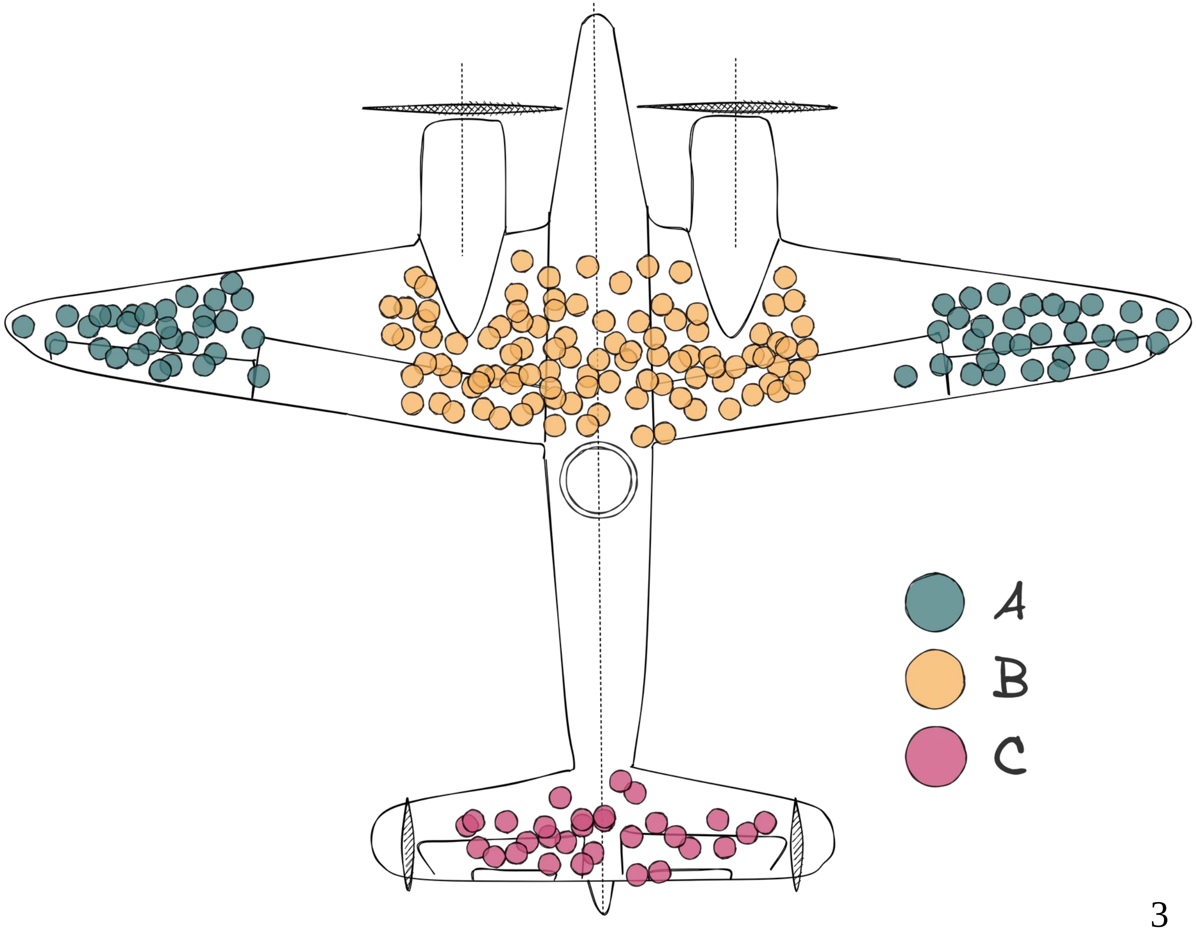
Plan

- Type of statistics
- Data types
- Inferential statistics
- Descriptive statistics
- Means, deviations and errors

Historical statistics case

This drawing is an illustration of a real problem that took place during WWII. The picture displays the frequency map of shots taken by airplanes. Each new shot is registered when the plane comes back from a mission.

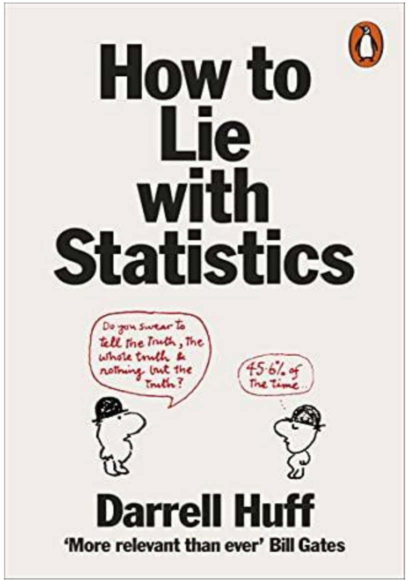
Given what you know about airplanes, what part of the plane would you equip with more armor?



Why is statistics important?

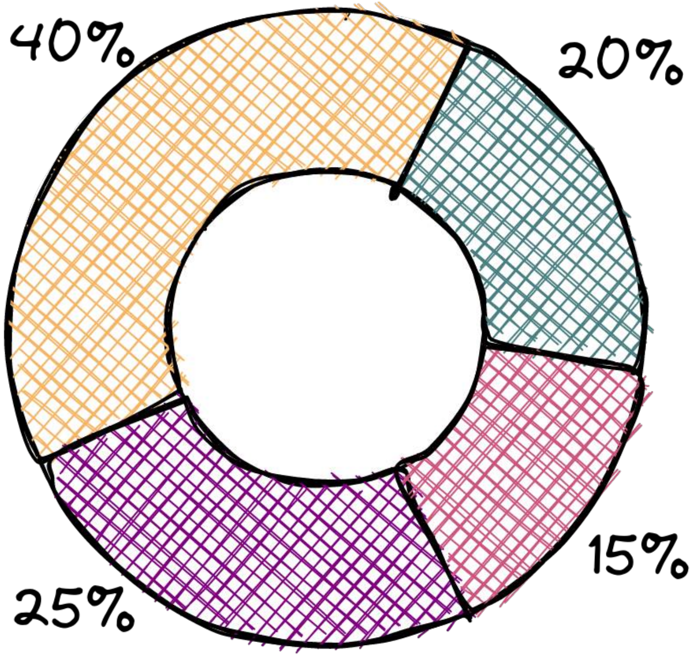
The field of statistics is the science of learning from data. Statistical knowledge helps you use the proper methods to

- collect the data
- employ the correct analyses
- and interpret results

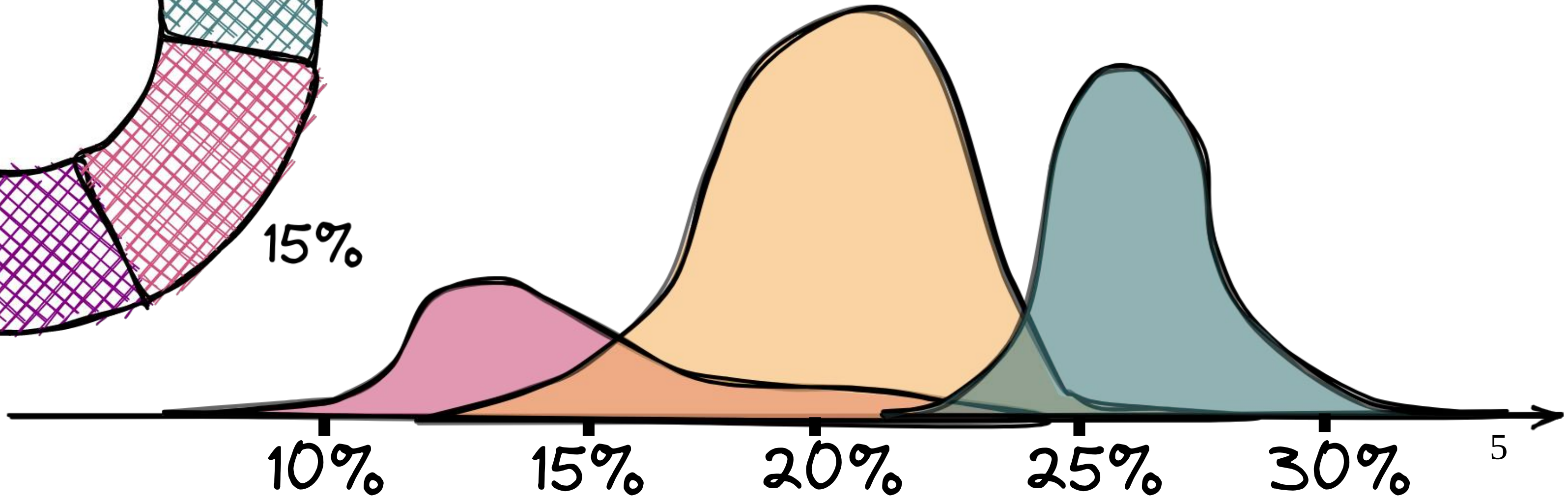


Types of statistics

Descriptive statistics



Inferential statistics

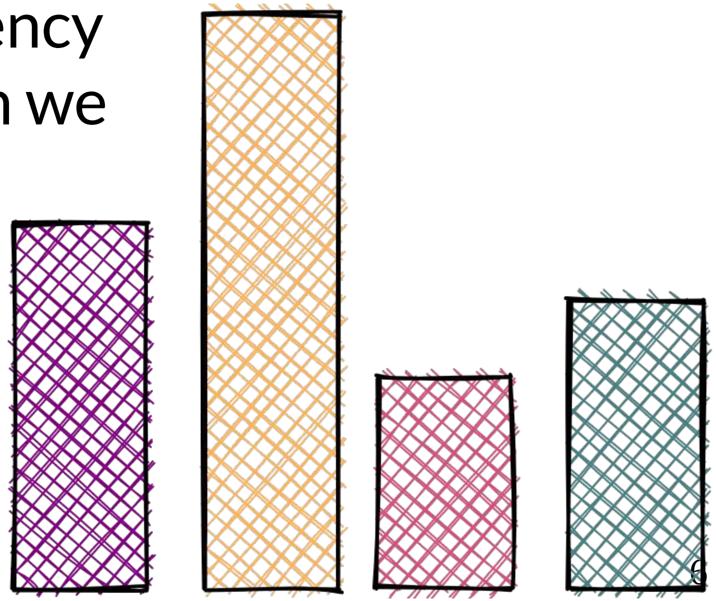


Descriptive statistics

Descriptive statistics describe the essential characteristics of your data.

Using the measures of the central tendency like mean, median, mode, and dispersion we can measure range, standard deviation, variance, etc.

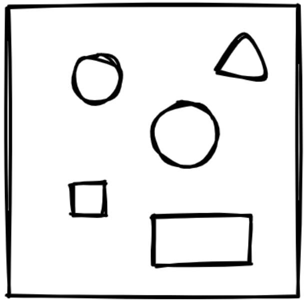
Data can be summarized using charts, tables, and graphs.

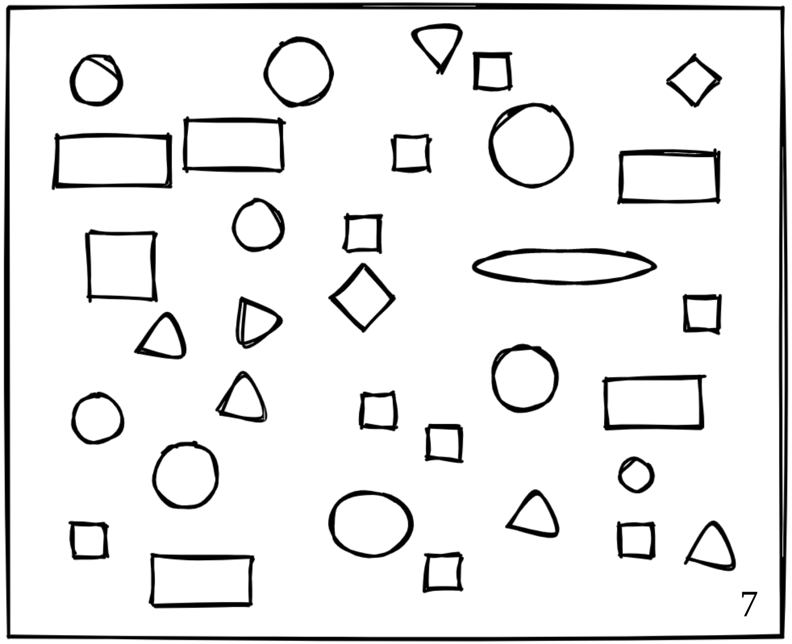


Inferential statistics

The goal of inferential statistics is to conclude a sample and generalize them to the population.

It is about using data from a sample and then making inferences about the larger population from which the sample is drawn





Descriptive or Inferential

We measured pulse of all football player from our team during the match. Their average was 120 bps with a maximum of 180

We measured the pulse of all football players from our team during the match and can conclude that the goalkeepers' pulse is lower than anyone's else

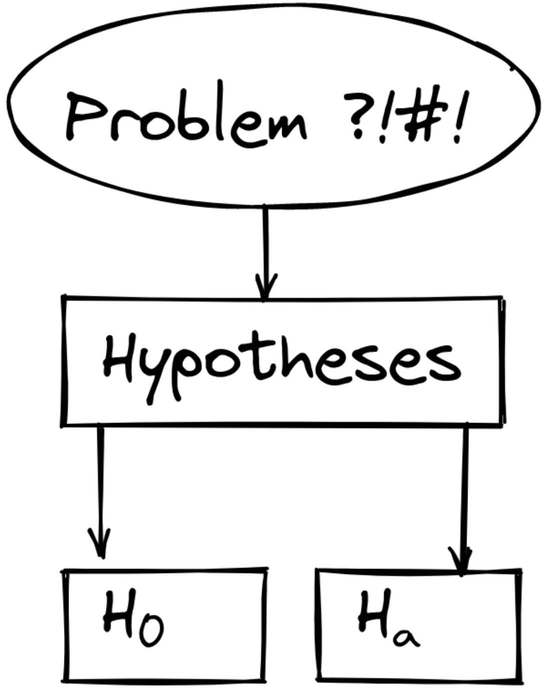
79% of employees of company A prefer to work at home

The average number of time spent by person on the beach in Dubai is 3.5 hours

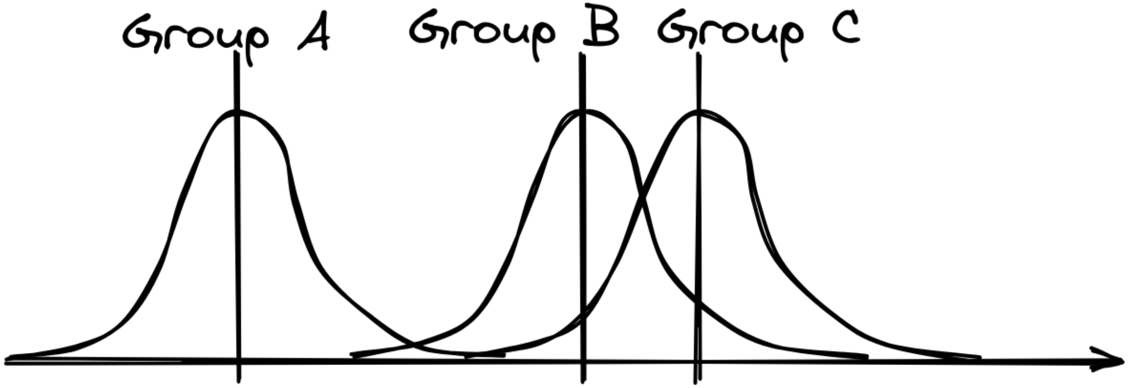
Inferential statistics tasks

The most common methodologies used are hypothesis testing and analysis of variance

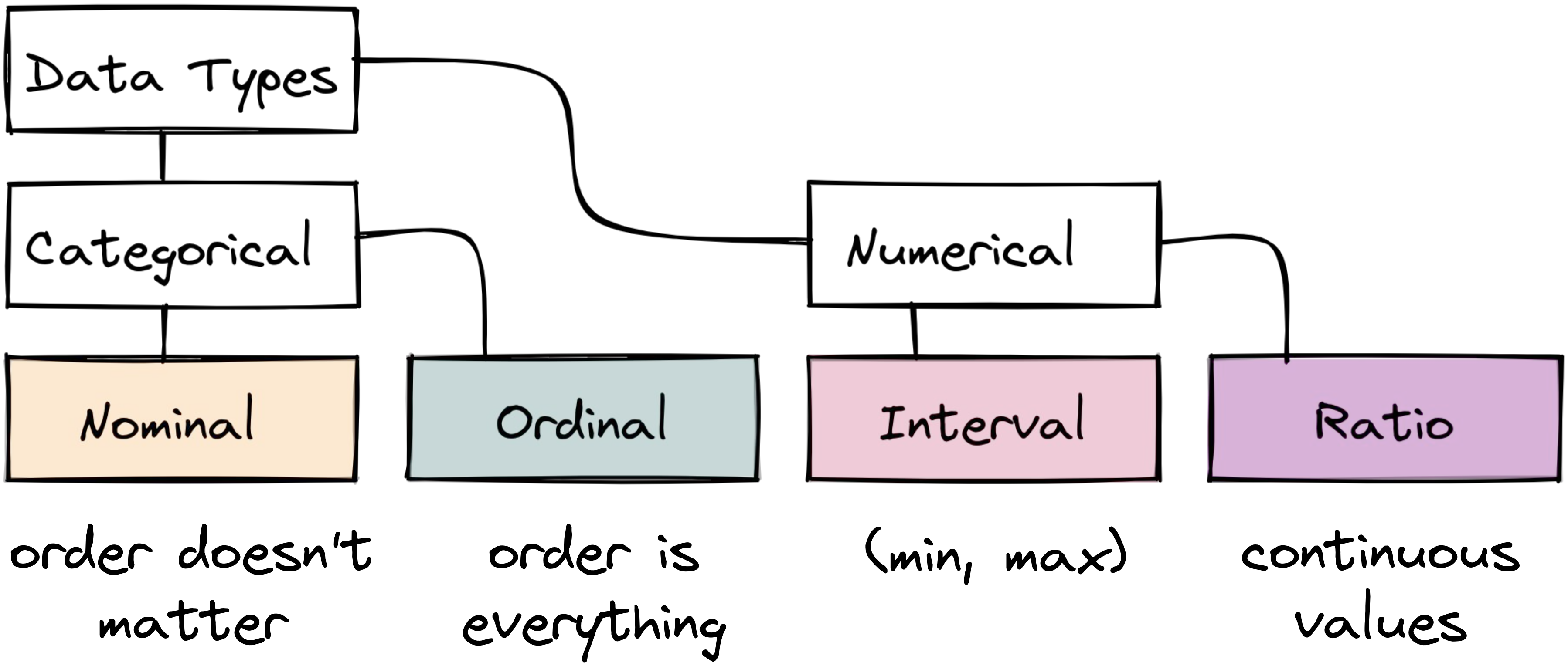
Hypothesis testing



Analysis of variance



Data Types


Data Types
Categorical
Numerical
Nominal
Ordinal
Interval
Ratio
order doesn't matter
order is everything
(min, max)
continuous values

Data Types

Nominal

Ordinal

Interval

Ratio

What visualizations would you use for certain data types?

Team	Seniority	Salary	YoE
Mobile	Lead	90-140	14
ML	Senior	100-110	8
Backend	Intern	40-50	0.5
Frontend	Middle	60-75	4
Mobile	Junior	40-45	2
Mobile	Senior	70-80	6
IoT	Middle	30-60	3
DevOPs	Senior	45-50	5
Manager	Assistant	35-40	3
ML	Lead	80-120	9

Descriptive statistics: Means

list1 = 1, 2, 3, 3, 4, 5, 10	list2 = 4, 9
Mean $= \dfrac{\Sigma(\text{list1}_i)}{\text{len}(\text{list1})} = 4$	Mean $= \dfrac{\Sigma(\text{list2}_i)}{\text{len}(\text{list2})} = 5.5$
Geometric Mean = $= \Pi(\text{list1}_i)^{1/\text{len}(\text{list1})} = 3.22$	Geometric Mean = $= \Pi(\text{list2}_i)^{1/\text{len}(\text{list2})} = 6$
Harmonic Mean $= \dfrac{\text{len}(\text{list1})}{\Sigma \frac{1}{\text{list1}_i}} = 2.58$	Harmonic Mean $= \dfrac{\text{len}(\text{list2})}{\Sigma \frac{1}{\text{list2}_i}} = 5.54$

Descriptive statistics: Means

Task1.
People contracting some pandemic virus increased
by 10% in day1,
by 20% in day 2
by 30% in day 3.
What's the average daily increase rate?

Task2.
A swimmer spends one lap of freestyle at 3km/h, then one lap of breaststroke at 2 km/h. What's his average speed?

Descriptive statistics: Means

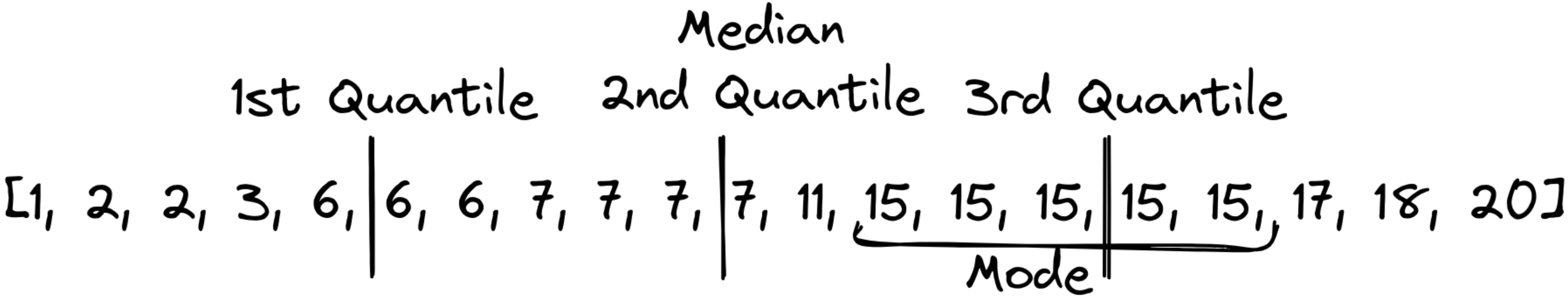
Task1.
day 0 – 1000 cases
day 1 – 1000 + 0.1×1000 = 1100 cases (+10%)
day 2 – 1110 + 0.2×1100 = 1320 cases (+20%)
day 3 – 1320 + 0.3×1320 = 1716 cases (+30%)
In this case it makes sense to use geometric mean => $(1716/1000)^{0.33} = 1.197$ => **19.7%**

Task2.
For simplicity, we'll assume that the lap is 3km. Freestyle will take 1hr, breaststroke will take 1.5. In total it'll take him 2.5 hrs for 6 km => 6/2.5 = **2.4 km/hr**.
Basically, we could've divided 2km/(1/2+1/3) = **2.4 km/hr**

Descriptive statistics: Means

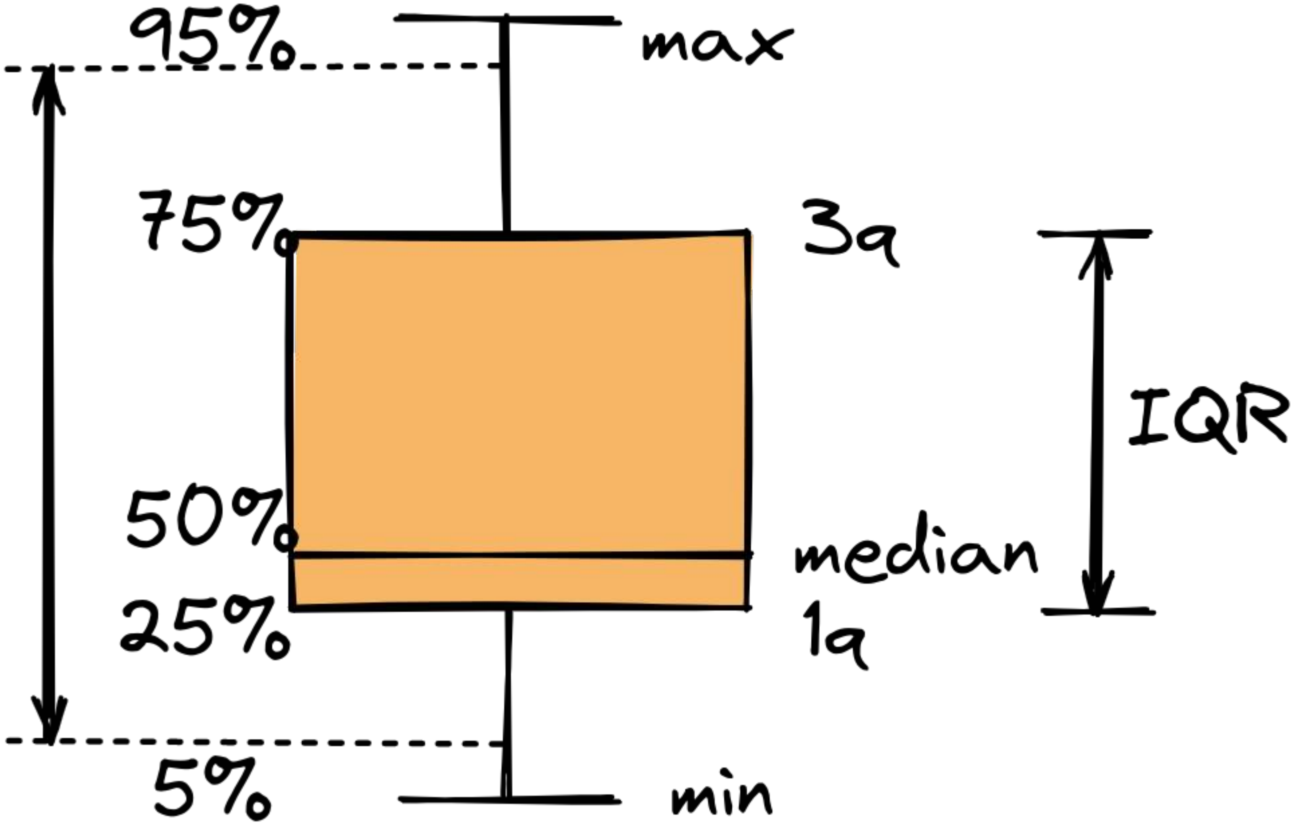
list 3 = [2, 15, 6, 2, 3, 17, 18, 20, 7, 1, 15, 6, 7, 11, 15, 6, 7, 7, 15, 15]

Good practice is to sort the list first:
list3 = [1, 2, 2, 3, 6, 6, 6, 7, 7, 7, 7, 11, 15, 15, 15, 15, 15, 17, 18, 20]

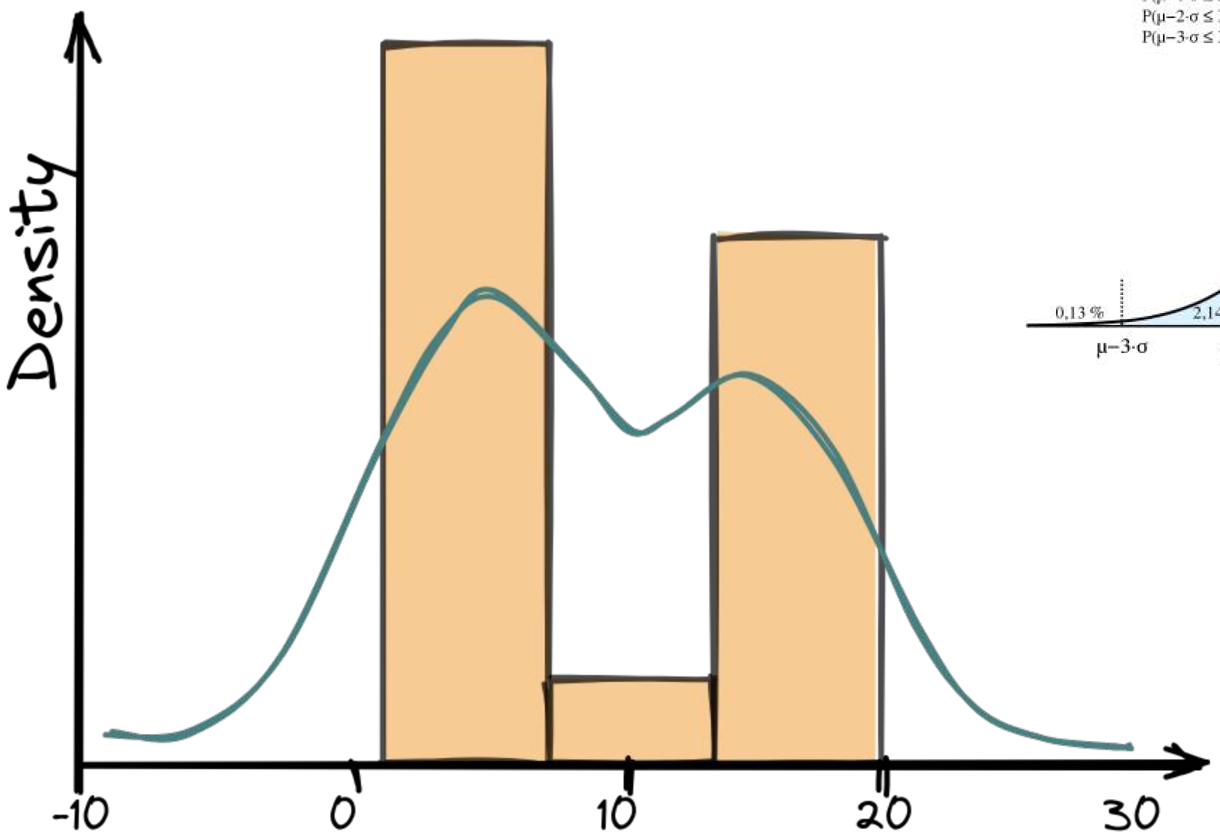


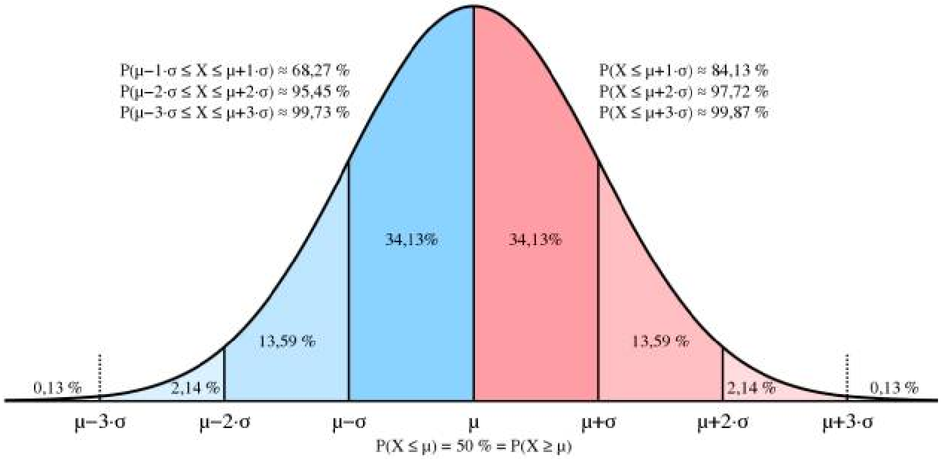
Mean = 9.5; Median = 7; Mode = 15

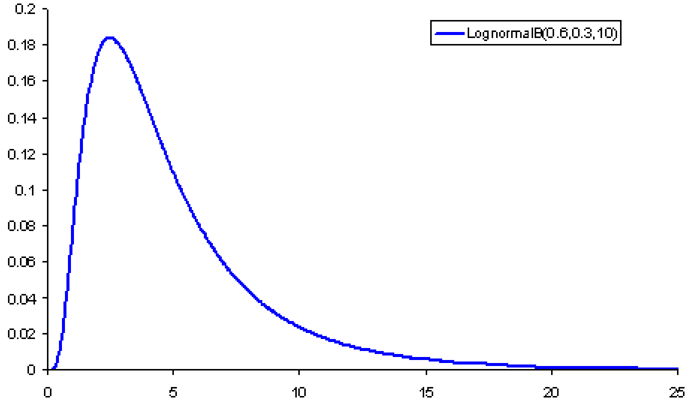
Boxplot



Distribution plots







Descriptive statistics

list3 = [1, 2, 2, 3, 6, 6, 6, 7, 7, 7, 7, 11, 15, 15, 15, 15, 15, 17, 18, 20]

$$MD = \frac{\Sigma(list1_i - \mu)}{len(list1)}$$

mean deviation

$$QD = \frac{Q_3 - Q_1}{2}$$

quartile deviation

$$\sigma^2 = \frac{\Sigma(list1_i - \mu)^2}{len(list1)}$$

variance

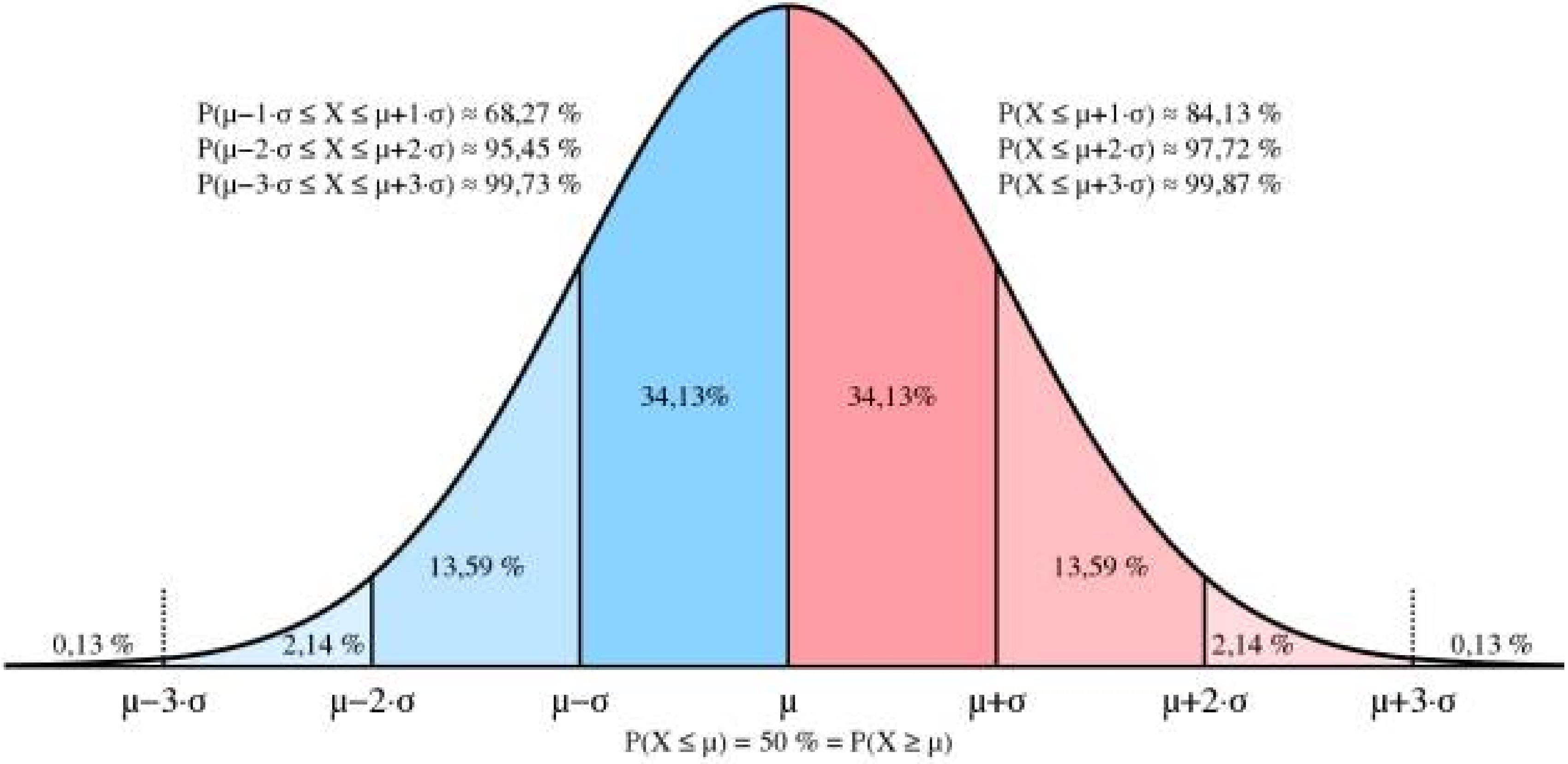
$$\sigma = \sqrt{\frac{\Sigma(list1_i - \mu)^2}{len(list1)}}$$

standard deviation

$$SE = \frac{\sigma}{len(list)}$$

standard error

Normal distribution



Mean, Median, Mode