Statistical Techniques for Data Science & Robotics

Week 1

Objectives (for today)

(1) to learn about the course: how to success in it

(2) to recall Probability theory

(3) to do short introduction to Statistics

Team and Communication

(1) Primary Instructor: Vladimir Ivanov
    (1) @nomemm
    (2) room 475
(2) Teaching Assistants:
    (1) Zamira Kholmatova
    (2) Sofya Muhamedzhanova
    (3) Maxim Evgrafov

Syllabus

(1) "Classical" Statistics: Tests, Hypotheses

(2) Non-parametric Statistics

Midterm

(3) Bayesian Statistics

(4) Sampling, MCMC

(5) Bandit algorithms

Structure

(1) Lectures (2 hours/week) + Quizzes (10 min. on a lecture)

(3) Labs (2 hours/week) + Assignments (aka Homework, up to 4 hours / week)

Books

(1) Introduction to Mathematical Statistics. By ROBERT V. HOGG AND ALLEN. B. CRAIG 4th Edition

(2) Bruce, Peter,Bruce, Andrew. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media.

(3) Hastie, T. Tibshirani, R. and Friedman, J. (2008) The Elements of Statistical Learning 2ed. Springer

(4) *Bishop Christopher. Pattern Recognition and Machine Learning. Springer, 2006. — 738 p.

Grading

(1) Quizzes: **10** %

(2) Labs: **10** %

(3) Assignments: **30** %

(4) Midterm (theory + practice): **20** %

(5) Final Exam: **30** %

85 - 100	A
70 - 84	B
55 - 69	C
0 - 54	D

Tools

(1) Pen and Paper

(2) R / Python (ver. 3+)

Course Prerequisites

(1) Linear Algebra, Calculus courses

(2) A course on Probability Theory

How to success?

(1) Assignments and Labs:

  (1) work hard (individually) + office hours

  (2) visit Labs to have enough practice with tools

  (3) visit Labs to solve pen and paper problems

(3) Exams:

  (1) read the books + office hours

  (2) solve exercises from books at home

Break, 5 min.

Recap of Probability Theory

Major Concepts

What is …?

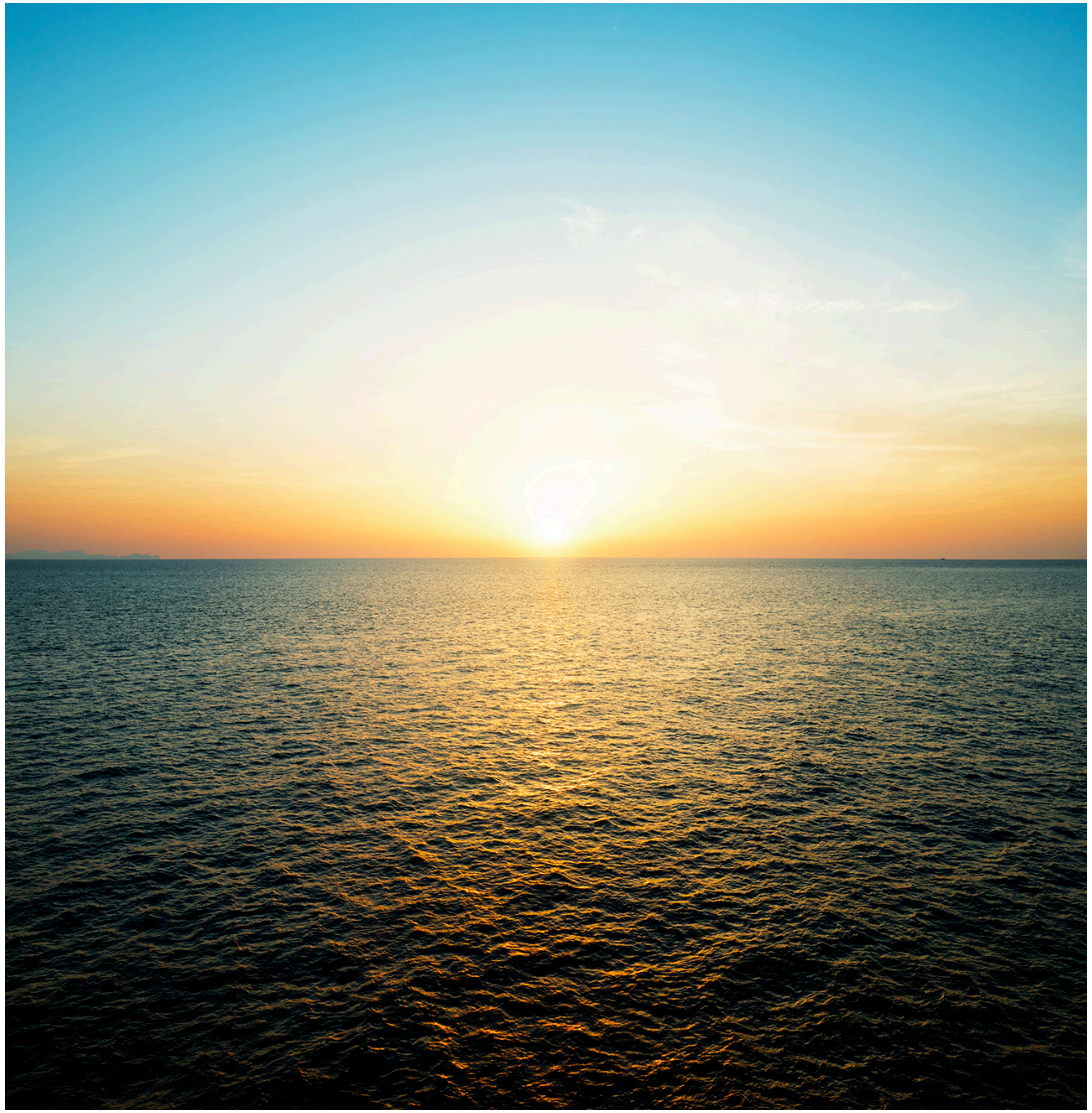
(1) Random Variable

(2) Expected value, variance

(3) Probability Density Function

(4) CDF

(5) Bayes Theorem



Random variables (r. v.)

(1) What are examples of random variables (YES/NO)?

A. Winning a lottery
B. Choosing a green ball from an urn with a large mixture of red and black balls
C. Total value from a roll of two dice
D. Two people in a classroom sharing the same birthday
E. The average exam score of a class if every student guesses answers
F. Winings from a game with a $1 gain/loss for each head/tail coin flip in a series of 10 flips

Random variables

(1) What are examples of random variables?

A. Winning a lottery

B. Choosing a green ball from an urn with a large mixture of red and black balls

C. Total value from a roll of two dice

D. Two people in a classroom sharing the same birthday

E. The average exam score of a class if every student guesses answers

F. Number of winings from a game with a $1 gain/loss for each head/tail coin flip in a series of 10 flips

Random variables, sample space

(1) Do not confuse these things!

  (1) Outcomes,

  (2) Events and

  (3) Random Variables

$$\Omega = \{H, T\}$$

$$X(H) = 1; \quad X(T) = 0$$

(1) A random variable assigns a numerical value to each outcome of a chance event (in a random experiment).

$$P(X = 1) = ?$$

Flip a coin 3 times

(1) Let N be the number of tails (T)

(2) N is a random variable

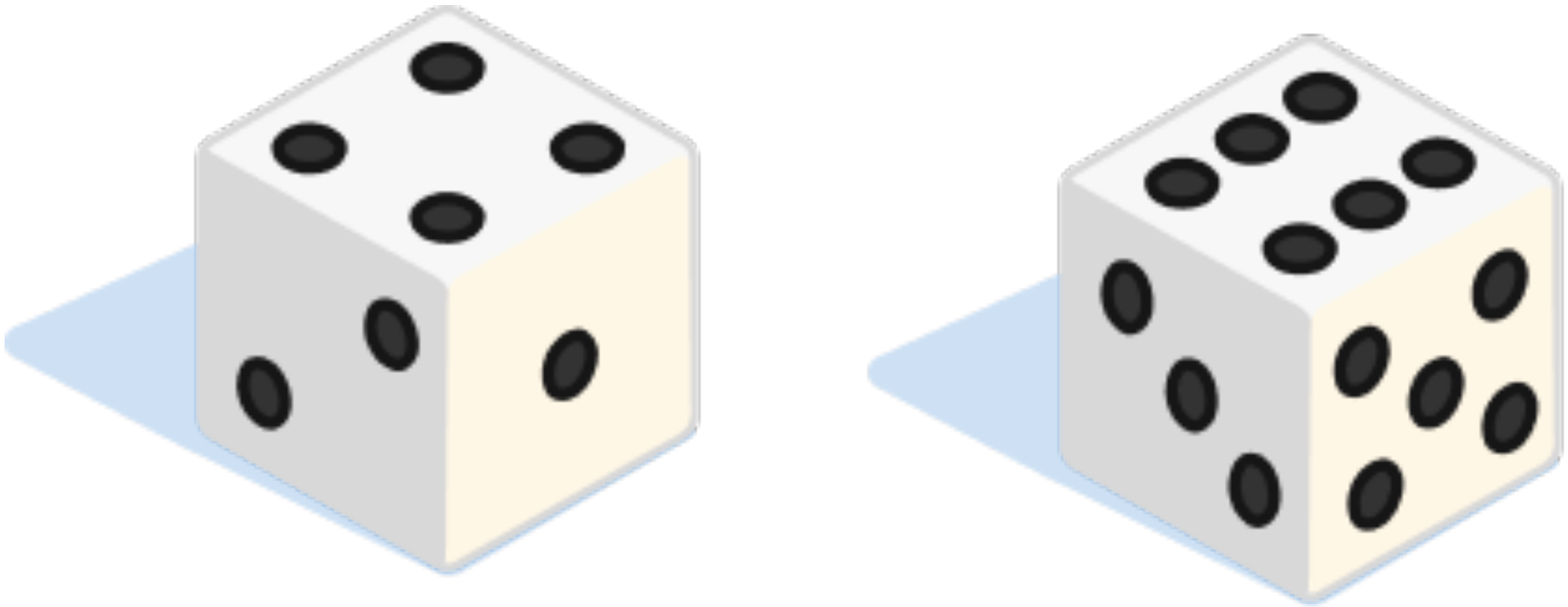
(3) N = 2 is an event

$$\Omega = \{HHH, HHT, HTT, TTT, THH, TTH, HTH, THT\}\ .$$

$$P(N = 2) = ?$$

$$P(N > 2) = ?$$

2 Dices



$$P(X \geq 10) = ?$$

Functions of random variables

(1) Functions of r.v. can produce new r.v.

(2) For example, given X, Y and Z are r.v.

$$A = \frac{1}{3}(X + Y + Z)$$

- <u>HOMEWORK exercise:</u>
  If the exam has three true/false questions and a score less than 20% doesn't pass the exam.
- What is the probability the group of three students fails on average?

Probability Distributions (of r.v.)

(1) **A random variable** quantifies chance events, and

(2) **its probability distribution** assigns a likelihood to each of its (r.v.) values.

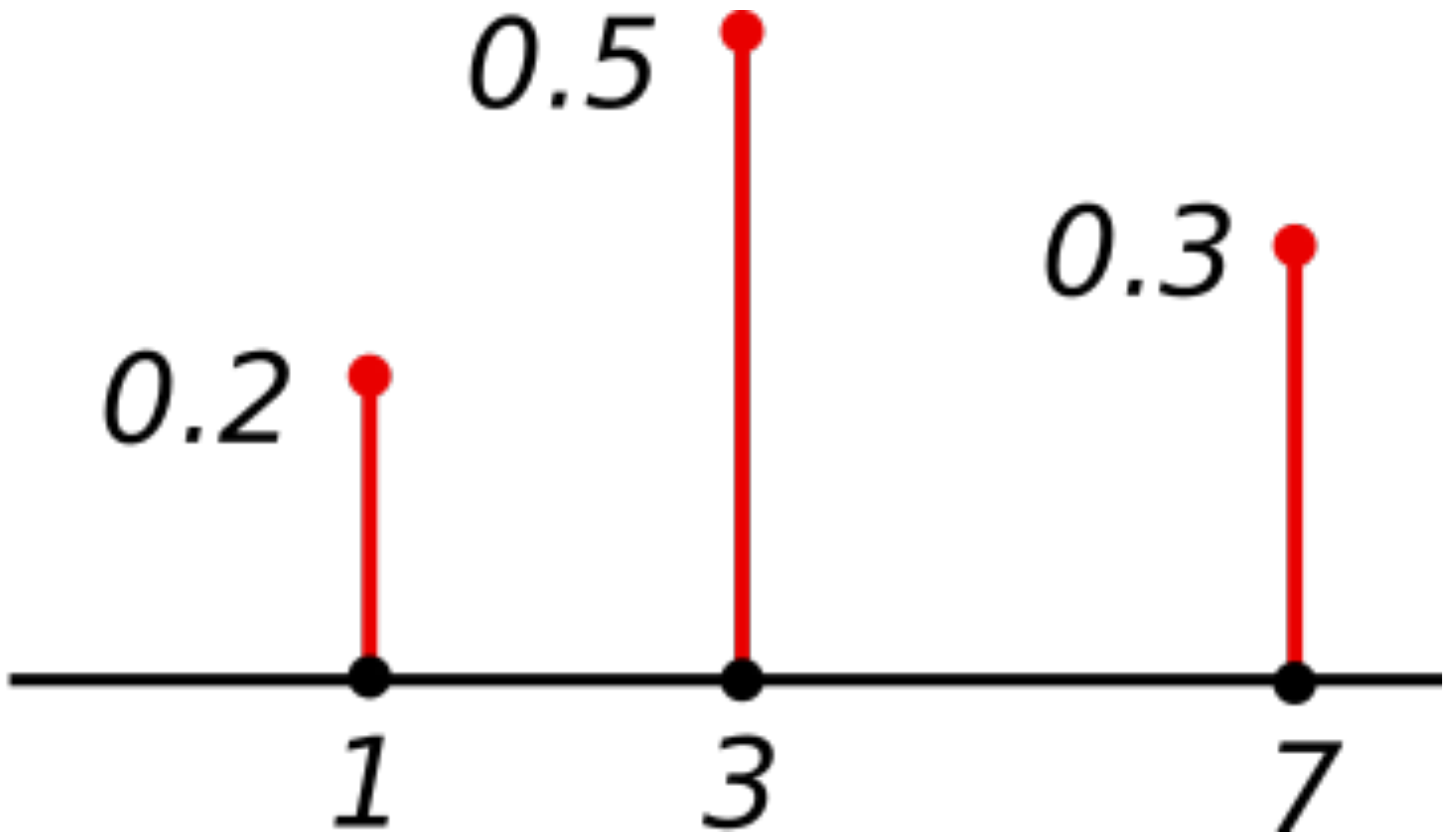
(3) Depending on nature of event the r.v. and distribution can be:

   (1) discrete

   (2) continuous

Probability mass function

For discrete random variable X: PMF



Probability density function

For continuous random variable X: PDF

$f_X(x)$ is a probability density function for r.v. X

$$P[a < X \leq b] = \int_a^b f_X(x)\, dx$$

x is just a value (argument) from the domain of $f_X(x)$

What about the **range** of $f_X(x)$?

Cummulative distribution function

CDF

(1) if $f_X(x)$ is continuous at x:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$

$$f_X(x) = \frac{d}{dx}F_X(x)$$

$$F_X(x)' = f_X(x)$$

Bernoulli and Binomial

(1) Any random experiment whose outcome can be classified as either a success or a failure is called a **Bernoulli trial**.

(2) If you run T trials (and probability of success is **p**), then the number of successes is a r.v. N that has a discrete distribution.

(3) Binomial distribution

$$P(N = k) = \binom{T}{k} p^k (1-p)^{T-k}$$

Expected value and Variance

(1) Given X is a random variable

(2) Well, try to understand the formulas:

$$E[X] = \sum_{n \in X's \quad range} nP(X = n)$$

$$V[X] = Var[X] = E[(X - E[X])^2]$$

Show (as an exercise) that

$$Var[X] = E[X^2] - (E[X])^2$$

Expected value (continuous r.v.)

(1) Well, try to understand the formula:

$$E[X] = \int_{x \in S} x f(x) dx$$

S is range of random variable X

Expected value and its Properties

- $E[X] = \sum xP(x)$

- $E[X] = \int xf(x)dx$

- $E[X + c] = E[X] + c$
- $E[cX] = cE[X]$
- $E[X + Y] = E[X] + E[Y]$

Poisson distribution

(1) Your phone glitches randomly

(2) with certain rate of k failures per second.

(3) X is the number of glitches in a time period *t* since release

(4) X is a r.v. that has Poisson distribution

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, n = 1,2,3,...$$

$$\lambda = kt$$

Poisson variables/distribution

(1) Find

$$P(X = n) = \frac{\lambda^n}{n!}e^{-\lambda}, n = 1,2,3,\ldots$$

$$\lambda = kt$$

$$E[X] = ?$$

$$Var[X] = ?$$

Exponential

(1) For instance, the time t it takes for the decay of a radioactive carbon-14 atom into stable nitrogen-14 is distributed **exponentially**

$$f(t) = ae^{-at}, t \geq 0, a > 0$$

What's the probability that a carbon-14 atom lasts at least as long as $\frac{1}{a}$?

Probability Theory

(1) It is about applications and for modeling of uncertainty

Axiomatic Probability Theory: Sample Space

The sample space Ω is the set of possible outcomes of an experiment.

Subsets of Ω are called events.

A class of events $\mathcal{E}$ is called a σ-algebra if:

(1) $\emptyset \in \mathcal{E}$

(2) $A \in \mathcal{E} \implies A^c \in \mathcal{E}$

(3) $A_1, A_2, \ldots \in \mathcal{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

Axiomatic Probability Theory

A probability measure is a function $\mathbb{P}$ defined on a σ-algebra $\mathcal{E}$ such that:

(1) $\forall A \in \mathcal{E}, \mathbb{P}(A) \geq 0$,

(2) $\mathbb{P}(\Omega) = 1$,

(3) if $A_1, A_2, \ldots \in \mathcal{E}$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Probability Space

(2) The triple $(\Omega, \mathscr{E}, \mathbb{P})$ is called a probability space

Random Variables and Events

So, what are events here?

A **random variable** is a map $X : \Omega \to \mathbb{R}$ such that, for every real x, $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{E}$.

σ-algebra

(1) Example of σ-albegra

A class of events $\mathcal{E}$ is called a σ-algebra if:

(1) $\emptyset \in \mathcal{E}$

(2) $A \in \mathcal{E} \implies A^c \in \mathcal{E}$

(3) $A_1, A_2, \ldots \in \mathcal{E} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

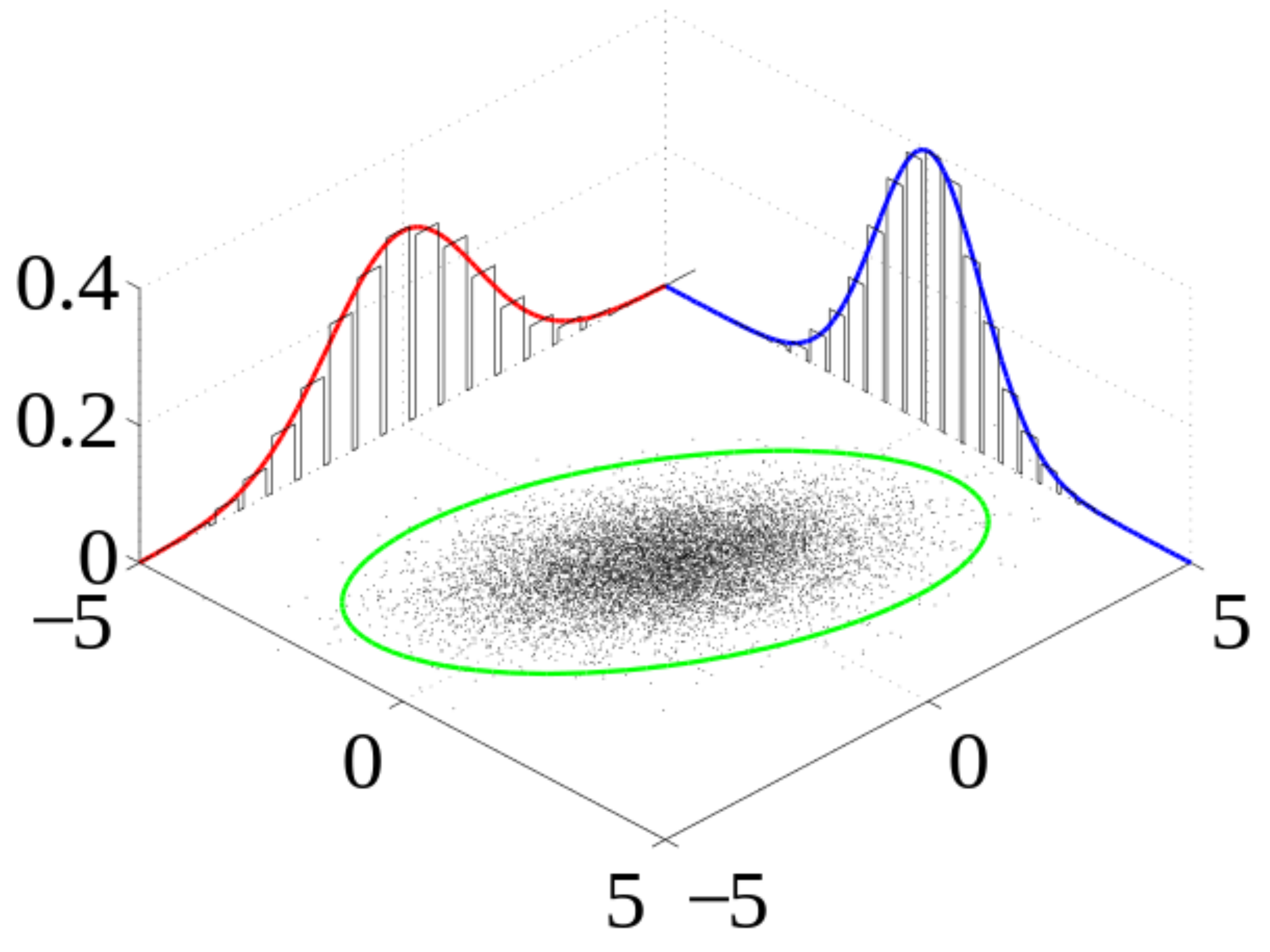
Still, you can figure it out!

Joint Distribution

(1) Given two random variables X,Y

(2) joint CDF is

$$F_{x,y}(x, y) = P(X \leq x, Y \leq y)$$



Independence of Events / Independence of r.v.

(1) Events A and B are independent iff

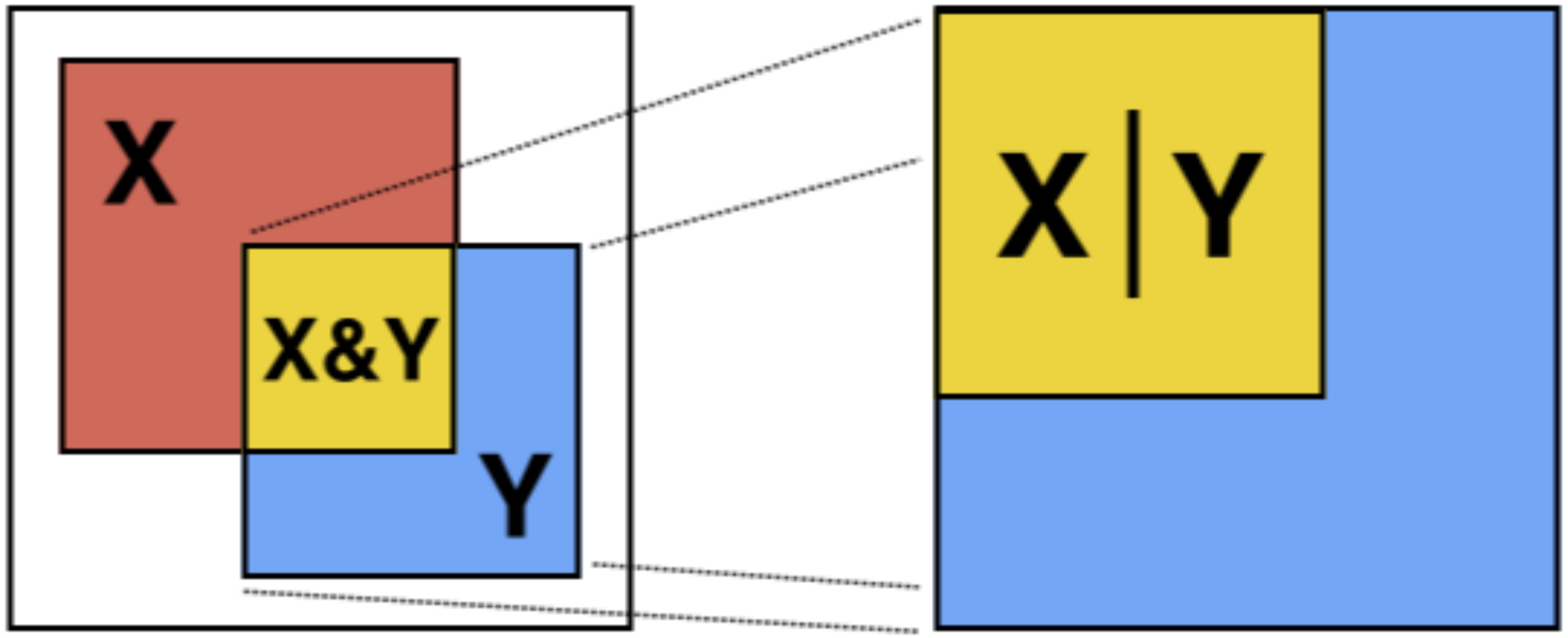
$P(A \cap B) = P(A)P(B)$

R.v. X and Y are independent iff

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \forall x, y$$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \forall x, y$$

Conditional Probability



Each large square has area 1, and the smaller squares represent probabilities.

$$P(X = x \mid Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Bayes' Theorem

(1) Philosophically, all probabilities are conditional probabilities.

Bayes' Theorem

(1) formula describes how to **update** the probabilities of hypotheses (H) when given evidence (E).

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H)$$

$$f_y(y \mid X = x) = \frac{f_X(x \mid Y = y)}{f_X(x)} f_Y(y)$$

Bayes' Theorem

(1) prior distribution,

(2) posterior distribution, and

(3) likelihood ratio

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H)$$

Conditional Expectation: discrete case

(1) The expected value of X given the event $Y = y$:

(2) $E[X \mid Y = y] = \sum_{x} xP(X = x \mid Y = y)$

HOMEWORK exercise:
What is the expected number of "heads" flips in 5 flips of a fair coin, given that the number of "heads" flips is greater than 2?

Conditional Expectation: cont. case

(1) The expected value of X with pdf $f(x)$

(2) Let S be the range of X given the event $Y = y$:

$$E[X \mid Y = y] = \frac{\int_{x \in S} x f(x) dx}{P(Y = y)}$$

HOMEWORK exercise: Let X be a continuous random variable

with density function $f(x) = 2x; 0 < x < 1$. What is $E[X \mid X < \frac{1}{2}]$?

PAUSE

Introduction to statistics

Major Concepts

What is …?

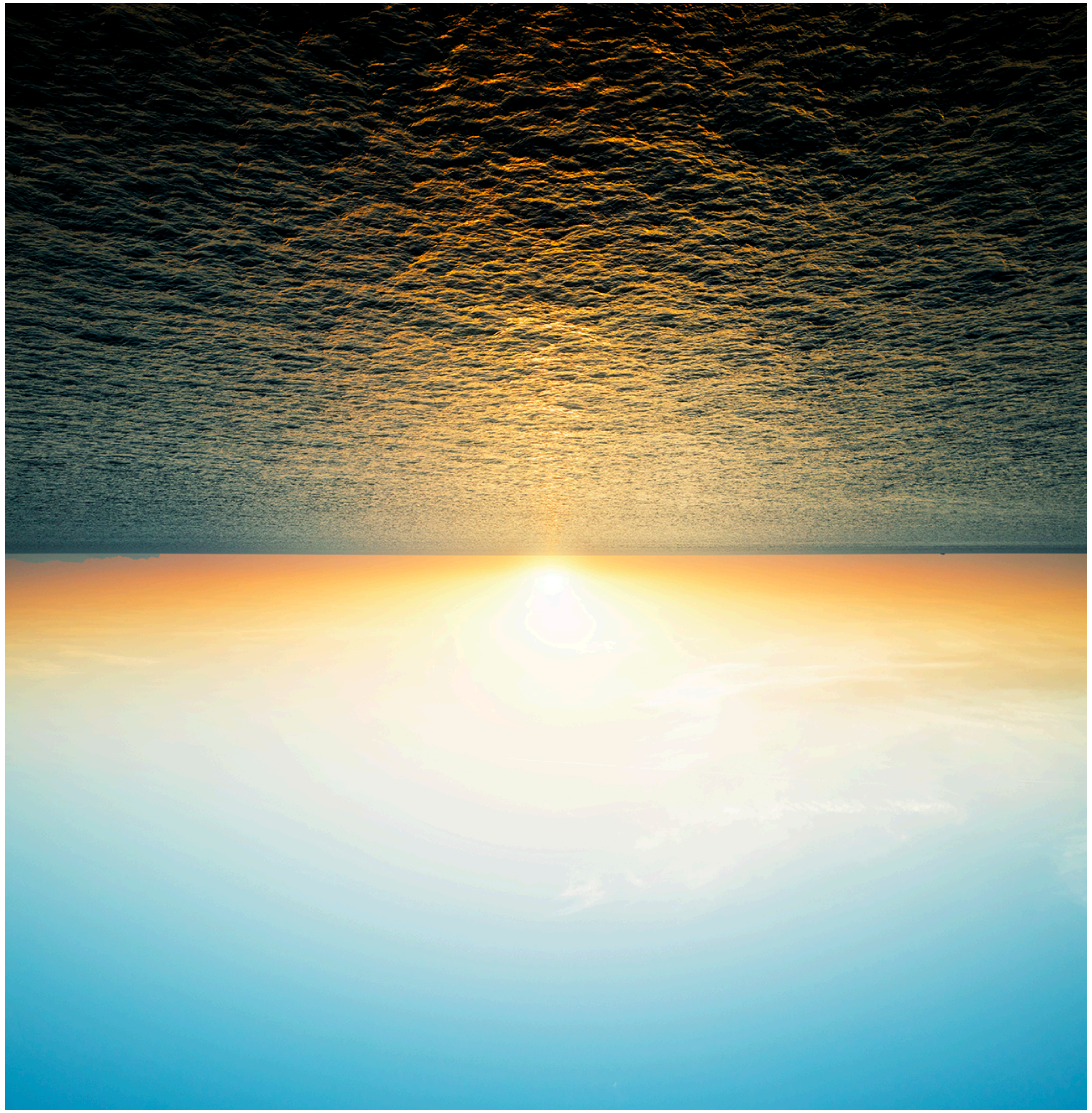
(1) a statistic

(2) measures of central tendency

(3) sample and population

(4) parameters and estimates

(5) confidence interval



Statistical modeling

(1) Statistical modeling is based on **optimization and simulation**

(2) While it is important to know optimization techniques, this is a topic of other courses

(3) In this course, we will study various techniques to estimation of parameters, also by means of resampling and simulation

The essence of Statistics

(1) Statistics solves a **backwards problem**

(2) It starts from data (observe) and then asks **what was used to generate the data**

(3) With statistics we can mathematically quantify predictions

(4) And also **helps to quantify our uncertainty**

Descriptive Statistics

(1) **Descriptive statistics** enables us to present the data in a meaningful way, which allows simpler interpretation of the data.

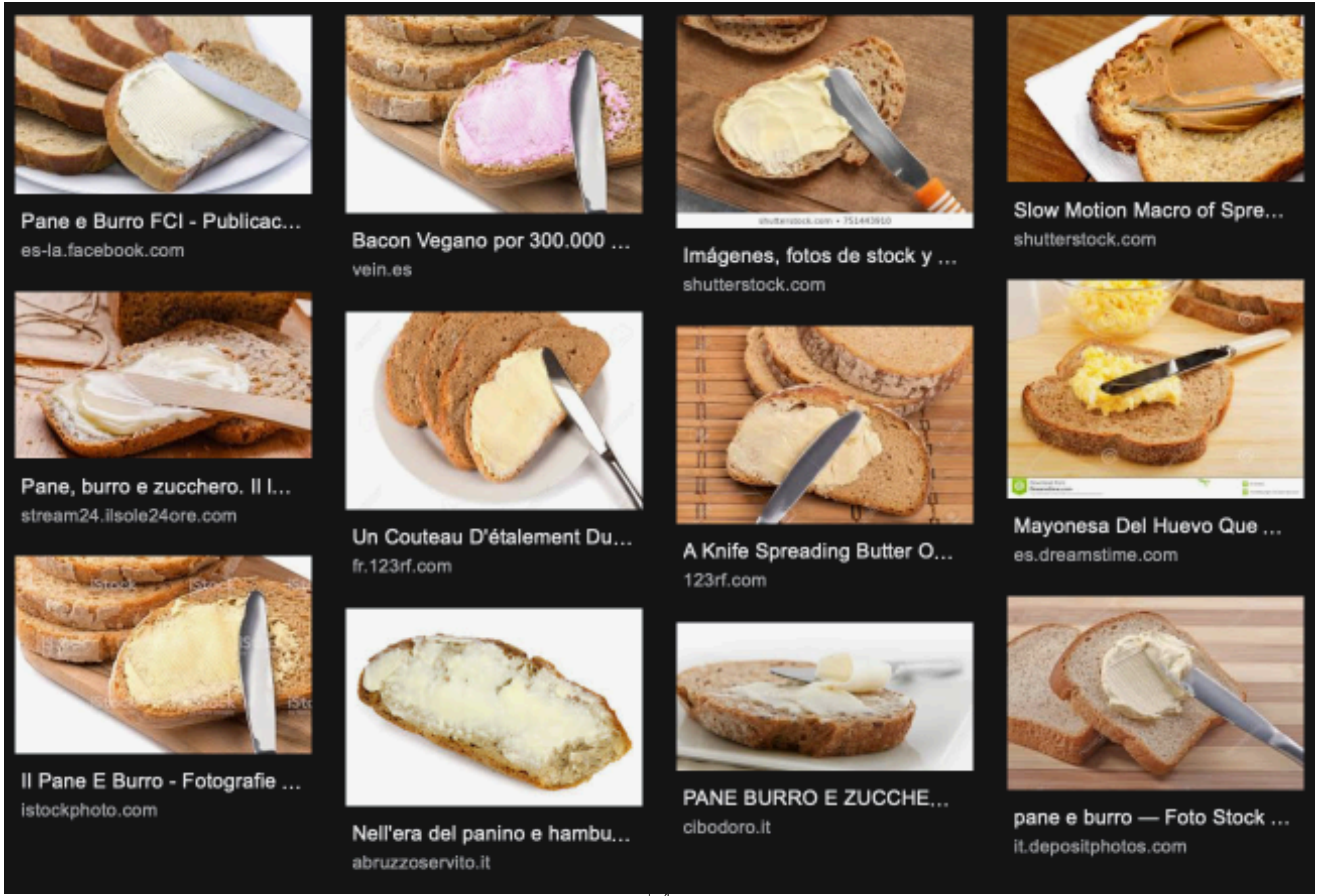
(2) Typically, two general types of statistic that are used to describe data:

(1) Measures of central tendency

(2) Measures of spread

Measuring the Central Tendency

Measuring the Spread



Inferential Statistics

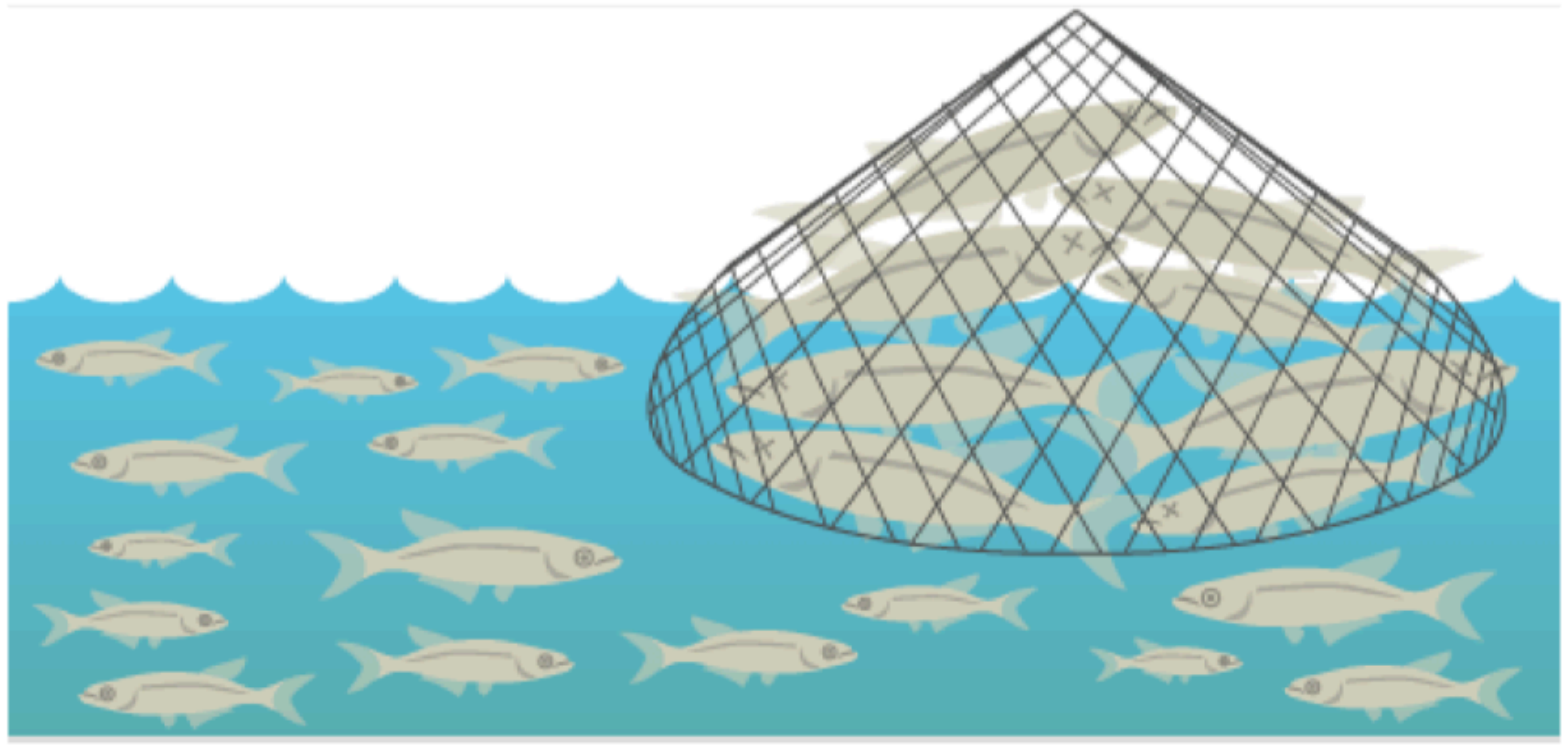
(1) **Inferential statistics** are techniques that allow us to use samples to make generalizations about the populations from which the samples were drawn.

(2) It is, therefore, important that the sample accurately represents the population.

Data and Estimators

Sample and Population

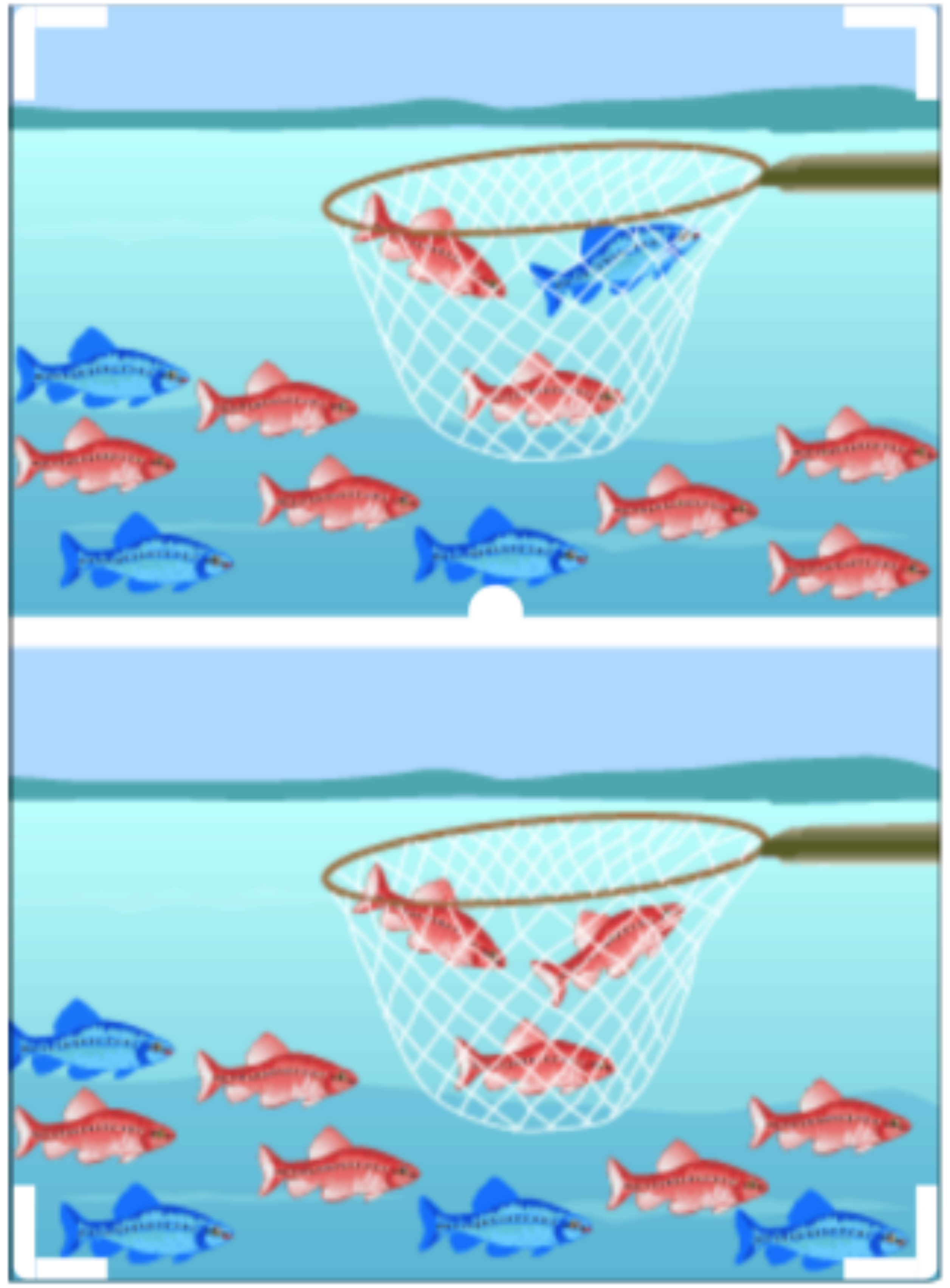
(1) In statistics, **a population** is all of the elements in a group and **a sample** is a part of a population chosen to represent the entire population.



Samples

(1) A sample should represent main properties of population (that are investigated in the research / analysis)

(2) Reliable statistical analysis deals with **representative samples.**



Simple Sample, Sample Size

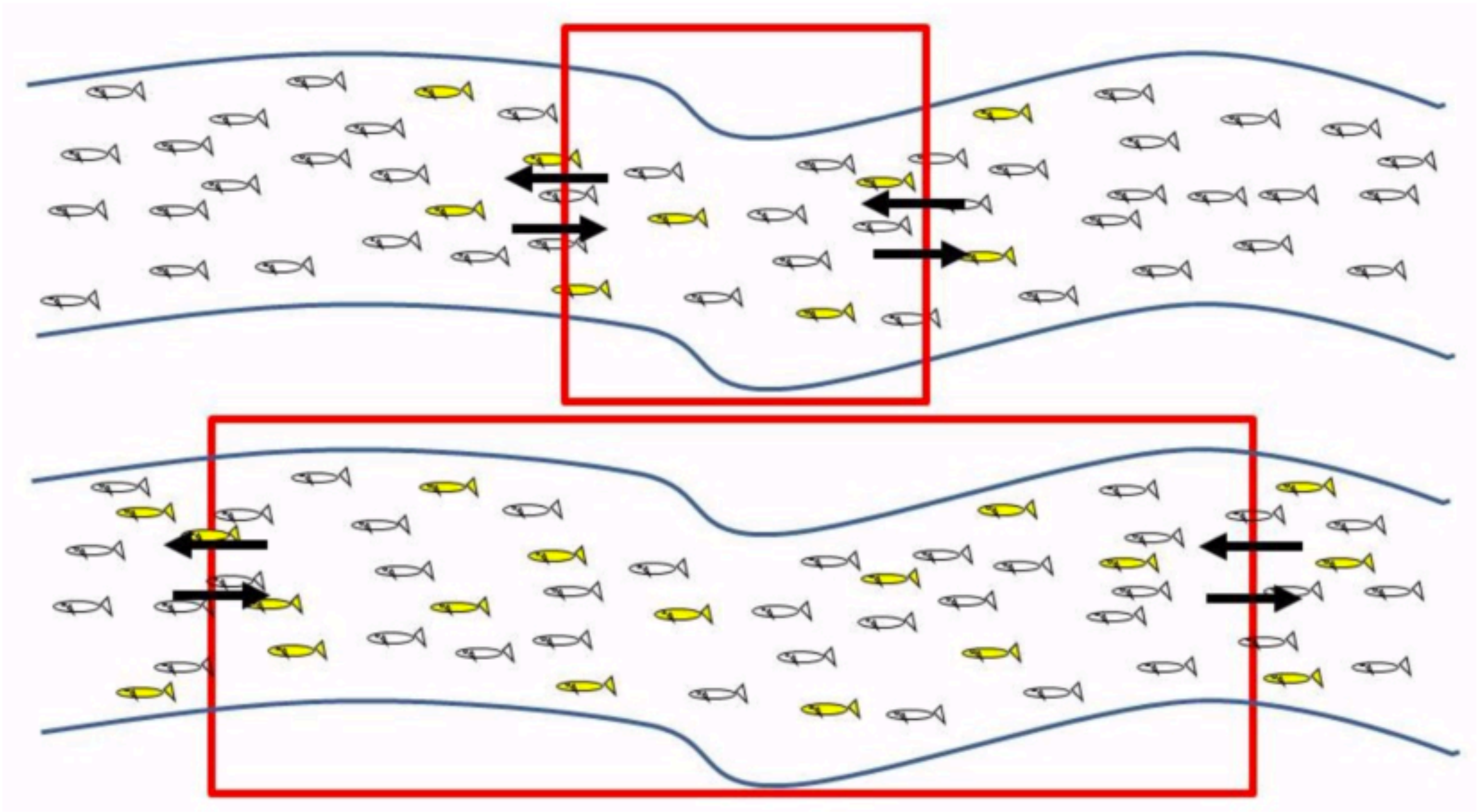
(1) X^n=$(X_1, \ldots, X_n)$ is a simple sample,

(2) if $X_1, \ldots, X_n$ are **independently identically distributed** (i.i.d.) random variables.

(3) Each X_i has the same density function f(x).

$$X^n = (X_1, \ldots, X_n).$$

n – sample size



Statistic. Definition

(1) A statistic is a function of a sample: $T(X^n)$

(2) Sample Mean:

(3) Sample Variance:

Let us consider at two examples:

Statistic or not?

(1) Example 1

(1) If $X_1 \sim N(\mu, \sigma^2)$, then

(2) $Y = (X_1 - \mu)/\sigma$ has a standard normal distr. ($Y \sim N(0,1)$)

(3) Y has dependency on parameters, because the functional form $(X_1 - \mu)/\sigma$ includes μ, σ

(4) Here we **have dependency** on parameters of the initial normal r.v. (X_1)

(2) Example 2

(1) If X has two values $X \in \{0,1\}$ and has a p.d.f. $f(x) = p^x(1-p)^{(1-x)}$

(2) Let $Y = \sum_{i=1}^{n} X_i$; then $Y \sim binomial(n, p)$

(3) Here we **do not have any dependency** on parameter **p** of the original Bernoulli,

because $\sum_{i=1}^{n} X_i$ deos not include **p**

A statistic. Definition

(1) Definition:
A function of one or more random variables that does <u>not</u> depend upon any unknown parameter is called a <u>statistic</u>.

(2) Question: Is statistic a random variable?
Spoiler: **Yes.**
Hence a statistic (as r.v.) has a distribution

(3) Note:

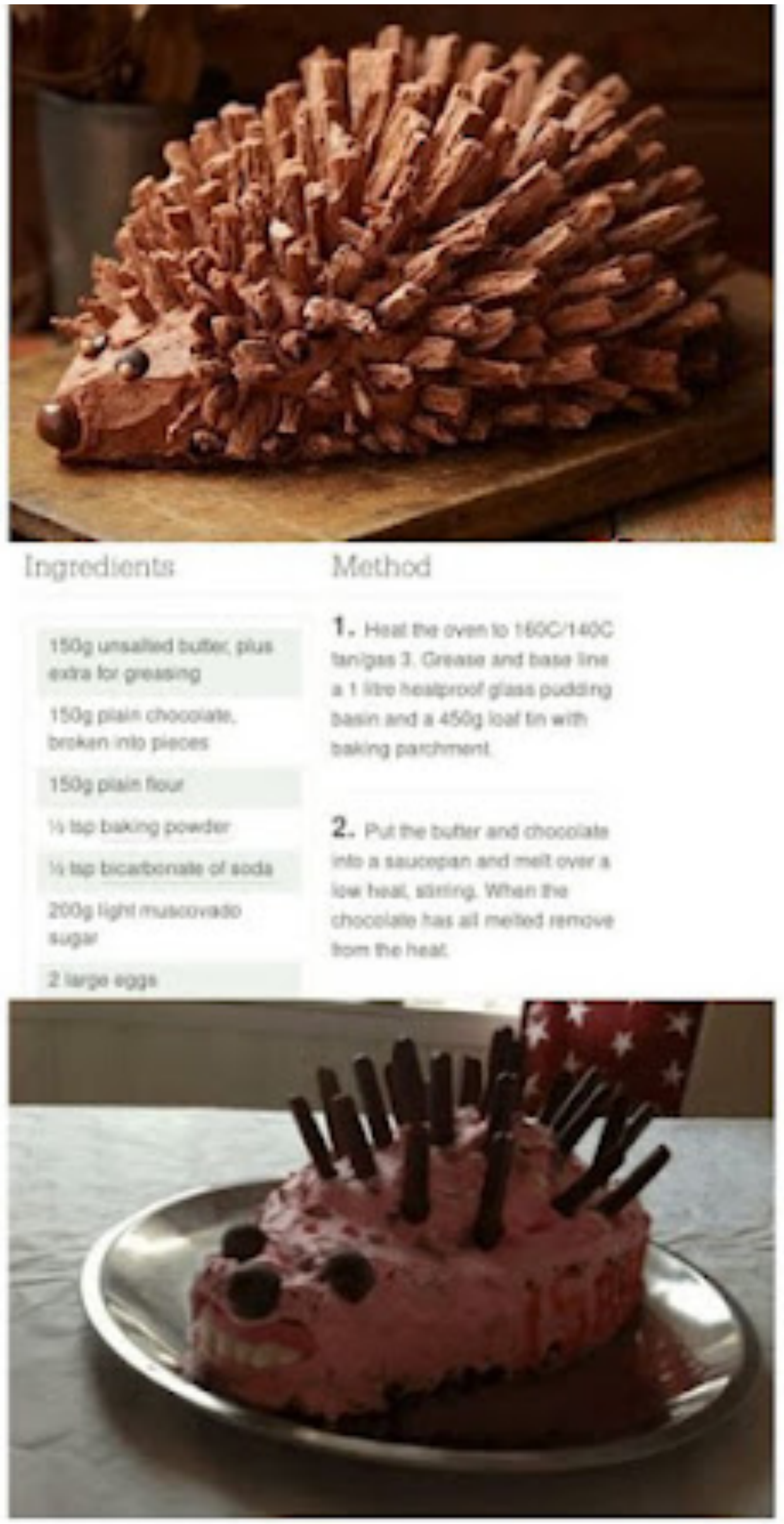
  (1) Important that although a statistic does not depend upon any unknown parameter, the <u>distribution</u> of that statistic may very well depend upon unknown parameters

Estimation

(1) **Inferential statistics** is focused on the estimation of the **population parameter** from **the sample statistic**.

(2) The **sample statistic is calculated** from the sample data and the **population parameter is inferred** (or estimated) from this sample statistic.

(3) **Again!** Statistics are calculated, parameters are **estimated**.



estimand

estimator

estimate

http://onbiostatistics.blogspot.com/2022/04/estimands-estimator-estimate-and.html

Point Estimates

The point estimate is the single best guess about the value of parameter

(1) A good estimator must satisfy three conditions:

   (1) **Unbiased**: The expected value of the estimator must be equal to the value of the parameter

   (2) **Consistent:** The value of the estimator approaches the value of the parameter as the sample size increases

   (3) **Relatively Efficient:** The estimator has the smallest variance of all estimators which could be used

Unbiased estimator. Example

(1) The expected value of the estimator must be equal to the value of the parameter

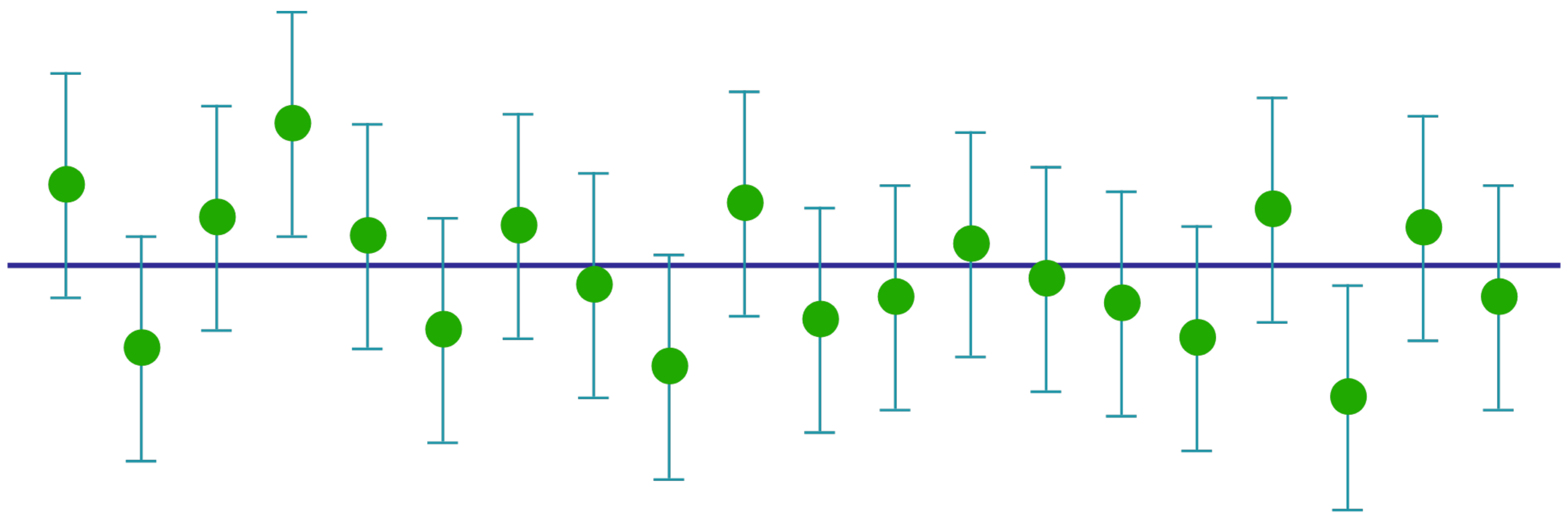
(2) Check whether $\bar{X}$ is unbiased or not

(3) Check whether $Var[X]$ is unbiased or not

Interval Estimates

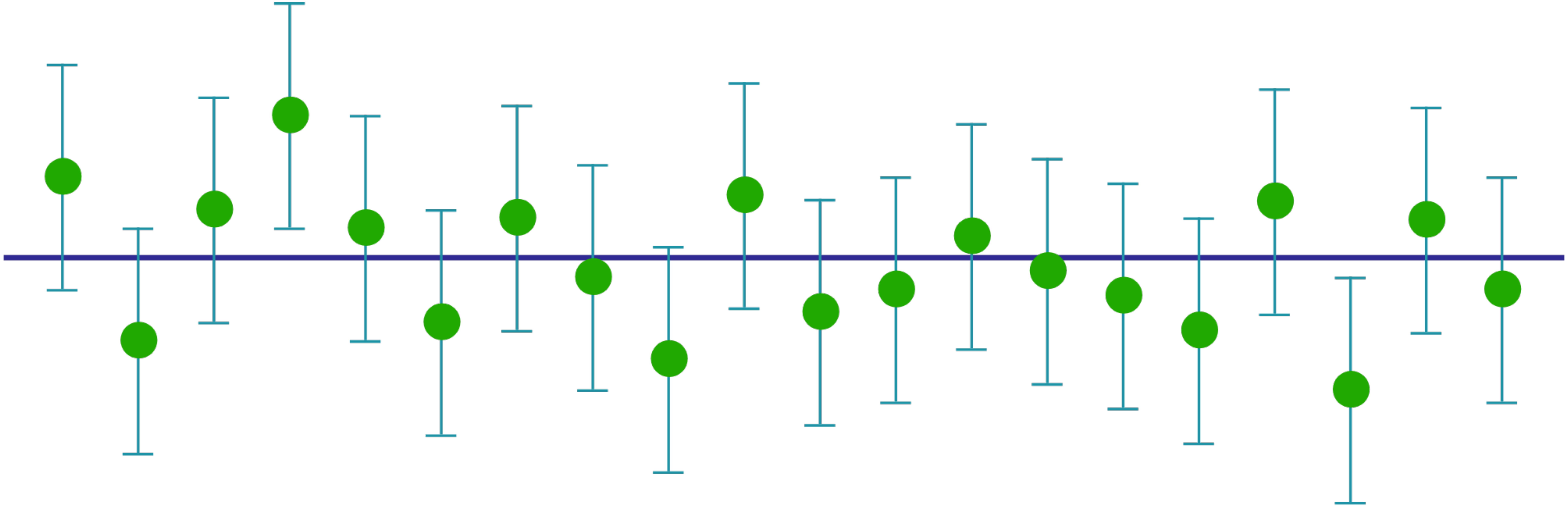
(1) A **<u>confidence interval</u>** contains the true value of the corresponding parameter with the specified probability

(2) Informally, if you run 100 experiments a 95%-confidence interval will contain the value of parameter 95 times (we will discuss it later in the course)



Confidence interval

What is random here? Where is the probability?



Properties of Parameter Estimates

Mean squared error: MSE

$$MSE(\hat{\lambda}) = \mathbb{E}(\hat{\lambda} - \lambda)^2 = Var(\hat{\lambda}) + [\mathbb{E}(\hat{\lambda}) - \lambda]^2$$