Statistical Techniques for Data Science (& Robotics)

Weeks 1 & 2

Objectives (for today)

(1) to learn about the course, including

- How to pass / fail
- How to get an “A”
- connections to 5+ courses from your program

(2) to recall Probability theory

(3) to do short recall basics of Statistics

Team and Communication

(1) PI: Vladimir Ivanov

- @nomemm
- room 475

(2) Teaching Assistants:

- Zamira Kholmatova
- Vladimir Bazilevich
- Zlata Shchedrikova

Syllabus

briefly

(1) “Classical” Statistics: Tests, Hypotheses

(2) Statistical view at Machine Learning models

(3) Non-parametric Statistics

Midterm

(4) Bayesian Statistics

(5) Sampling, MCMC

(6) Bandit algorithms

Structure

(1) **12-14** Lectures (2 hours/week) + Quizzes (10 min. on a lecture)

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

(3) **12-14** Labs (2 hours/week)

(4) **1** Case Study instead of the Midterm

(5) **1** Final Exam

(6) **0...n** retakes

Books

(1) Bruce, Peter,Bruce, Andrew. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media. — Simple and short, an **Intro-level book**

(2) Bishop Christopher. Pattern Recognition and Machine Learning. Springer, 2006. — 738 p. — **CoreBook, hard, but worth**

(3) Introduction to Mathematical Statistics. By ROBERT V. HOGG AND ALLEN. B. CRAIG 4th Edition — **MathStats book**

(4) Probability & statistics for engineers & scientists/Ronald E. Walpole . . . [et al.] — 9th ed. — **your ProbStat book (from the Fall semester)**

Grading

(1) Assignments: **30** % (aka Homeworks)

(2) Midterm : **30** % (Case study)

(3) Final Exam: **30** % (Written + Oral)

(4) Labs (grading during labs): **10 pts** (0, 1/2, 1 scale; max 1 point per lab)

86 - 100	A
70 - 85	B
55 - 69	C
0 - 54	D

Tools

(1) Pen and Paper

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

Course Prerequisites

(1) Linear Algebra, Mathematical analysis courses

(2) A course on Probability Theory

How to success?

(1) Assignments and Labs:

- work hard (individually) + office hours
- visit Labs to have enough practice with tools
- visit Labs to solve pen and paper problems

(3) Exam:

- read the books + office hours
- solve exercises from books at home

Discussion

(1) Which courses this course is connected to?

Break, 5 min.

Example

Fitting of a curve

Recap of Probability Theory

Major Concepts

What is …?

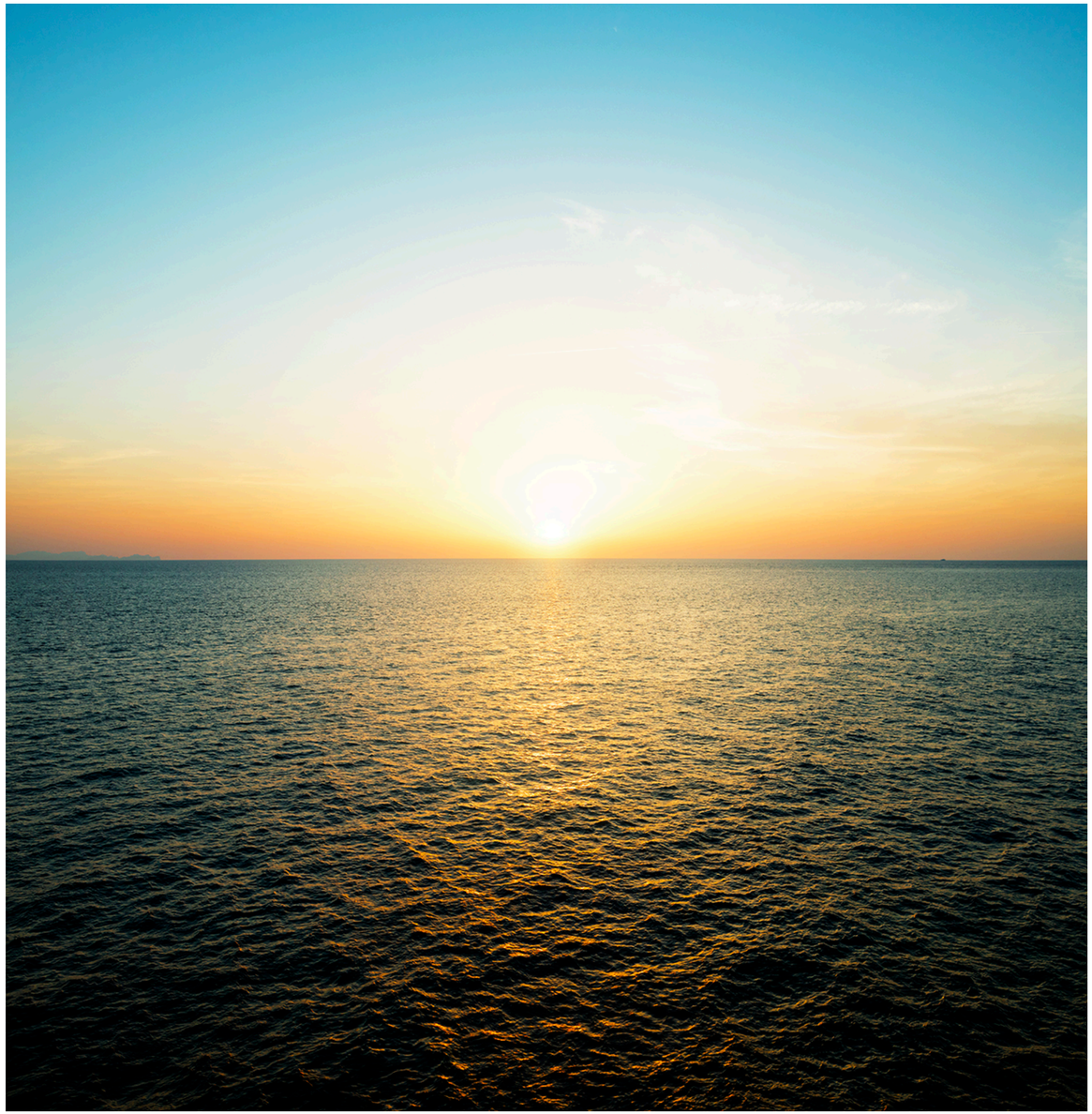
(1) Random Variable

(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

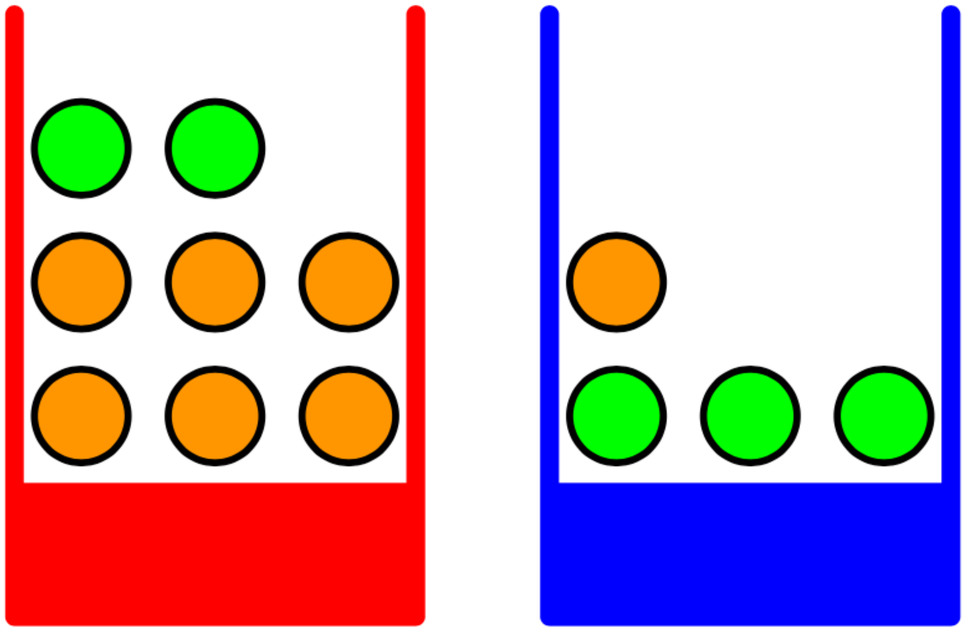
F. Number of winings from a game with a $1 gain/loss for each head/tail coin flip in a series of 10 flips

2 Rules of Probability

Apples and Oranges

(1) A - fruit type

(2) B - box color



- Q: probability of …
- given that we have chosen an orange, what is the probability that the box we chose was the blue one?

General Case

Sum rule

(1) N is a number of all experiments

(2) n_{ij} is a number of trials when

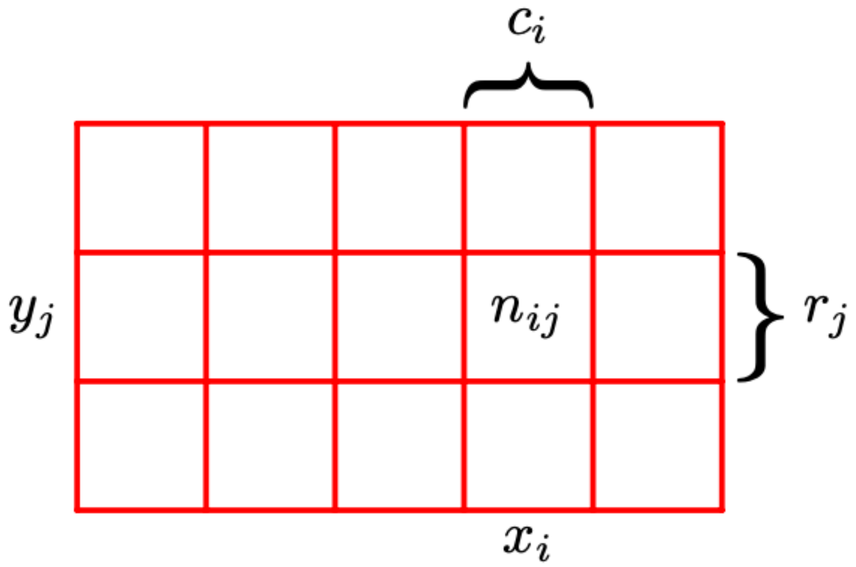
- $X = x_i$ and $Y = y_j$

(3) c_i is the sum of i-th column

(4) $P(X=x_i) = \ldots$

(5) Sum rule:

- $P(X = x_i) = \ldots$



General Case

product rule

(1) n_{ij} is a number of trials when

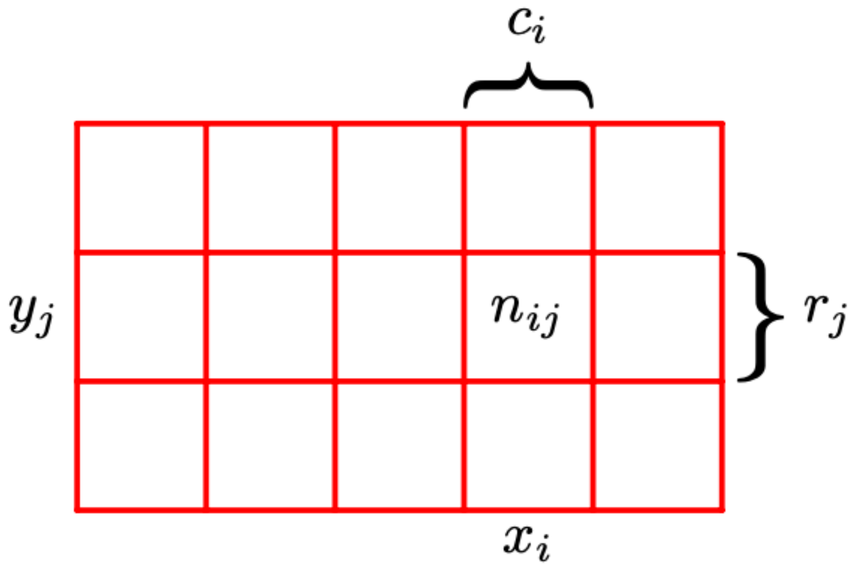
- $\mathrm{X} = x_i$ and $\mathrm{Y} = y_j$

(2) c_i is the sum of i-th column

(3) $\mathrm{P}(\mathrm{X}=x_i) = c_i/N$

(4) Product rule:

- $\mathrm{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = \ldots$



2 Rules of Probability

sum rule $$p(X) = \sum_{Y} p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X).$$

Probability Distributions (of r.v.)

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

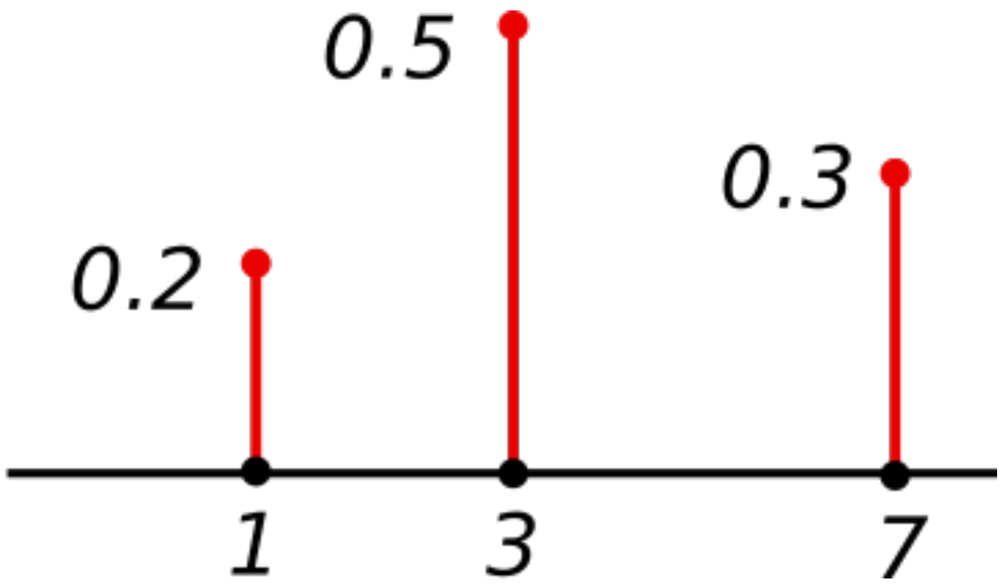
(2) **its probability distribution** assigns a likelihood to each of its (r.v.) values.

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

- discrete
- 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

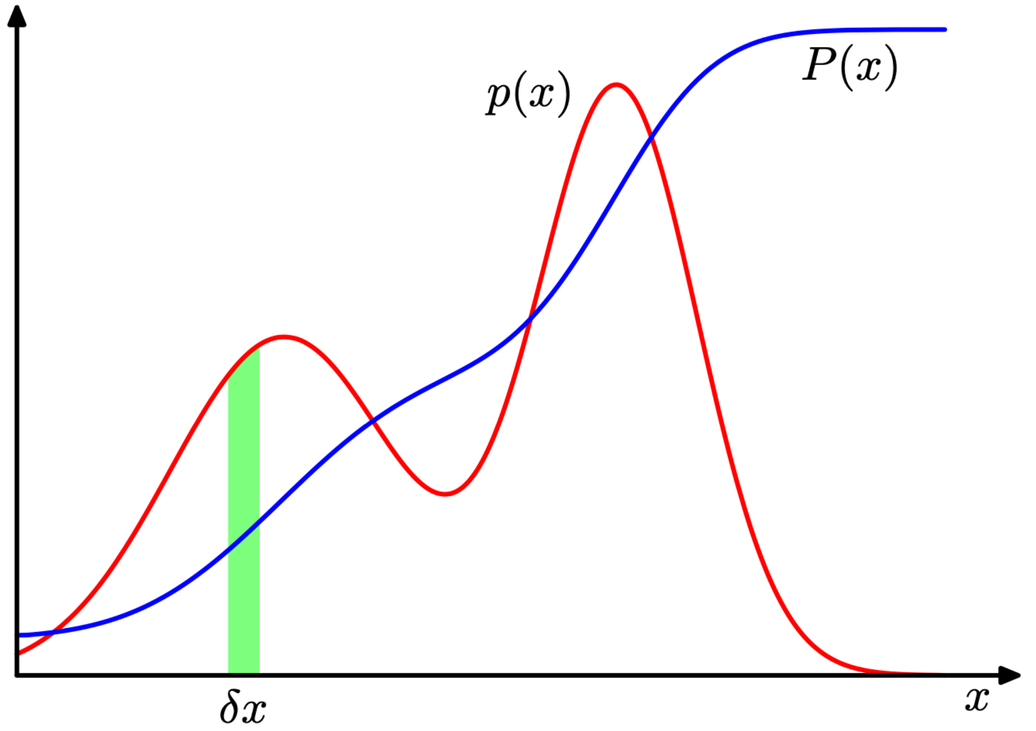
$$P[a < X \leq b] = \int_a^b f_X(x)\,dx$$

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

Question

$$p(x) \geqslant 0$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$



What about the range of a pdf $p(x)$?

Cummulative distribution function

CDF

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

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

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

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

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:

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

S is range of random variable X

Expected value and its Properties

- $E[X] = \sum xp(x)$
- $E[X] = \int xp(x)dx$
- $E[X + c] = E[X] + c$
- $E[cX] = cE[X]$
- $E[X + Y] = E[X] + E[Y]$

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 $p(x)$

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

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

HOMEWORK exercise: Let X be a continuous random variable with density function $p(x) = 2x; 0 < x < 1$. What is $E[X \mid X < \frac{1}{2}]$?

Distributions

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

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

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

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

$$\lambda = kt$$

$$E[X] = ?$$

$$Var[X] = ?$$

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 $\frac{1}{a}$?

About Normal Distribution

Univariate case

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$\beta = \frac{1}{\sigma^2}$ is a precision

Normal Distribution

Multivariate case (aka D-dimensional vector r.v.)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Σ is a $D \times D$ covariance matrix

Bayesian Probabilities

(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

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

Alternative view

(1) a model, with parameters $\mathbf{w}$

(2) data observed: $\mathbf{D}$

(3) Likelihood of data, given a model, **likelihood function**: $p(\mathbf{D} \mid \mathbf{w})$

(4) **Prior** probability : $p(\mathbf{w})$

(5) **Posterior** probability $p(\mathbf{w} \mid \mathbf{D}) =$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

PAUSE

Statistical Techniques for Data Science (& Robotics)

Week 2

Quiz

10 min

Q1: 2 Rules of Probability

Q2: Given the dataset **D** and parameters of a model **w** one can compute the posterior distribution:

$$p(w \mid D) = \frac{p(D \mid w)p(w)}{p(D)}$$

Write the expression for the $p(D)$ in case when **w** has a discrete distribution.

Objectives

(1) Intro to Statistics

(2) Estimators

(3) MLE

Alternative view

(1) a model, with parameters $\mathbf{w}$

(2) data observed: $\mathbf{D}$

(3) Likelihood of data, given a model, **likelihood function**: $p(\mathbf{D} \mid \mathbf{w})$

(4) **Prior** probability : $p(\mathbf{w})$

(5) **Posterior** probability $p(\mathbf{w} \mid \mathbf{D}) =$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Introduction to statistics

Major Concepts

What is …?

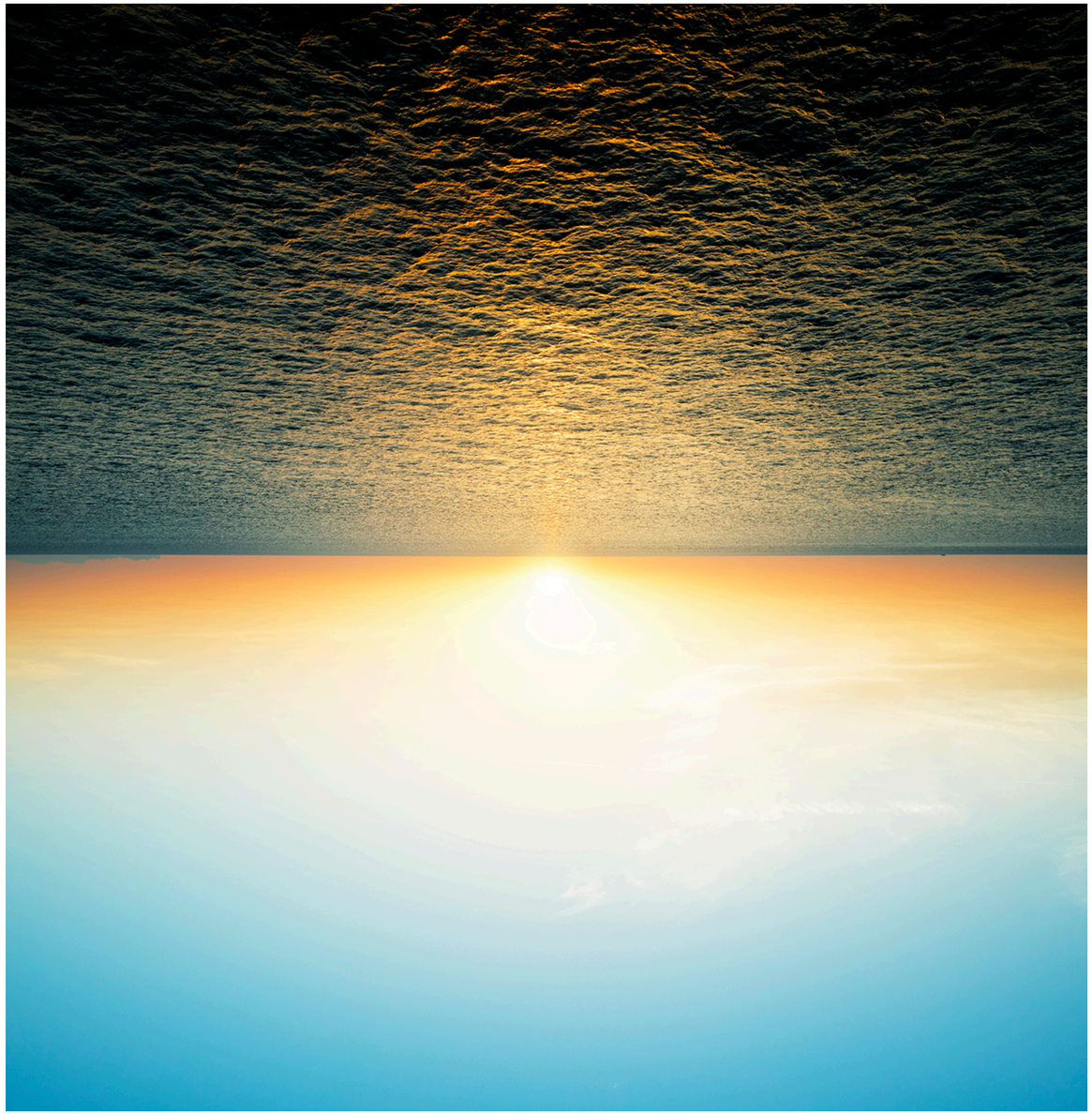
(1) a statistic

(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

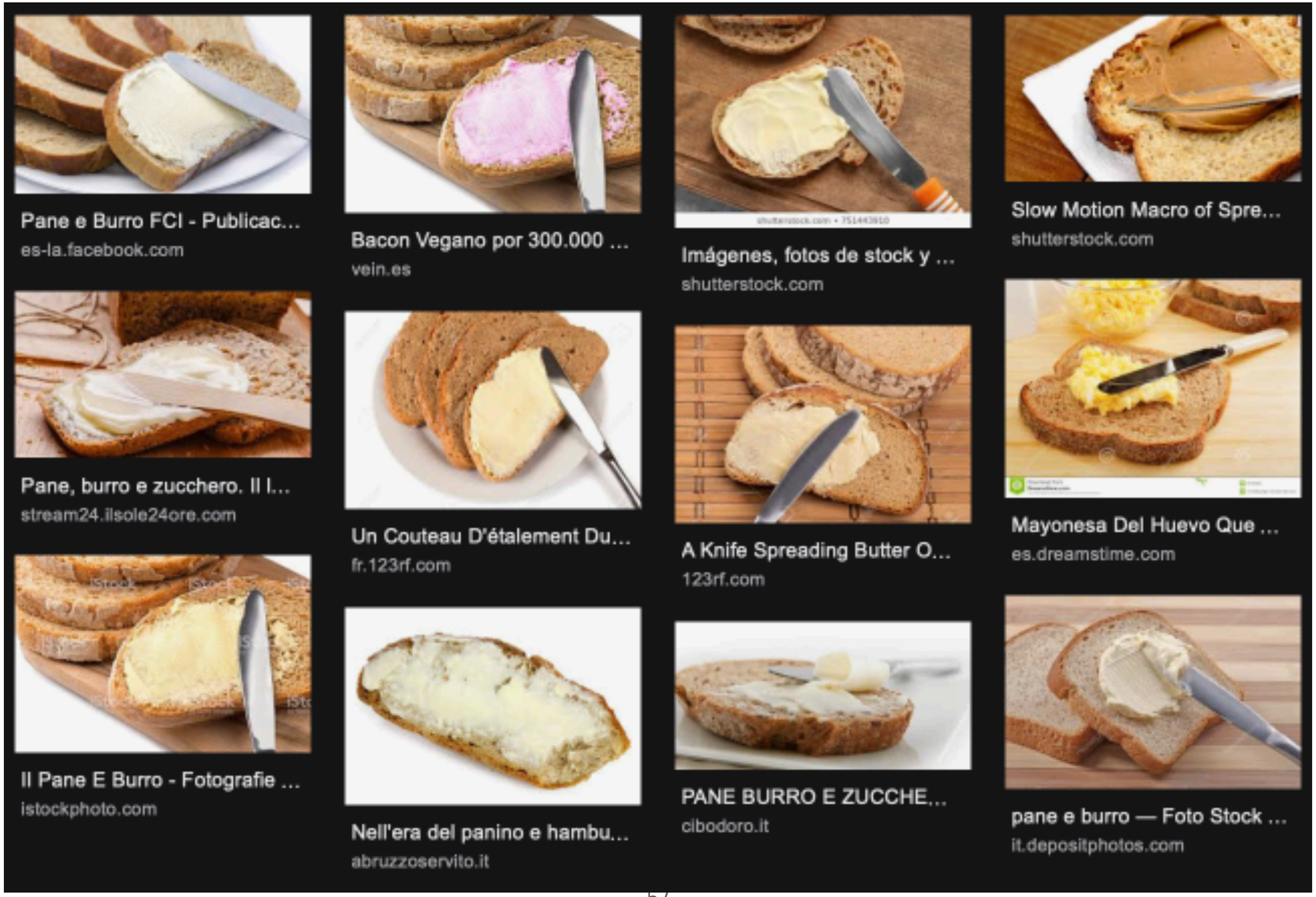
(1) **Descriptive statistics** enables us to present the data in a meaningful way, which allows simpler interpretation of the data.

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

- **Measures of central tendency**
- **Measures of spread**

Measuring the Central Tendency

Measuring the Spread



Inferential Statistics

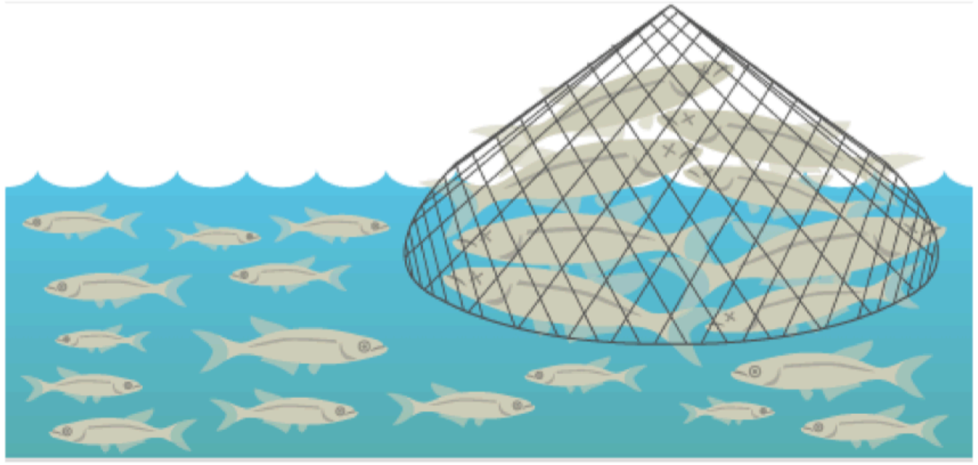
(1) **Inferential statistics** are techniques that allow us to use samples to make generalizations about the populations from which the samples were drawn.

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

Data and Estimators

Sample and Population

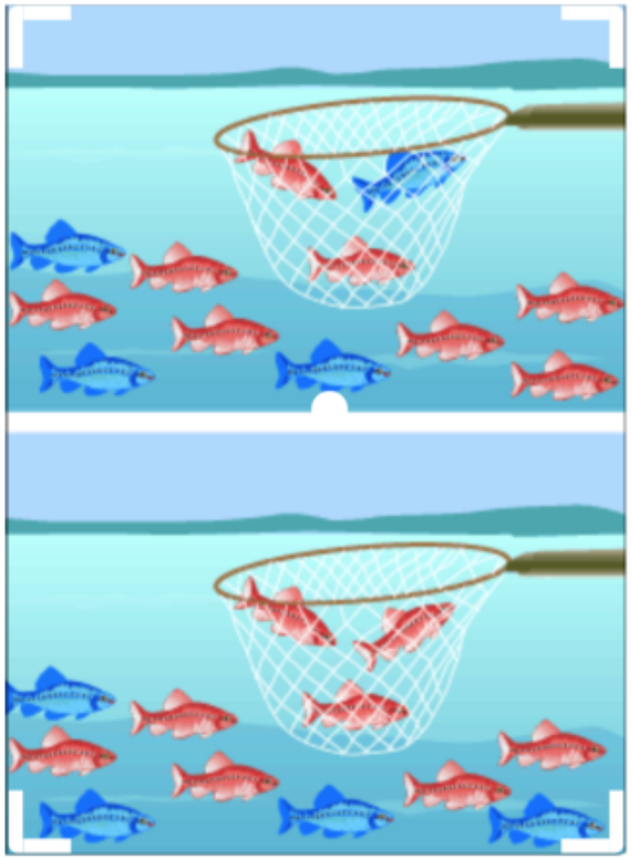
(1) In statistics, **a population** is all of the elements in a group and **a sample** is a part of a population chosen to represent the entire population.



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

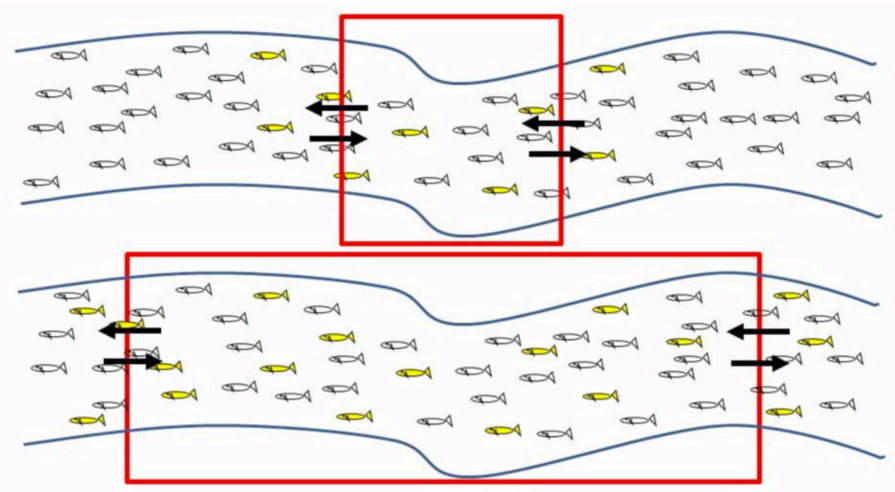
(1) $X^n=(X_1,\ldots,X_n)$ is a simple sample,

(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:

A statistic. Definition

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

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

(3) Note:

- Important that although a statistic does not depend upon any unknown parameter, the distribution of that statistic may very well depend upon unknown parameters

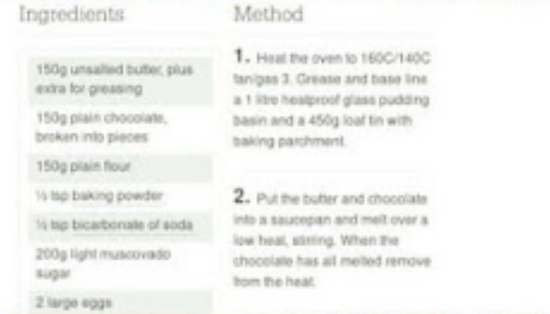
Estimation

(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:

- **Unbiased**: The expected value of the estimator must be equal to the value of the parameter
- **Consistent:** The value of the estimator approaches the value of the parameter as the sample size increases
- **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

Maximum Likelihood Estimation

Maximum likelihood estimation: MLE

(1) We have a sample $X^n = (X_1, X_2, \ldots, X_n)$

(2) all x_i are independently and identically distributed (i.i.d.)

(1) with respect to a p.d.f. $f(x; \lambda)$,
where λ is a parameter

(3) The method searches among all distributions to find the one that places the highest chance on the observed data.

MLE: Example of the Poisson distr.

(1) p.d.f. $f(x,\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$, where $x \in 0,1,2,\ldots$

(3) Likelihood function for a random sample

$$\frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \times \cdots \times \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{\lambda^{\sum x_i}}{\prod x_i!}e^{-n\lambda} = L(\lambda)$$

(9) MLE -> maximize likelihood function over set of parameters

MLE: Example

(1) maximize likelihood function -> maximize the logarithm of likelihood function

$$\begin{aligned}
\frac{\partial l}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left[\sum x_i \log(\lambda) - n\lambda - \sum \log(x_i!)\right] \\
&= \sum x_i/\lambda - n, \\
\frac{\partial^2 l}{\partial \lambda^2} &= -\sum x_i/\lambda^2.
\end{aligned}$$

Exercise: pen&paper

(1) Exponential distribution with unknown parameter θ

(2) p.d.f.: $f(x;\theta) = (1/\theta^2)xe^{-x/\theta}, \quad 0 < x, \theta < \infty$

(3) Given a sample: $X = (x_1, x_2, \ldots, x_n)$

(4) Find for θ using MLE

Convergence and CLT

Towards the Central Limit Theorem

(1) The formal statement requires a definition of "converging in distribution"

(2) A sequence of random variables X_n converges in distribution to a random variable Z if

$$\lim_{n\to\infty} P(X_n < x) = P(Z < x)$$

(4) for any real number x at which the function $P(Z \leq x)$ is continuous.

Towards the Central Limit Theorem

Let $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with the same probability density function f, mean μ and variance σ^2.

Then the mean of the sample is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Towards the Central Limit Theorem

Let $X_1, \ldots, X_n$ is a sequence of i.i.d. random variables with the same probability density function f, mean μ and variance σ^2.

Then the mean of the sample is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Question: The **mean** of the sample is also a random variable. How is it distributed?

The Central Limit Theorem (Classical formulation)

(1) Let X_i be i.i.d. random variables with **parameters:**

  (1) $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ (this holds for each X_i)

(2) Let

  (1) $\bar{X}(n) = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$

(3) Then the variables defined as $Y_n = \sqrt{n}(\bar{X}(n) - \mu)$

(4) C**onverge in distribution** to $Z \sim N(0,\sigma^2)$ the normal distribution with mean 0 and variance σ^2

Demo

(1) check:
http://onlinestatbook.com/stat_sim/sampling_dist/

Application to Estimation of Error / Variability

(1) Bernoulli:

  (1) a trial with “0” or “1” outcomes

  (2) p is a parameter (a probability of success)

(2) We know that parameters are $\mu = p; \sigma^2 = p(1-p)$

(3) But can we estimate the p as a function of sample?

  (1) $\hat{p} = ?$…. well, it is just mean,

  (2) and we can find $\hat{\sigma}^2$…

Application of the CLT

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Application of the CLT

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to CLT (approximately) $Z \sim \mathcal{N}(0, 1)$:

$$Z = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

Application of the CLT:

For Bernoulli distribution

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to CLT (approximately) $Z \sim \mathcal{N}(0,1)$:

$$Z = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$$

- Derive basic 95% Confidence Interval for p using CLT

$$\bar{X} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \bar{X} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Estimation result

(1) So, just by tossing a coin many (n) times we can build upper and lower bounds for the probability of HEAD:

$$\bar{X} - z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \bar{X} + z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

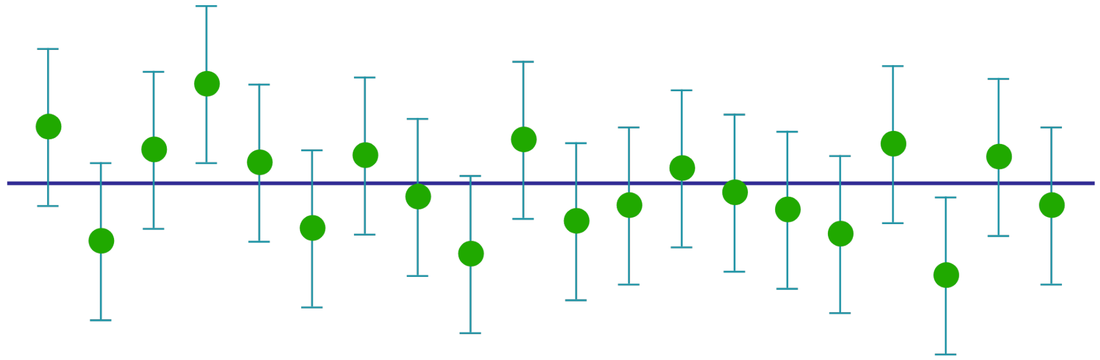
(1)

The lower and upper bounds are RANDOM VARIABLES !
(as they depend on a random sample)

Interval Estimates

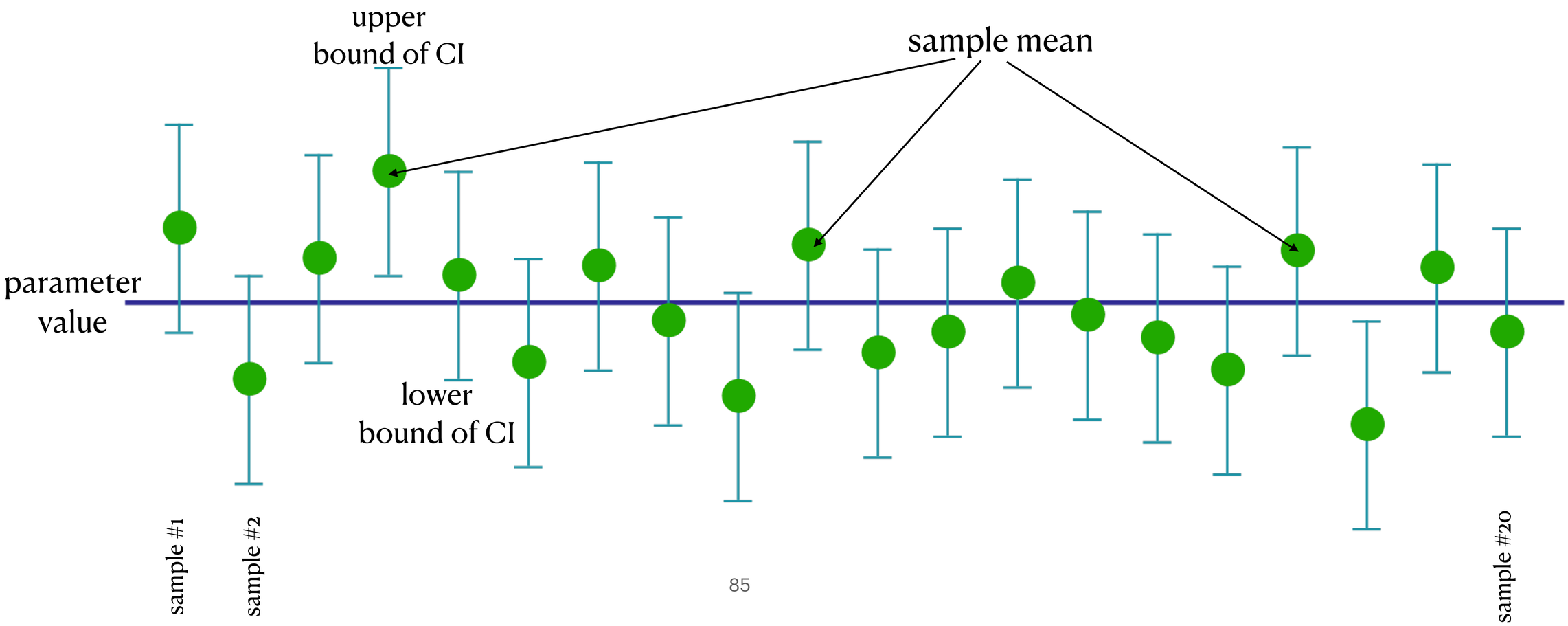
(1) A **confidence interval** contains the true value of the corresponding parameter with the specified probability

(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?



final meme