Tutorial 2

Biological Data Analysis

Spring 2023

Outline

- Numpy
- Population and sample
- Parameter vs estimator vs statistic
- Matplotlib / Presentation
- Distribution as a model

Basic concepts in statistics

- **Population**: a “pool” from which we take a **sample**.

The group is large enough to allow a independent observation.

We have no access to the entire population but want to describe it using certain parameters.

We use sample to find the parameters

- The goal of statistical analysis: to infer the population parameter using sample

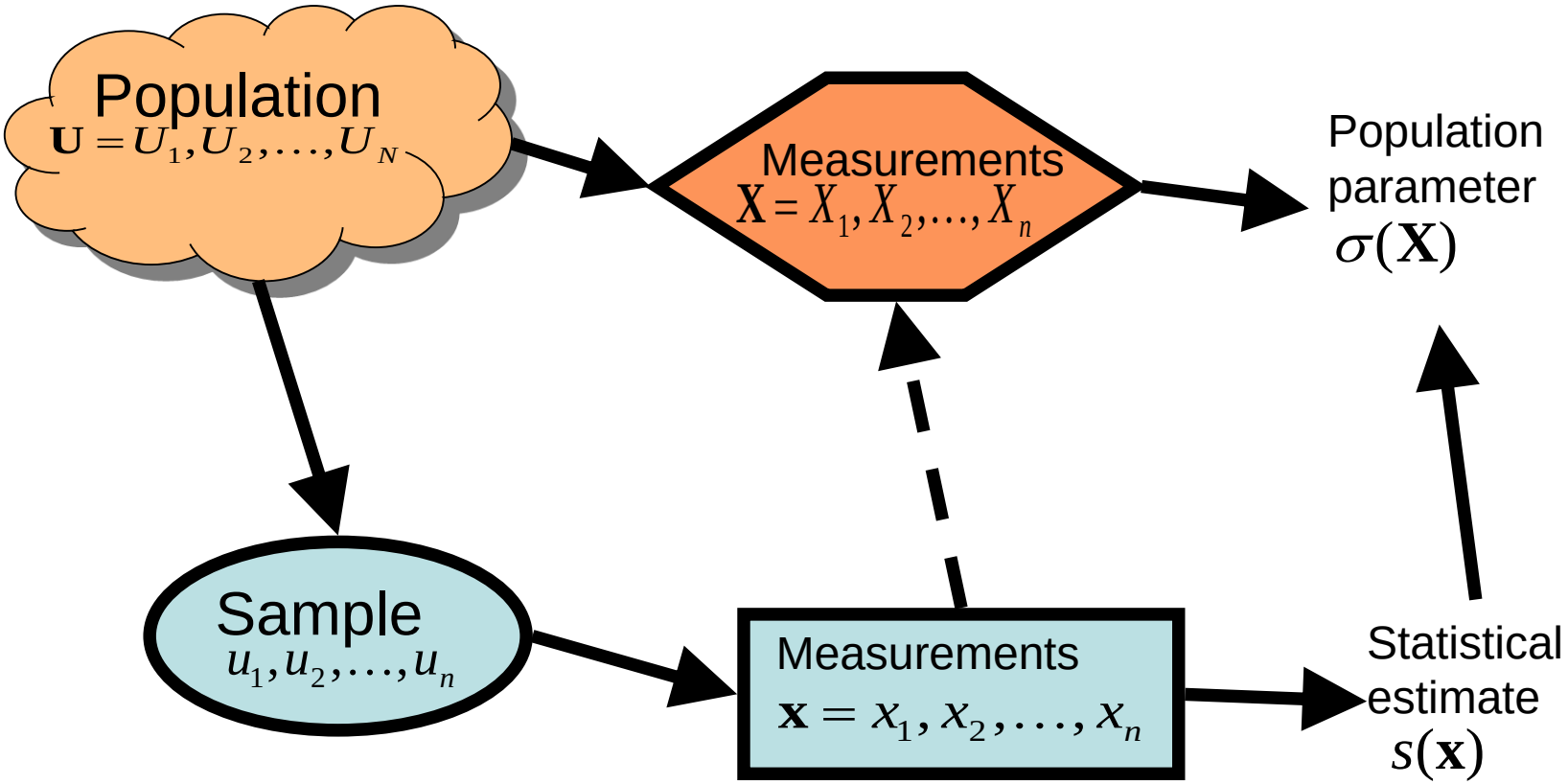
Basic concepts in statistics

- **Parameter**: numerical value that describes a population. It is fixed and unknown, but can be estimated

- **Statistic**: numerical value that describes a sample. It varies from sample to sample and can be calculated from sample data

- **Estimator**: estimator is a formula or rule that uses sample data to estimate an unknown parameter. Provides a best guess of the true value of the parameter, based on the available data

Population and sample



Donchin, Biological data analysis lecture

Good sample

- Representative
- Independent
- Identically distributed

Types of variables: numerical, categorical, ordinal

Example 1

A study conducted on 200 cats in Beer Sheva found that they sleep 14 hours a day.

What is sample in this study?

What is population?

Example 2

Are the following observations independent?

Number of pathologies found in a number of biopsies taken from one tissue from the same individual?

If the population is all cells in the body (different tissues) of the individual – observations are not independent

If the population is all cells in this specific tissue – observations are independent

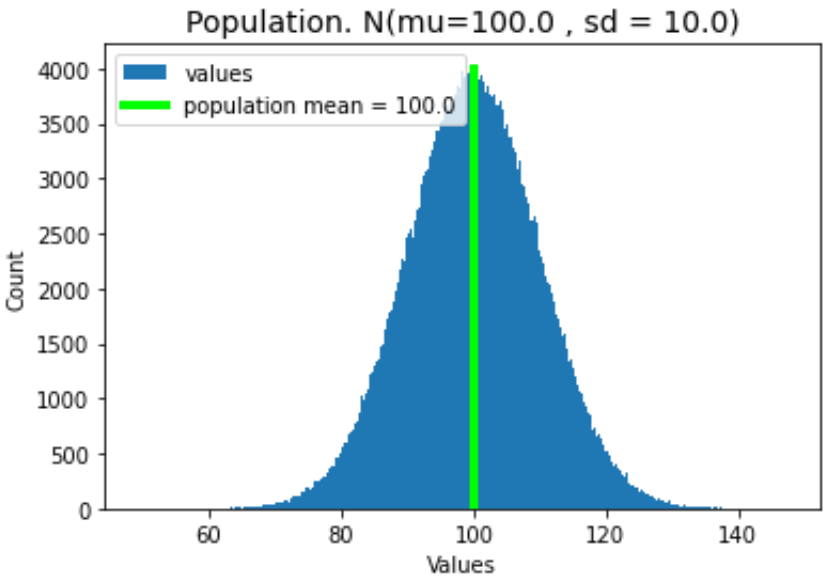
Parameter vs Estimator

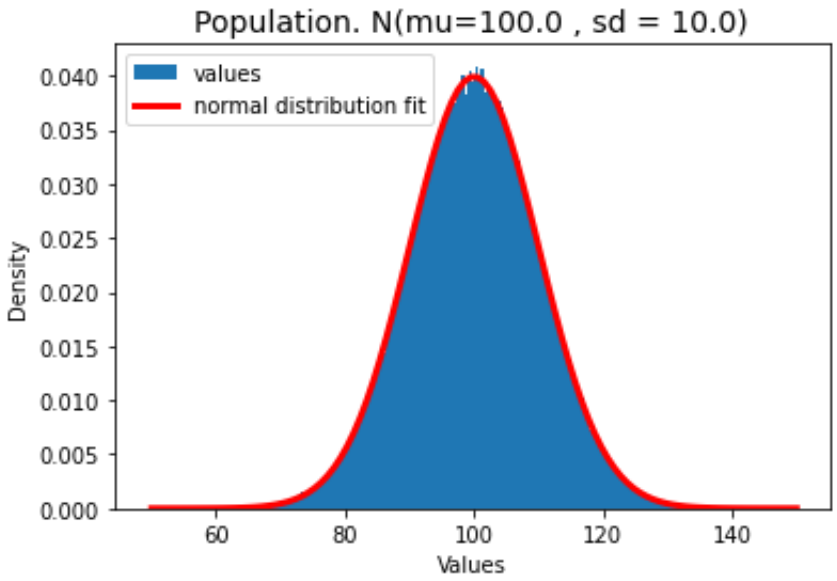
- **Parameter – a value that describes a population**
  - Not a random variable

- **Estimator – random variable that depends on a sample**
  - Used to estimate population parameter
  - Calculated using sample statistic

Sampling simulation

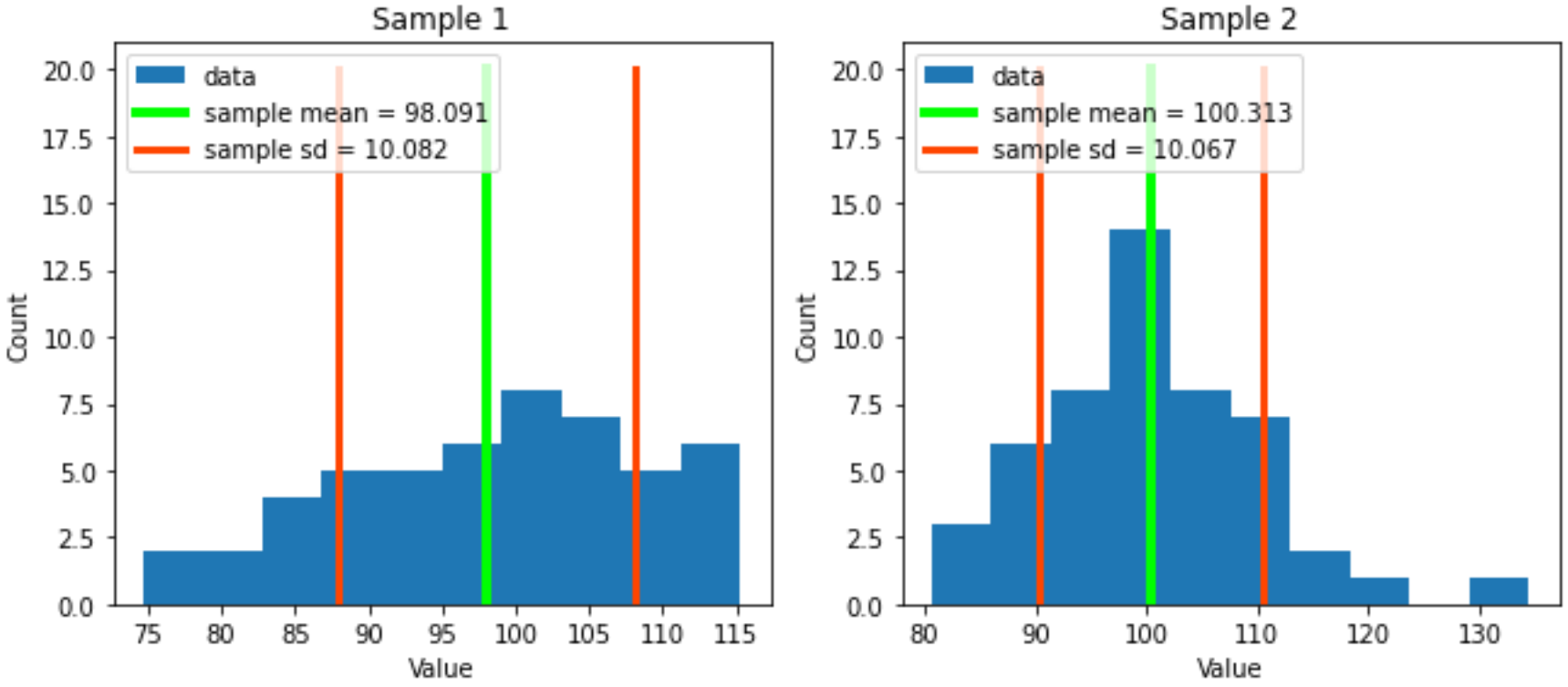
Generate 1000000 random numbers from
Normal Distribution with mean 100 and standard deviation 10





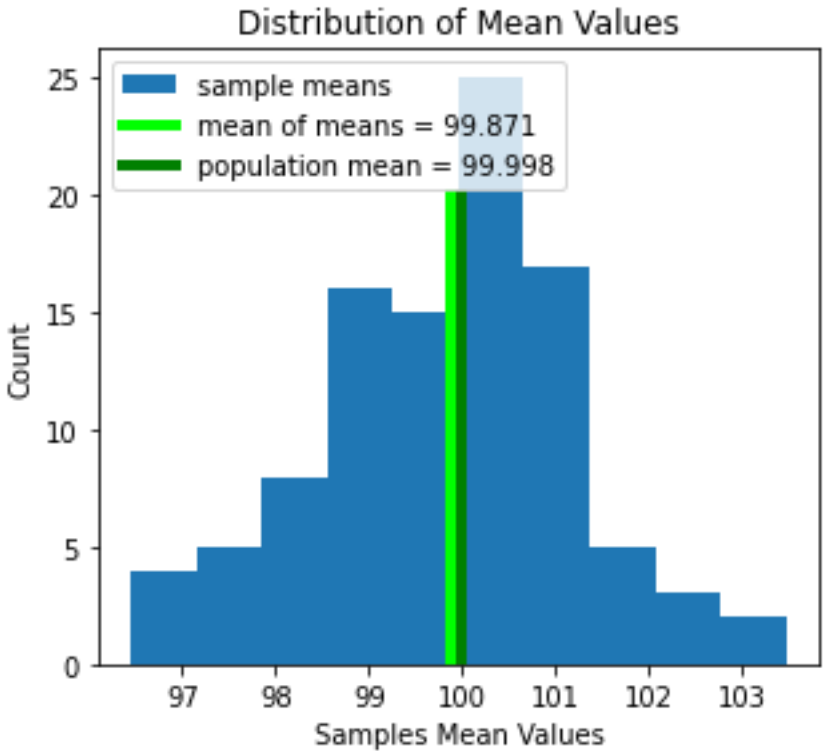
Sampling simulation

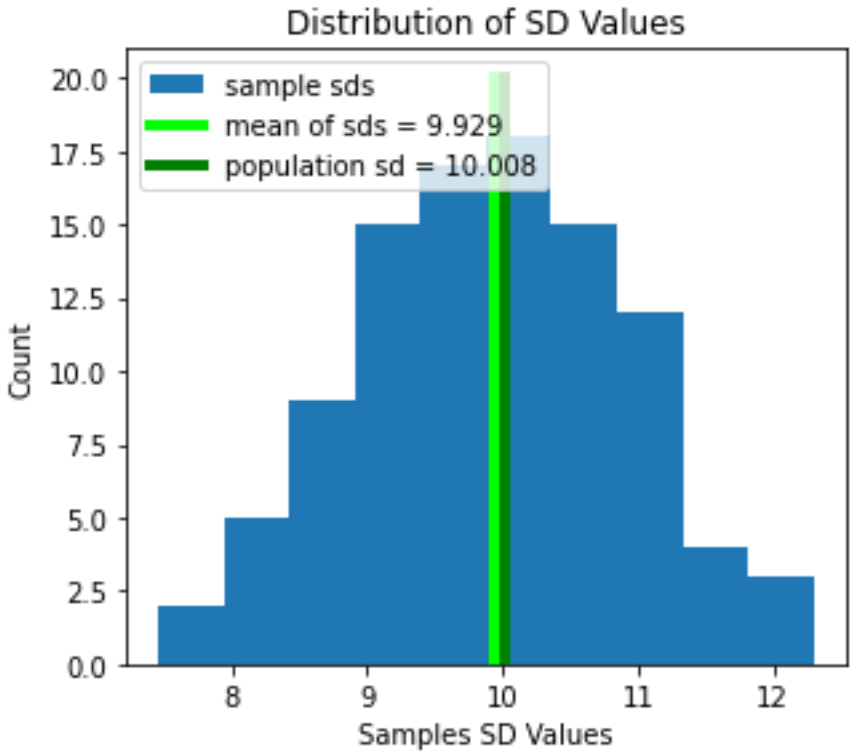
Sample from population: 2 samples with size 50



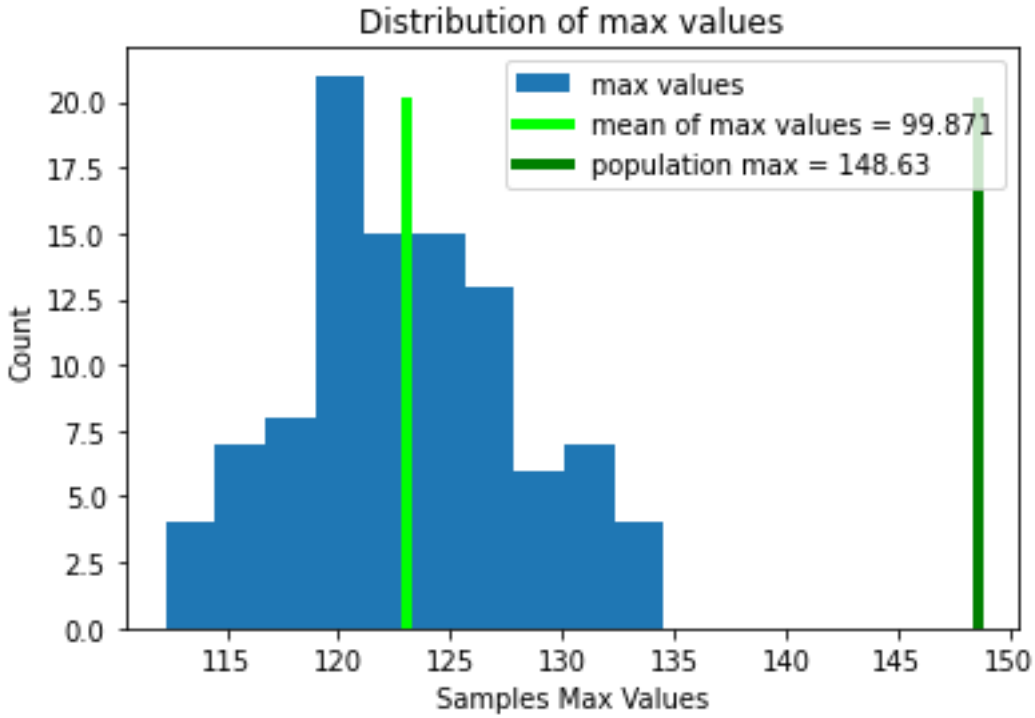
Mean and SD of Samples

Sample from population: 100 samples with size 50





Maximum Value of Samples



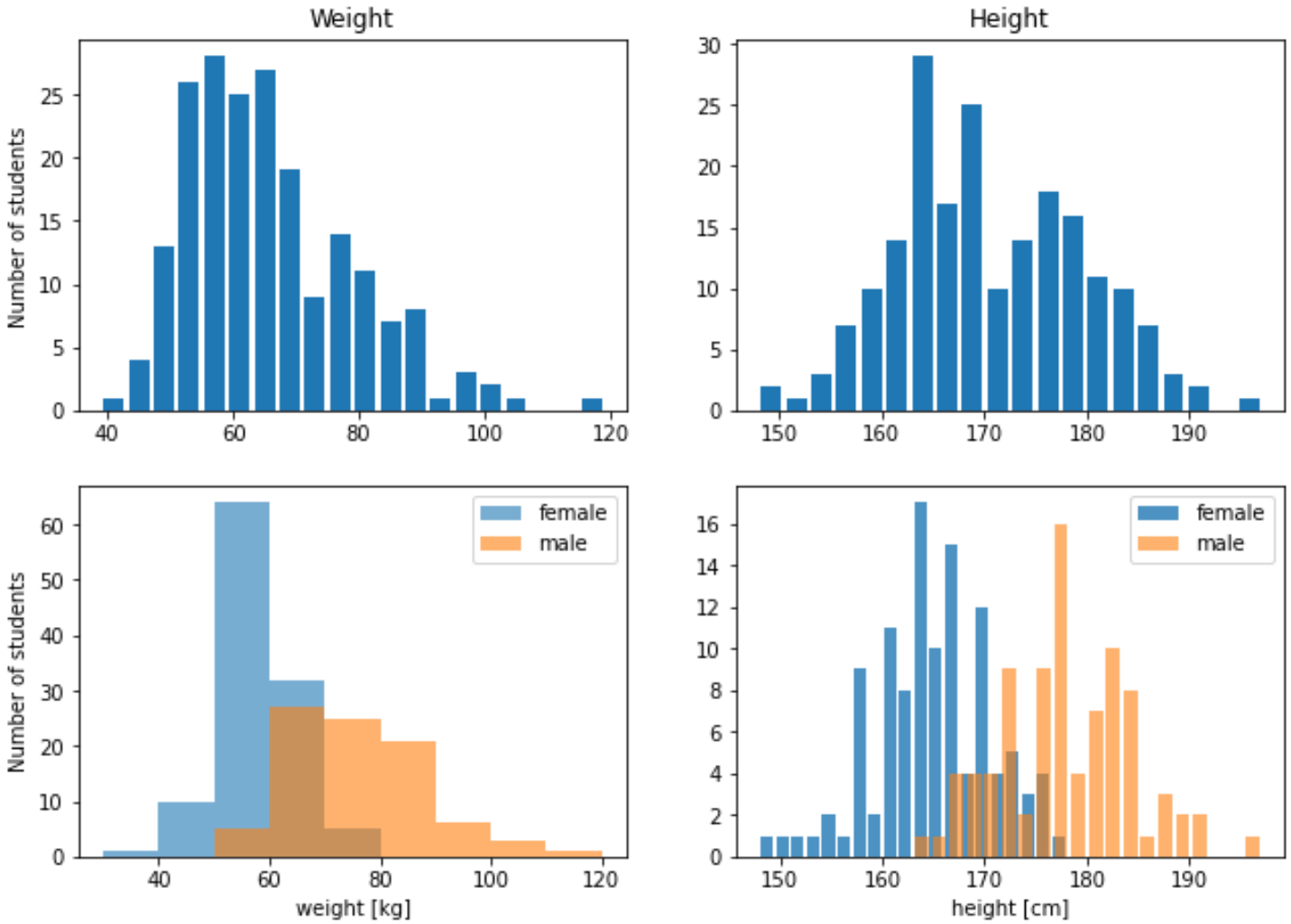
Example dataset 1

Self-reported and real weight and height of 200 students – 88 males and 112 females

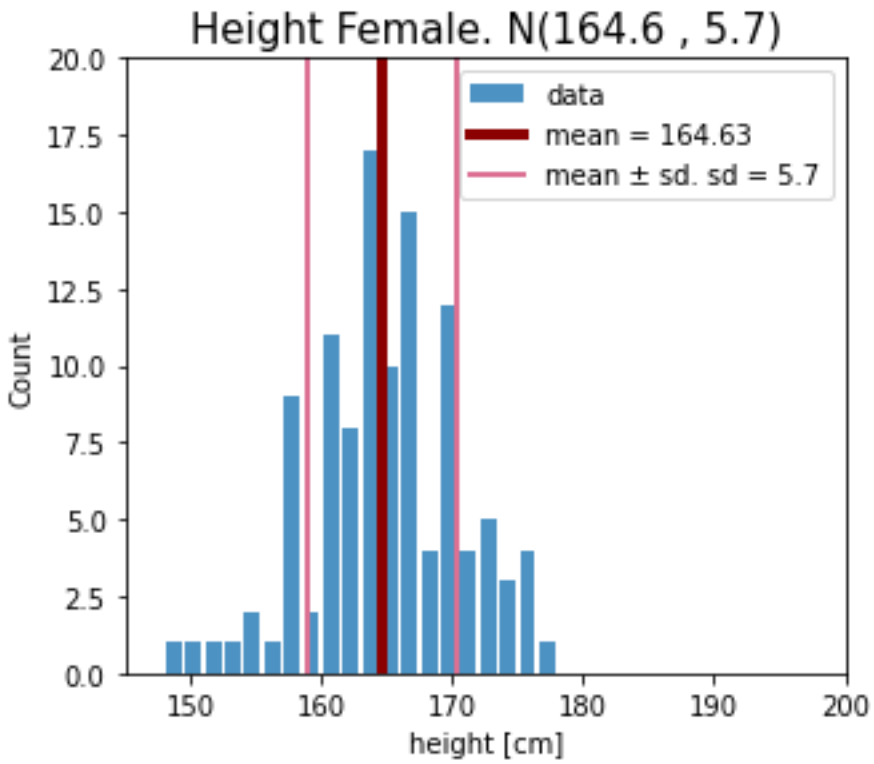
	subject	sex	weight	height	repwt	repht
0	1	M	77	182	77.0	180.0
1	2	F	58	161	51.0	159.0
2	3	F	53	161	54.0	158.0
3	4	M	68	177	70.0	175.0
4	5	F	59	157	59.0	155.0
..	...	..	...	...	...	...
195	196	M	74	175	71.0	175.0
196	197	M	83	180	80.0	180.0
197	198	M	81	175	NaN	NaN
198	199	M	90	181	91.0	178.0
199	200	M	79	177	81.0	178.0

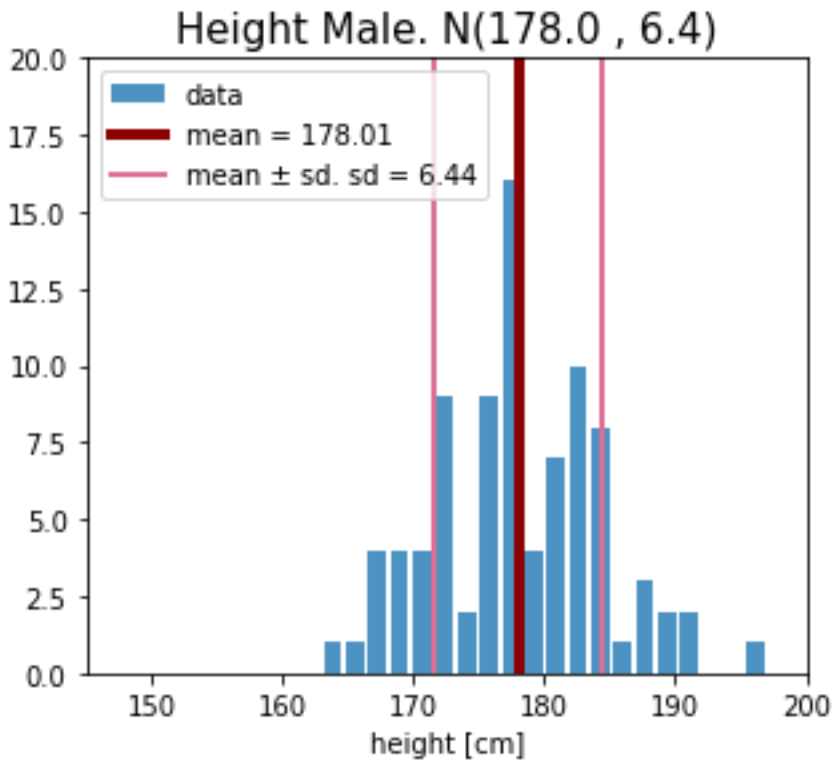
[200 rows x 6 columns]

Example dataset 1 Histogram

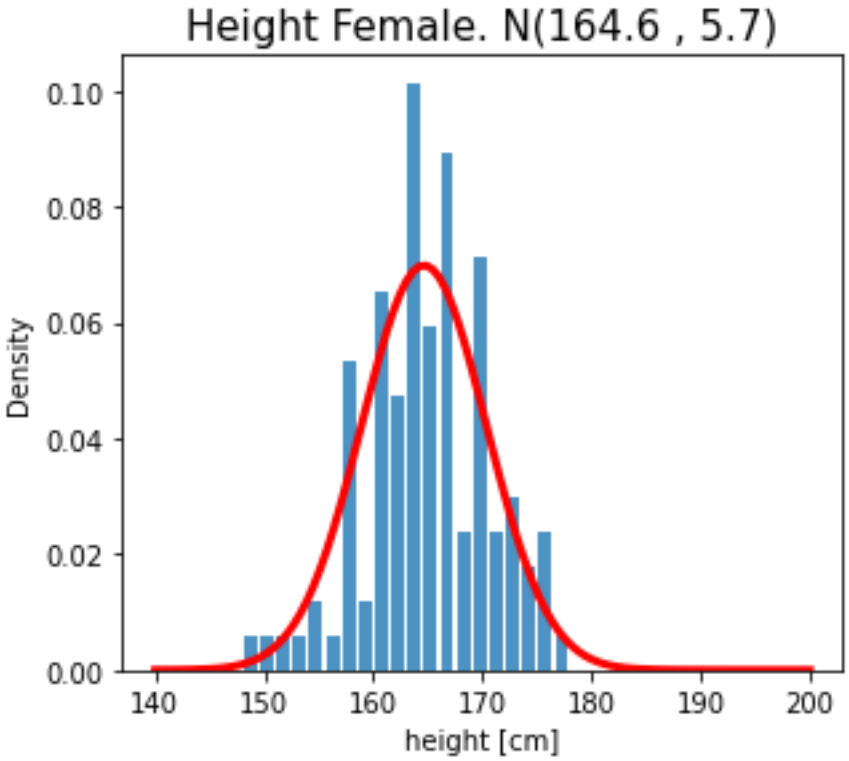


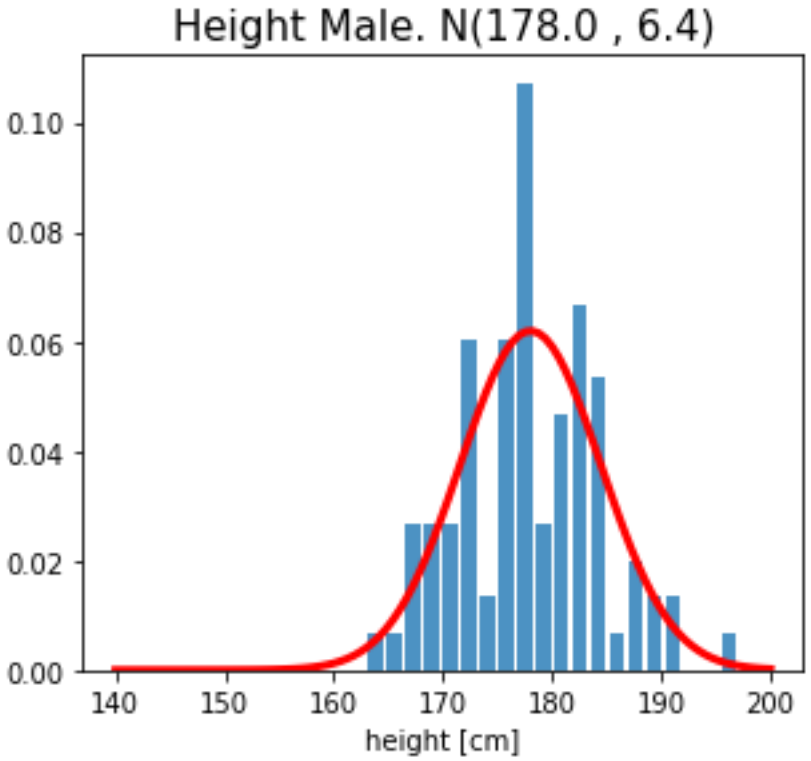
Example dataset 1
Histogram + Fit





Example dataset 1
Histogram + Fit





Example dataset 1

Distribution that fits heights:

female – N(165 , 5.7)

male – N(178 , 6.4)

Do they model the population?

Normal(mu , sigma)

? ?

There are methods for estimation of mean (mu) and standard deviation (sigma) of the population

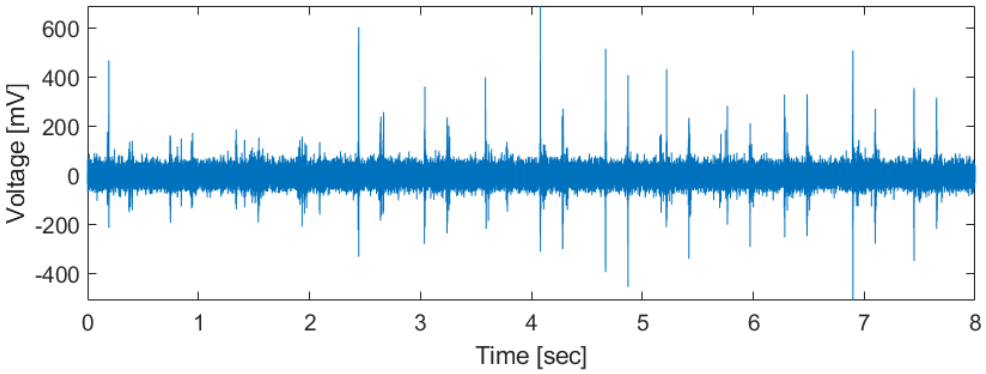
Example dataset 2

We recorded activity of one neuron and counted a number of spikes (action potentials) per second

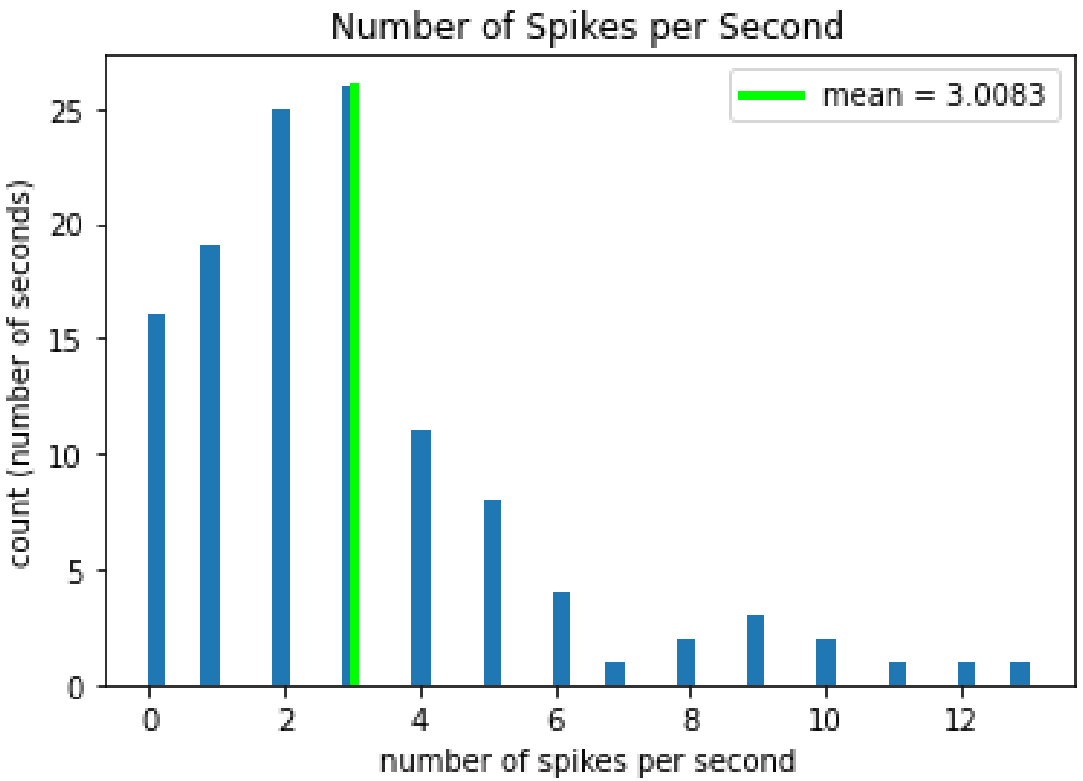
How can we describe the behavior of the this neuron?

	numSpikes
0	1
1	1
2	0
3	3
4	1
..	...
115	1
116	2
117	3
118	7
119	0

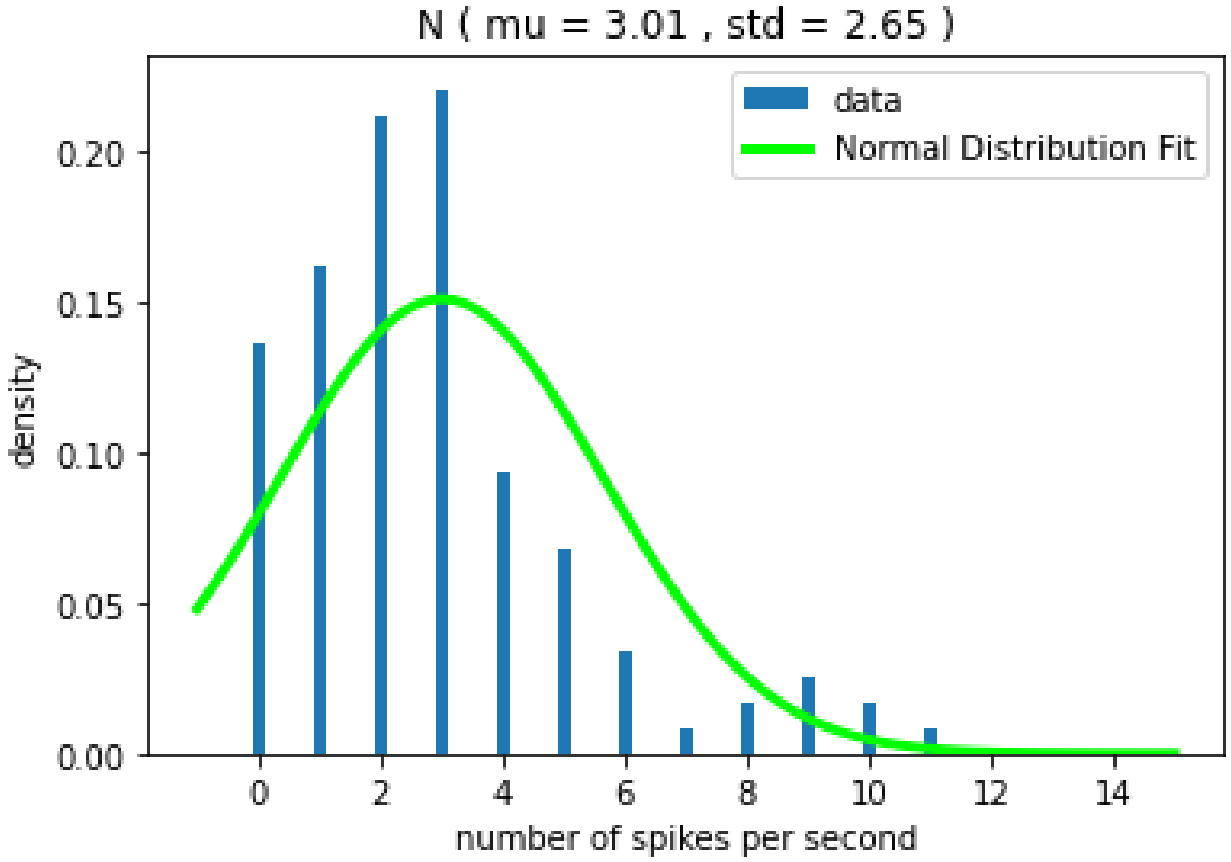
[120 rows x 1 columns]



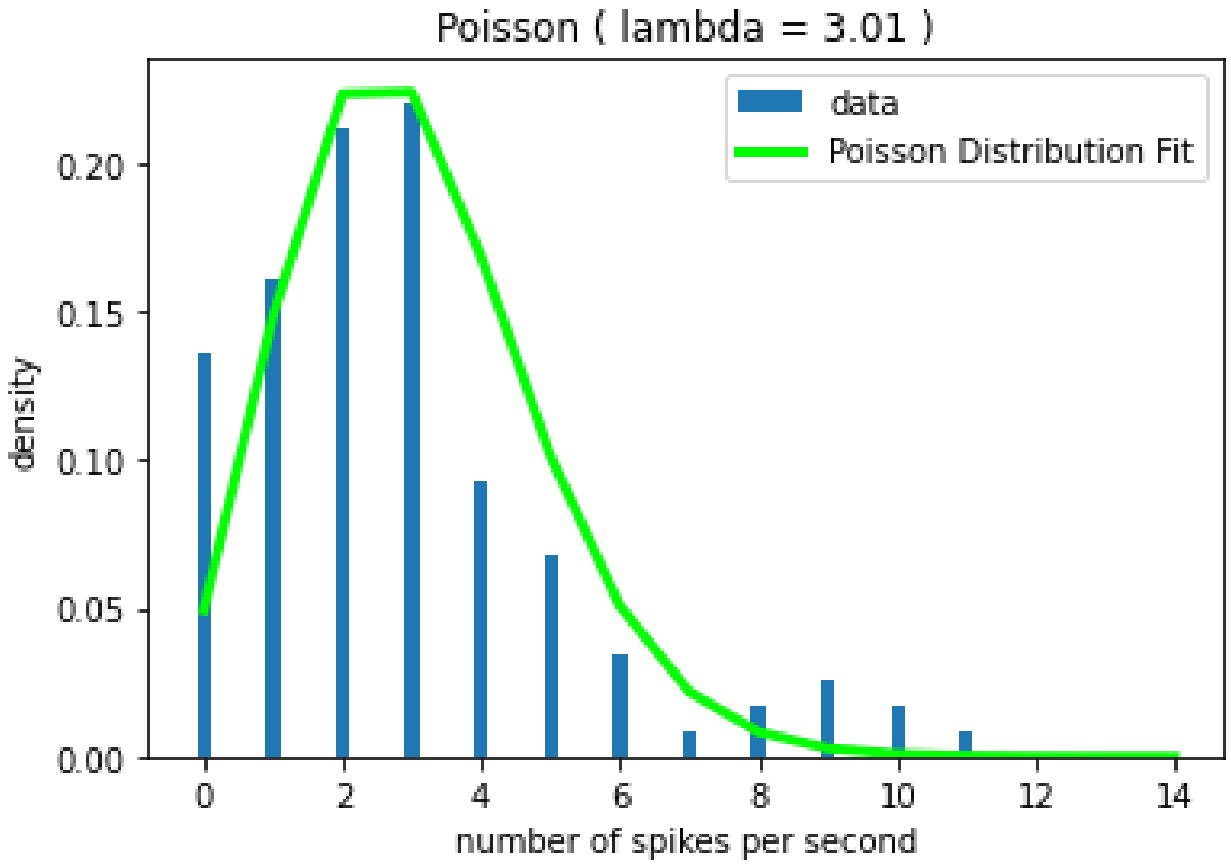
Example dataset 2 Histogram



Example dataset 2
Normal Distribution Fit



Example dataset 2
Poisson Distribution Fit



Example dataset 2

Distribution that fits neuron activity:

$$\text{spike frequency} \sim \text{Pois}(3.01)$$

Homework submission guidelines

- Zip file named ex#_########## (exercise number + ID number) must be submitted to the course website

- Zip file should contain:
    1. word document with your solution
    2. pdf of this document
    3. code file – .py or .ipynb (with comments)
    4. data files

- Every claim or conclusion you make should have an explanation
- Every graph must have a title, axis labels, and legend if there is more than one plot on a graph
- When you display the graph explain its content below
- Code should be annotated
- The homework and the project must be written in English.