Tutorial 3

Biological Data Analysis

Spring 2023

Outline

- Plug-in principle
- Biased / unbiased estimator
- Sample size significance
- Bootstrap:

  - parametric

  - non-parametric/regular

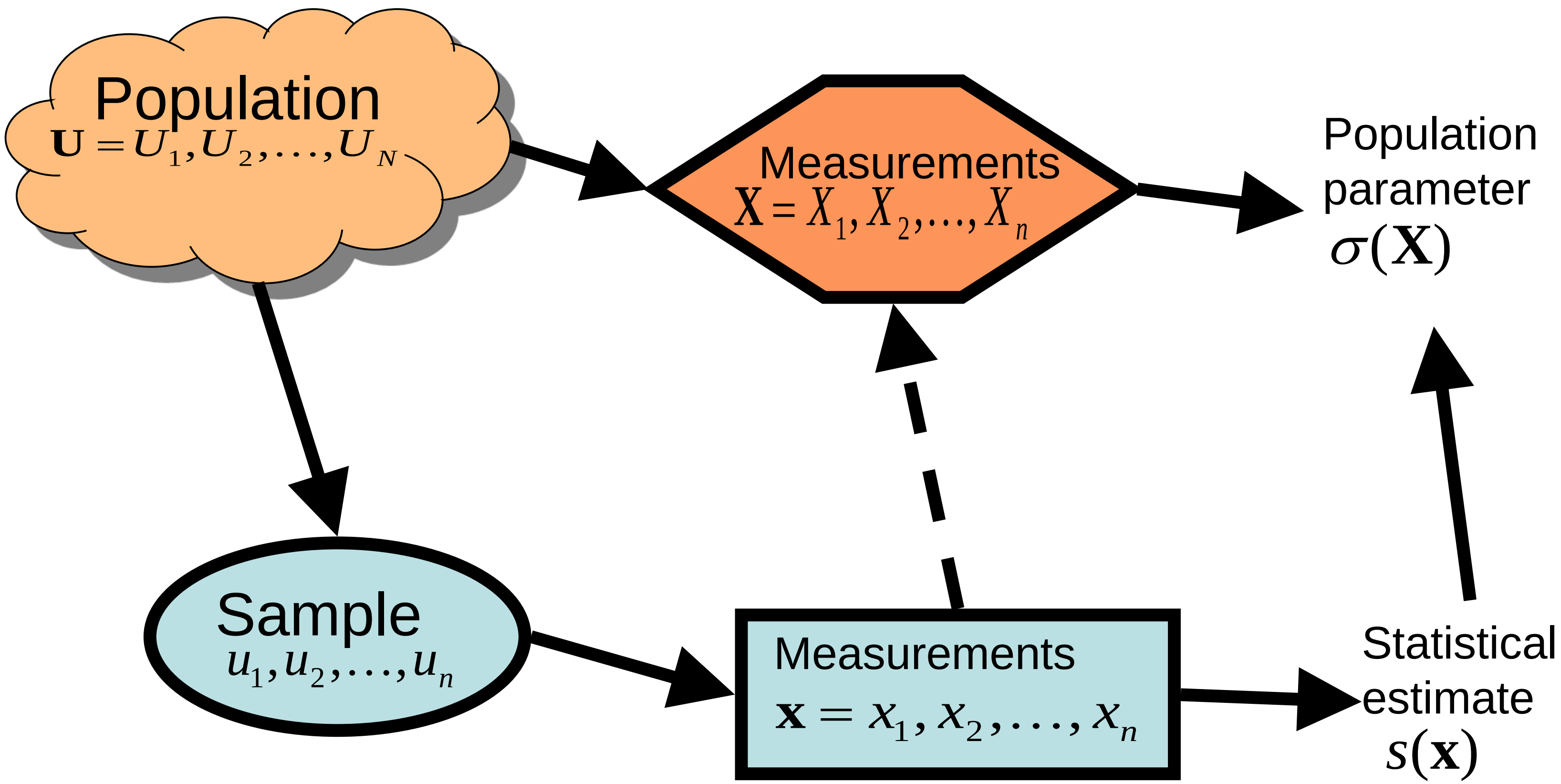
  - paired/unpaired
- Visualization

Plugin principle

Applying the same method/rule on sample and on population.

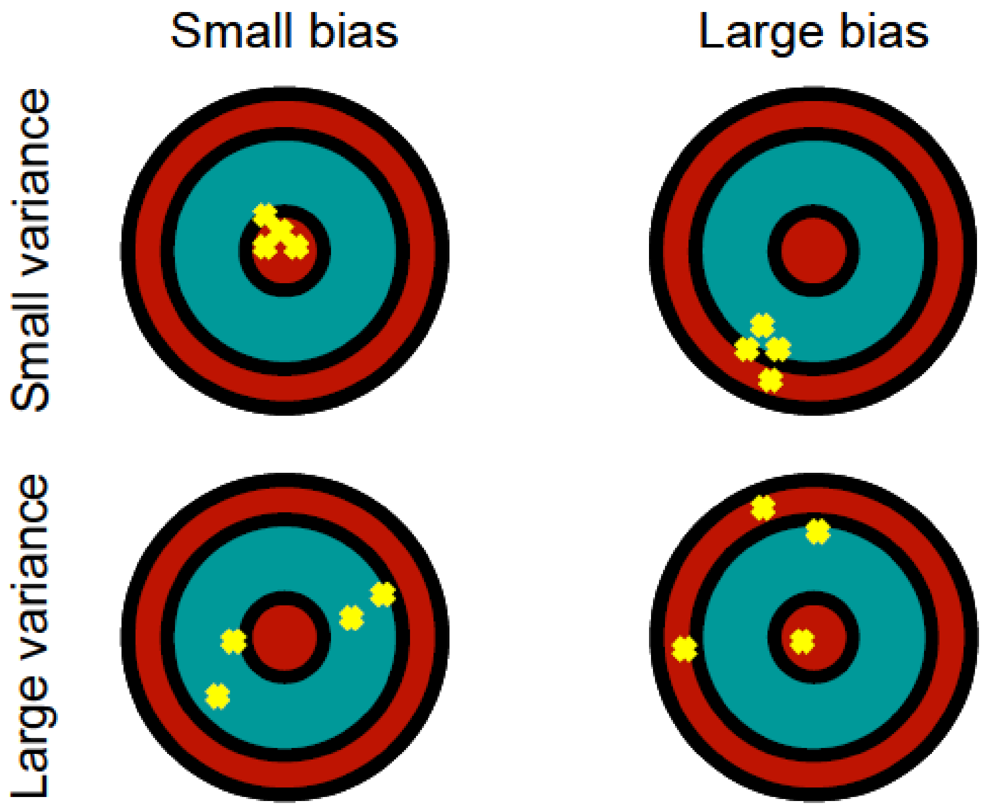
Statistic value is calculated on a sample using the same formula as population parameter.

Population and sample



Good estimator

Small bias and small variance



Bias

- **Unbiased estimator** $E\left[\hat{\theta}\right] = \theta$

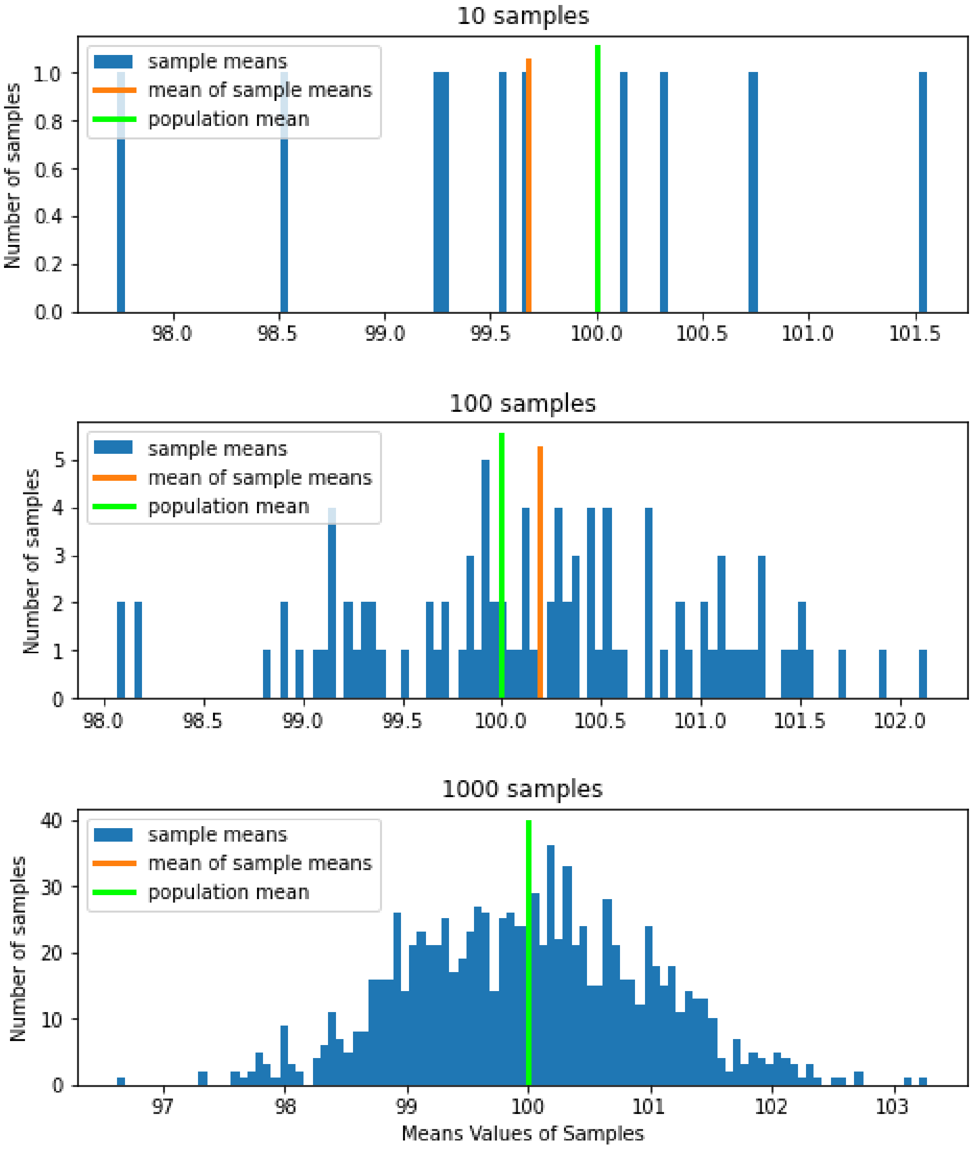
  expected value of the statistic =

  expected value of the population

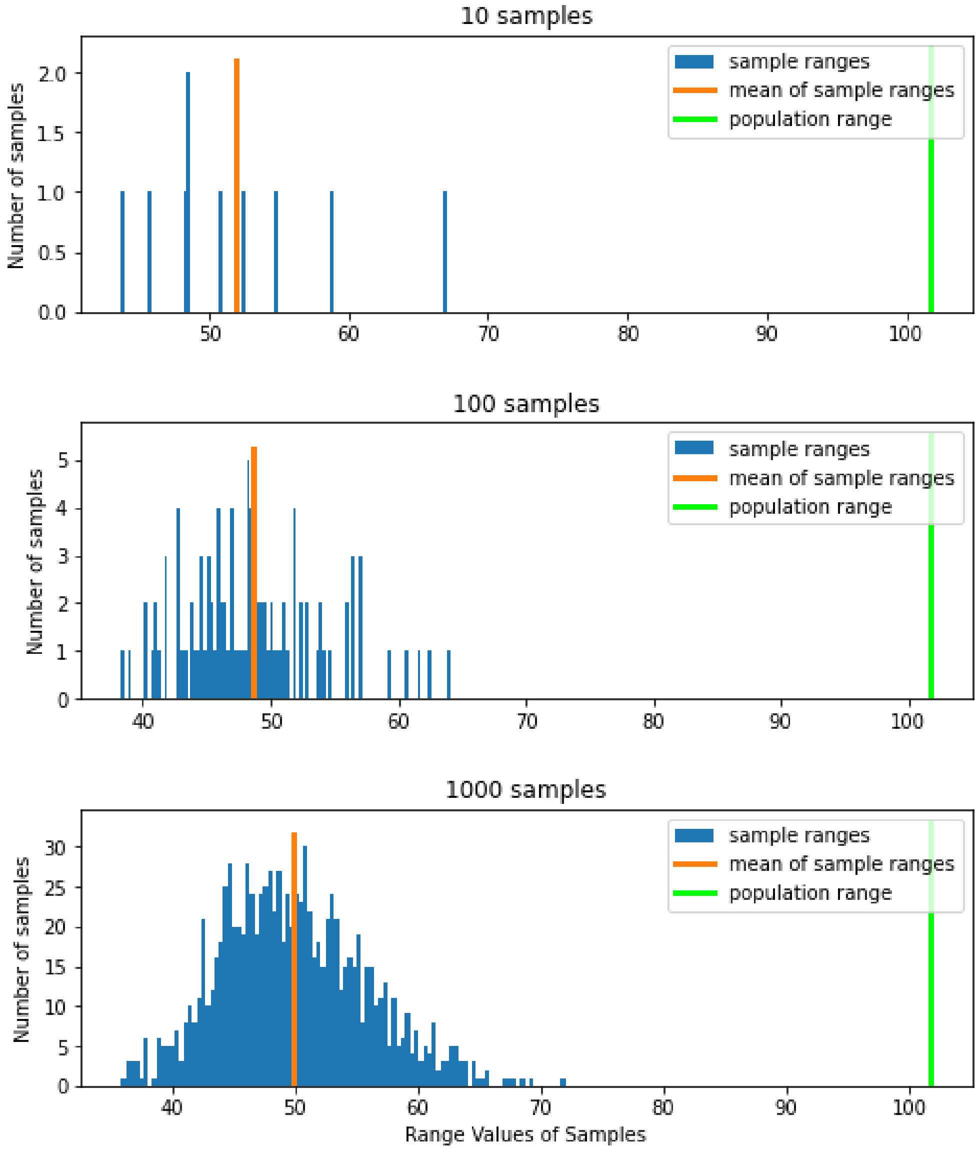
- **Biased estimator**

  - does not meet the condition $E\left[\hat{\theta}\right] = \theta$

Example – mean of samples



Example – range of samples



Biased and non-biased statistics

Mean – unbiased estimator to the population mean

also: not plug-in variance and standard deviation

Range – biased estimator to the population range

also: maximum, minimum, plug-in variance and standard deviation

Parameter vs statistic

Parameter	Statistic
$\mu = E(X)$	$\bar{X} = \dfrac{\sum_{i=1}^{N} x_i}{N}$
$\sigma^2 = E(X^2) - E(X)^2$	$s^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$
$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Variance and standard deviation

Population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

Plug-in estimator is biased

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2 \qquad E(S^2) \neq \sigma^2$$

Estimator unbiased

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

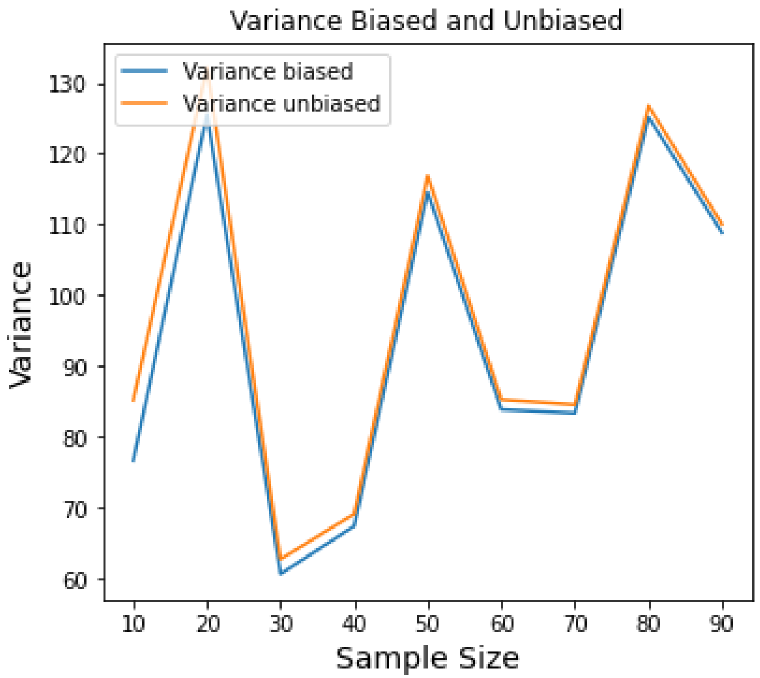
Standard deviation

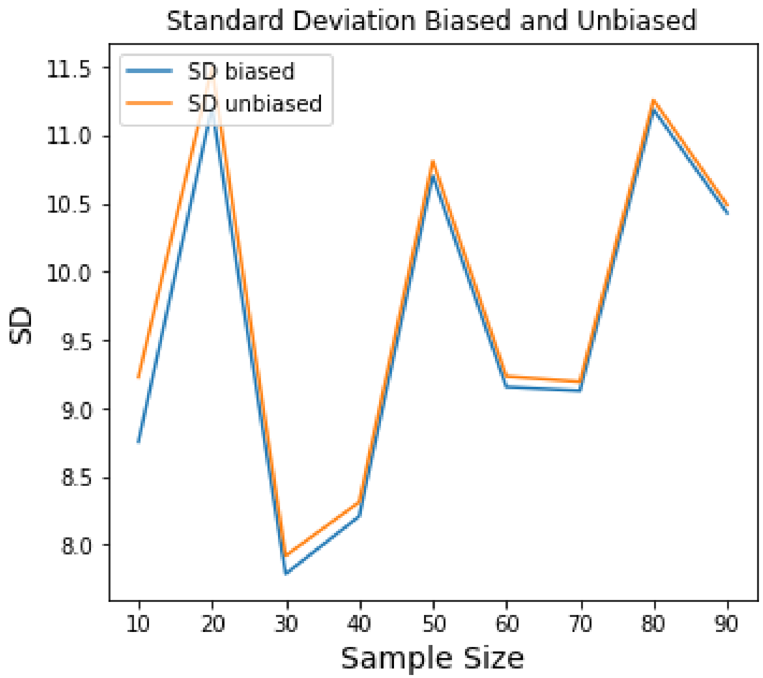
$$s = \sqrt{s^2}$$

Variance and Standard Deviation

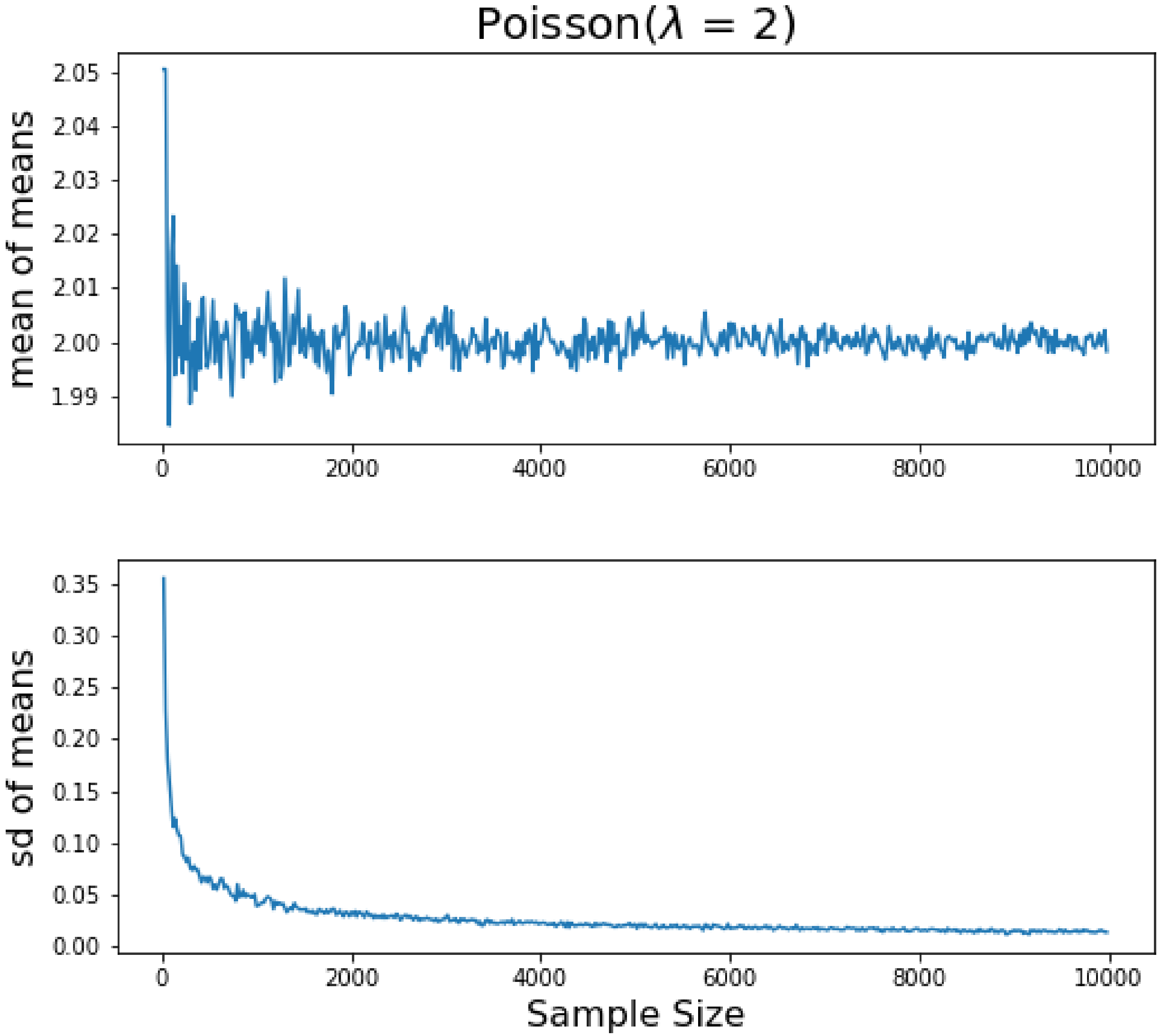
Variance biased $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

Variance unbiased $$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

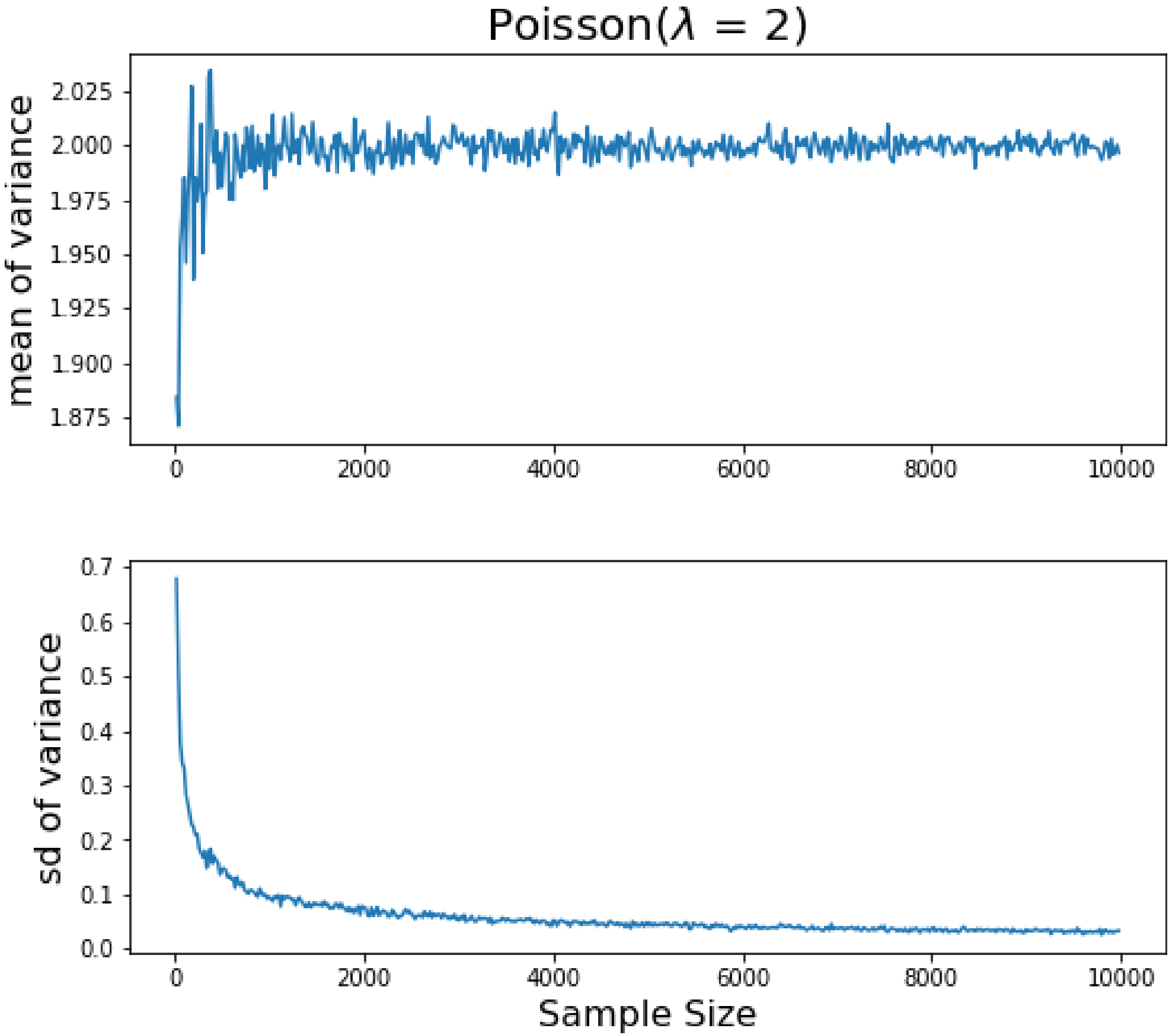




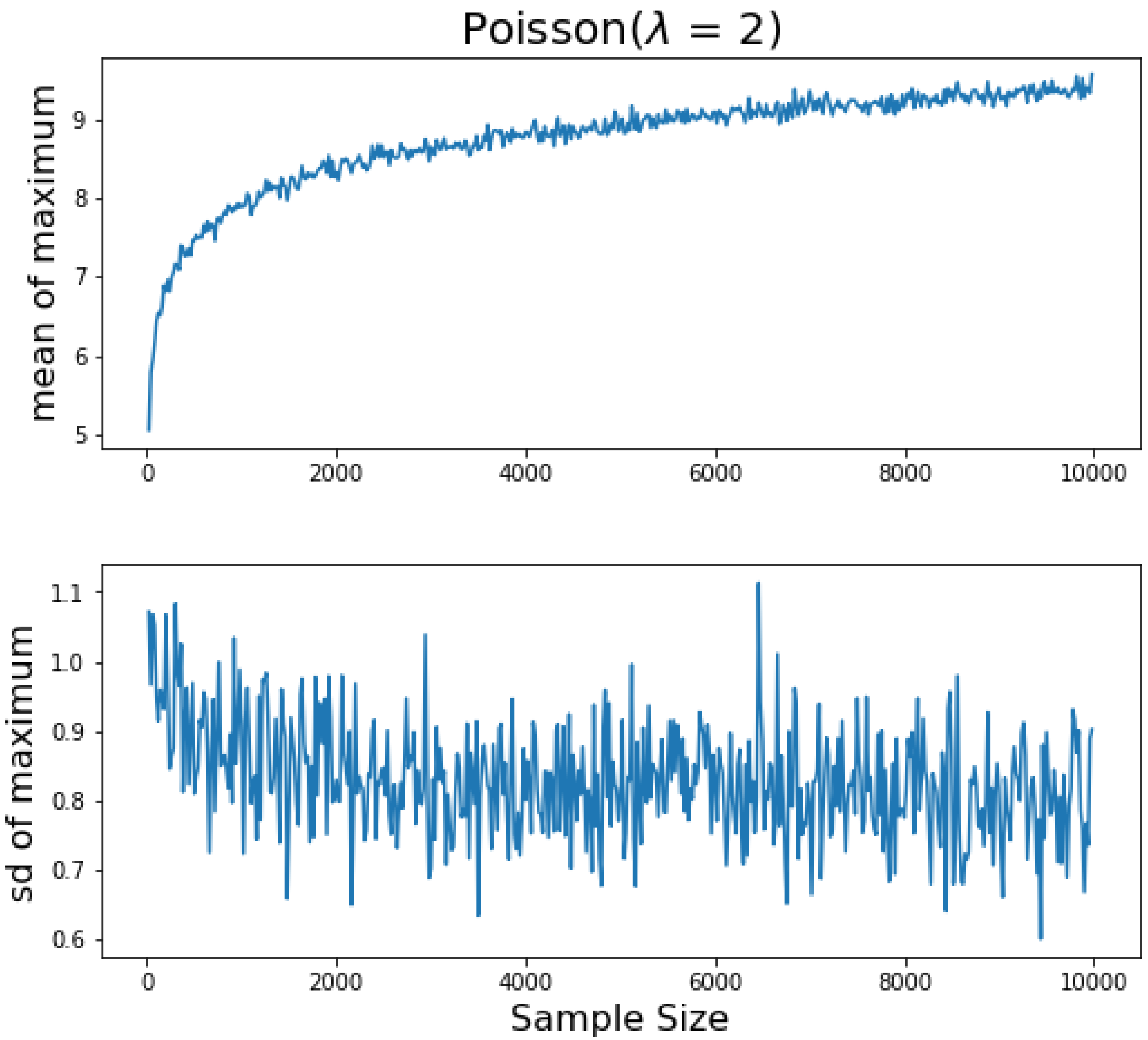
Effect of sample size on mean estimation



Effect of sample size on variance estimation



Effect of sample size on maximum estimation



Simulation conclusions

- All estimators get closer to the population parameter value as the sample size gets bigger:
- Including the biased estimators of SD. But the value is always smaller the that of the unbiased

- Variance of the estimators gets smaller as the sample size get larger

Bootstrap

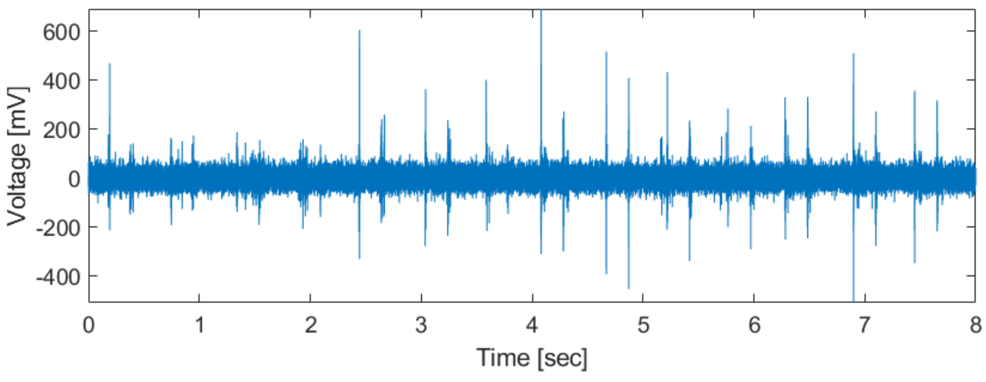
Creating new samples using existing sample

- Parametric: assume the sample data comes from a known distribution and the sample represents the population
    - estimate the distribution parameters from sample
    - generate new samples from the distribution using the calculated parameters
- Non-parametric (regular): assume the populations has the same distribution as the sample
    - generate new samples from the sample data

Example1

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

We want distribution of mean estimator (lambda of Poisson distribution)



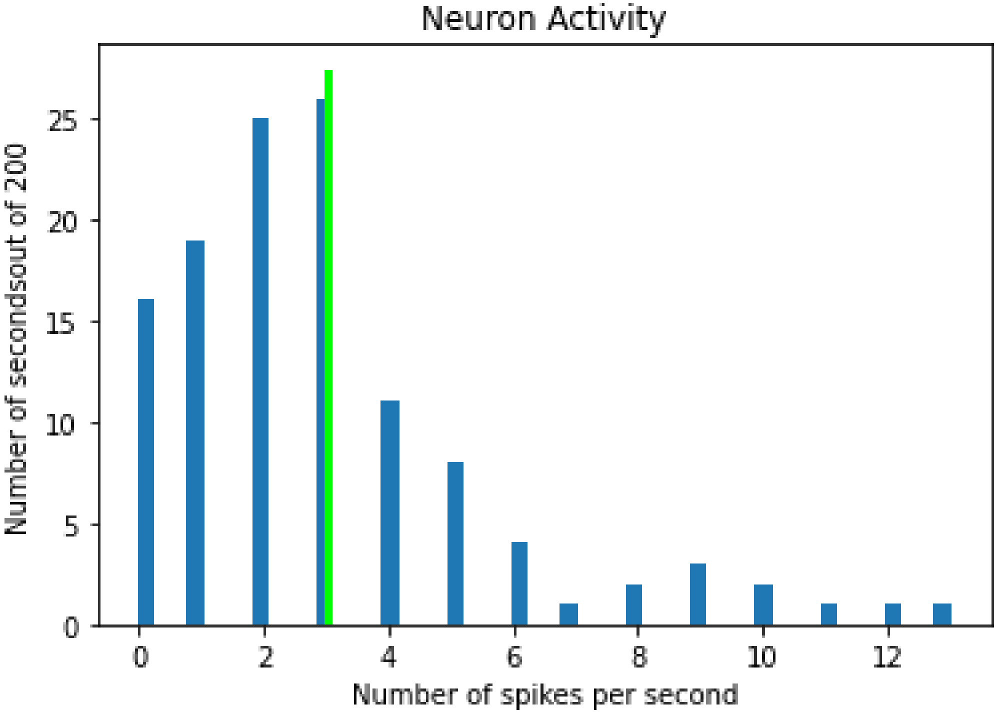

```
     numSpikes
0            1
1            1
2            0
3            3
4            1
..         ...
115          1
116          2
117          3
118          7
119          0

[120 rows x 1 columns]
```

Example1

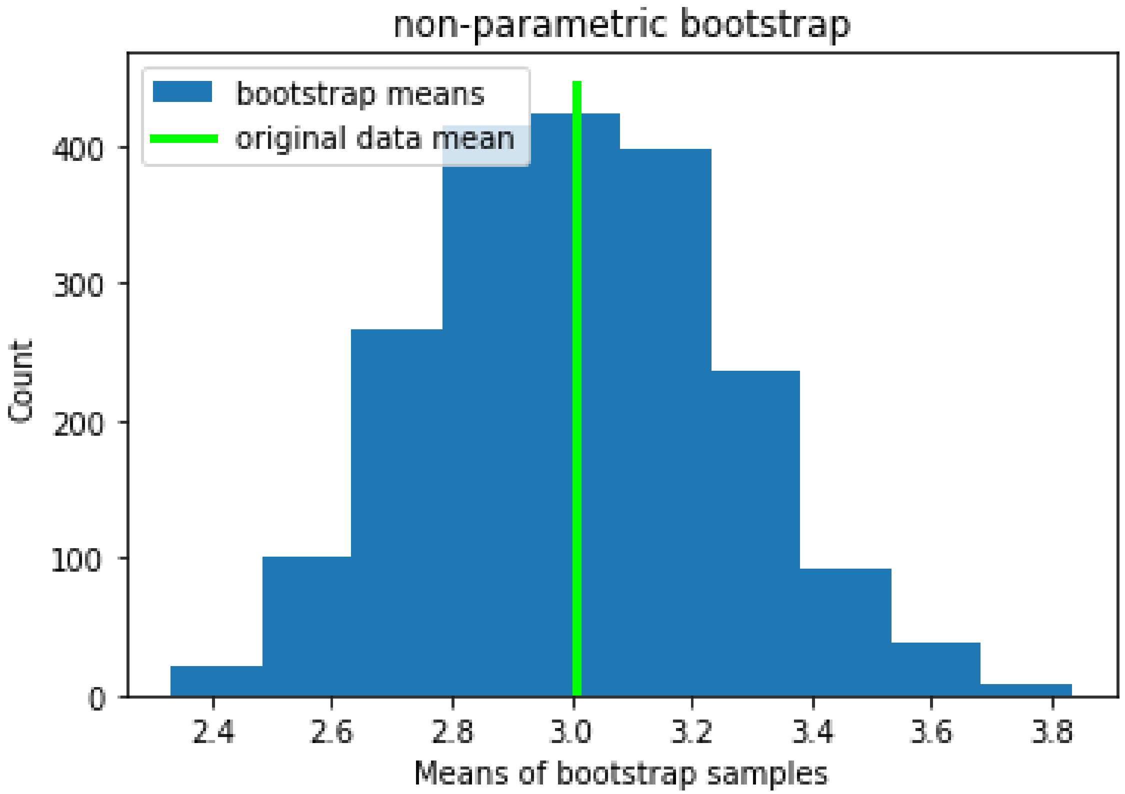
Histogram of original data

Data mean = 3.01



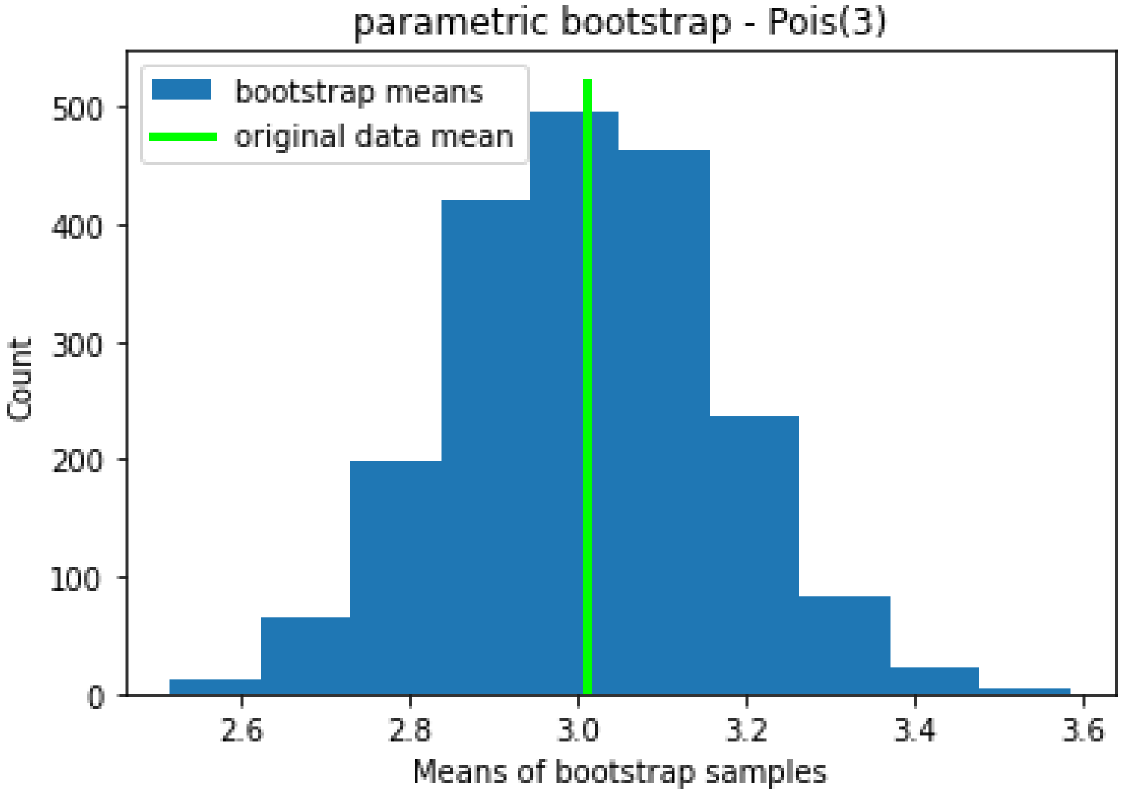
Non-parametric Bootstrap

Sample from data with repetitions: 2000 samples of 200

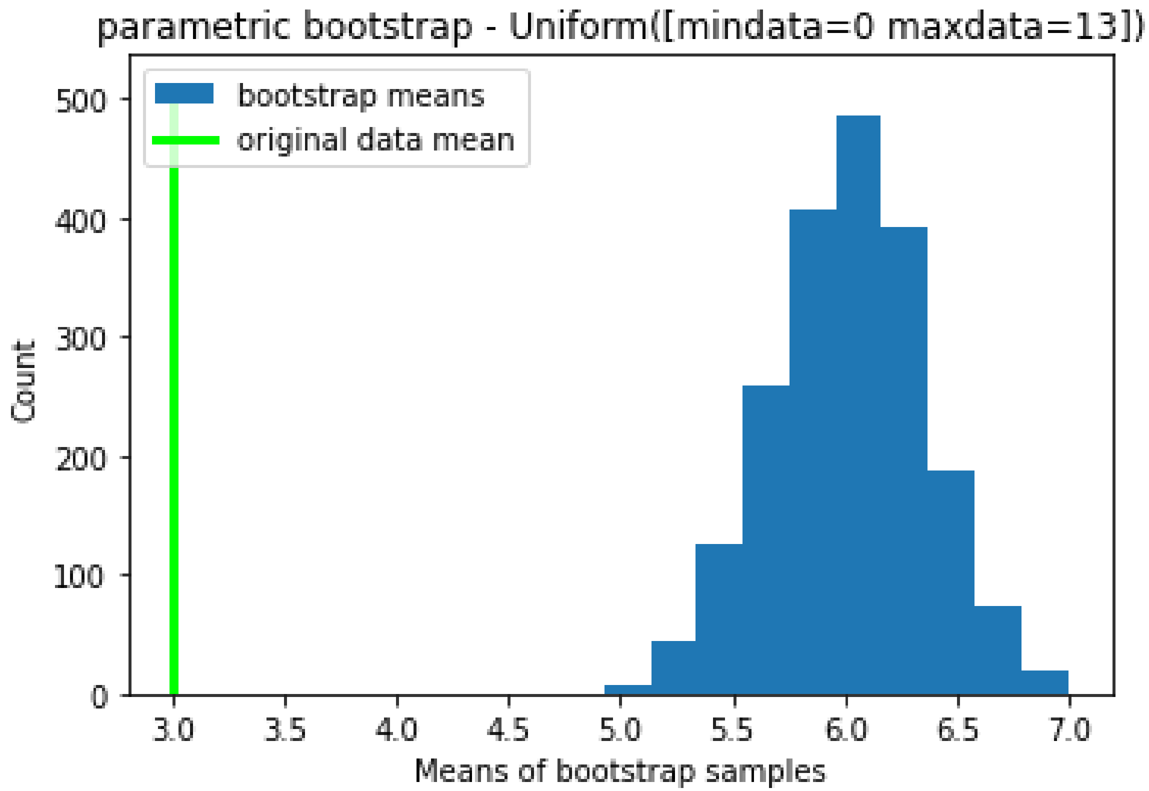


Parametric Bootstrap

Sample from distribution: 2000 samples of 200



Parametric Bootstrap wrong distribution



Exercise 2
Bootstrap Paired and Unpaired

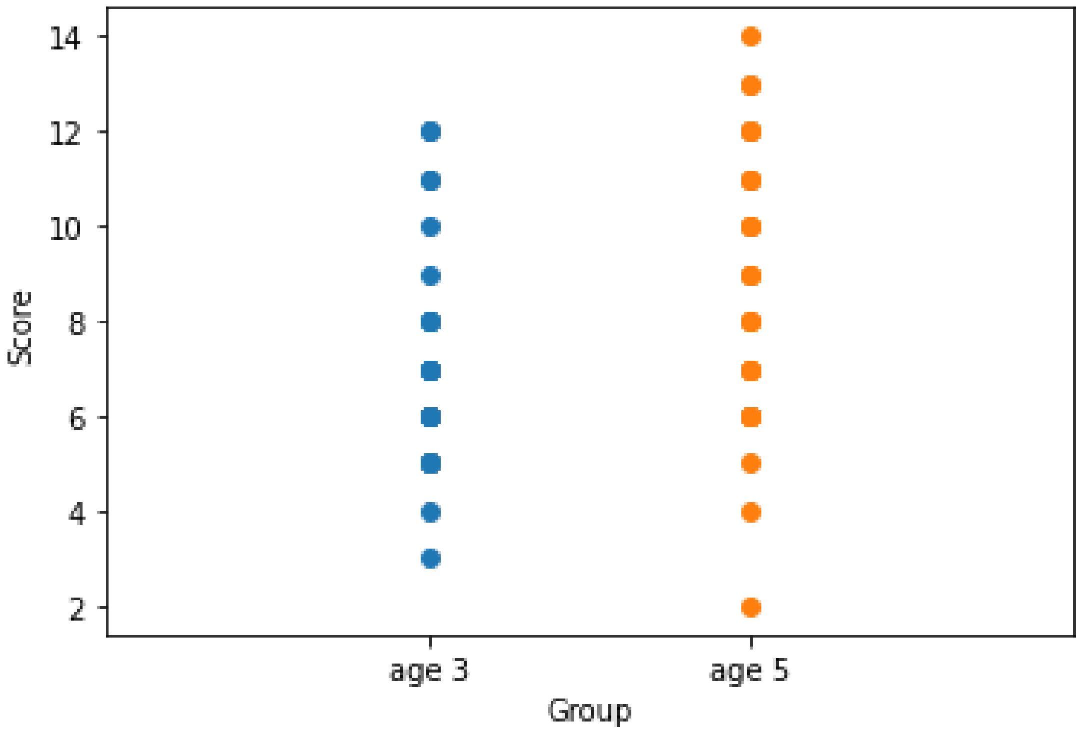
We want to measure the difference in verbal comprehension in children aged 3 and 5.

We are given a table with comprehension scores of two groups of children: 30 – aged 3, 30 - aged 5

```
comprehension = [[5,6,8,12,11,9,7,6,8,7,8,12,5,6,7,8,11,7,7,5,3,4,10,6,6,5,7,6,5,6],
                 [6,7,9,13,12,10,11,8,10,9,8,13,6,14,6,11,12,10,9,4,5,12,11,2,7,8,7,6,6,7]]
```

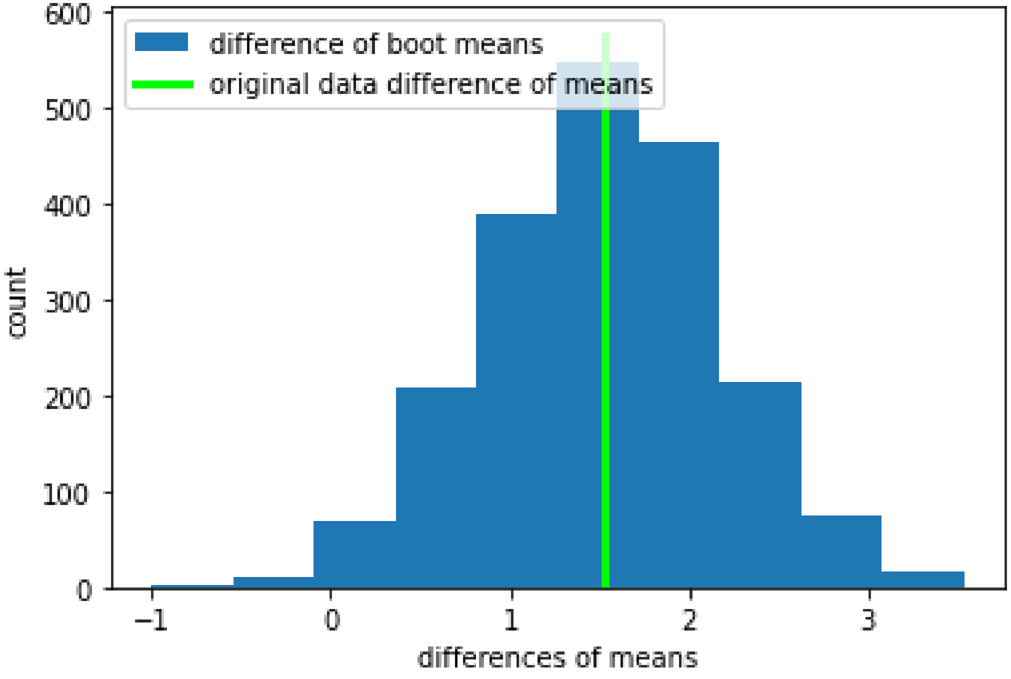
Exercise 2

Unmatched pairs: different children in two groups



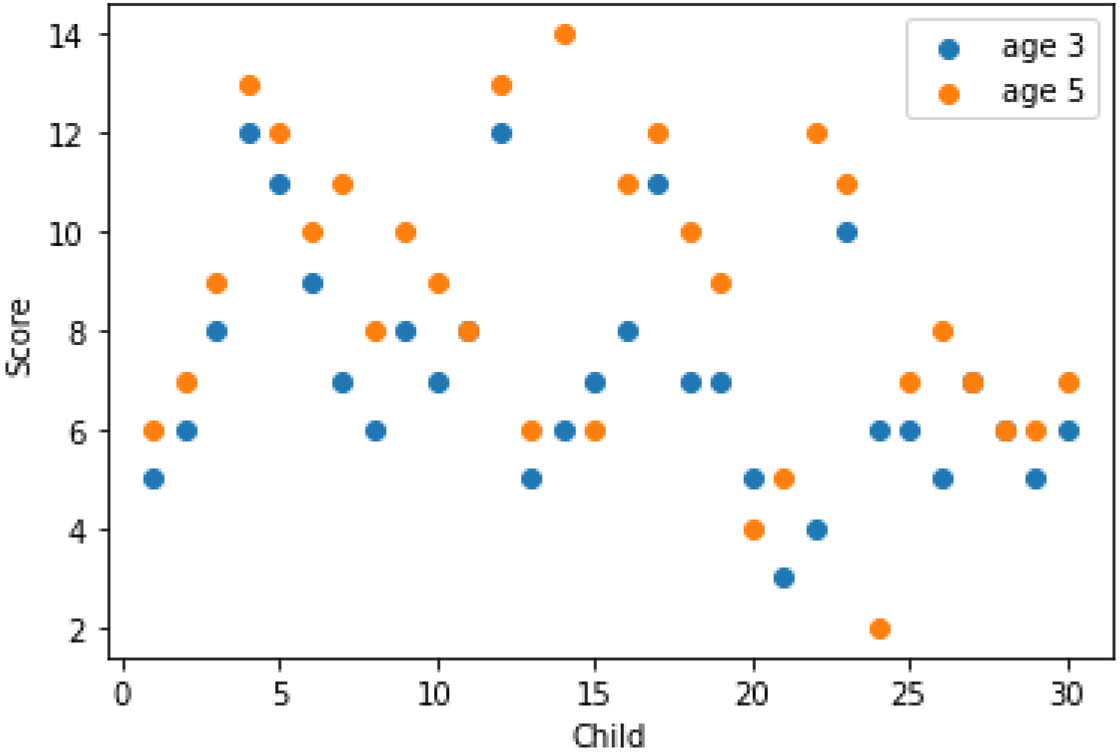
Exercise 2
Unmatched pairs

Regular bootstrap:
sample from group 1
sample from group 2
calculate differences of means



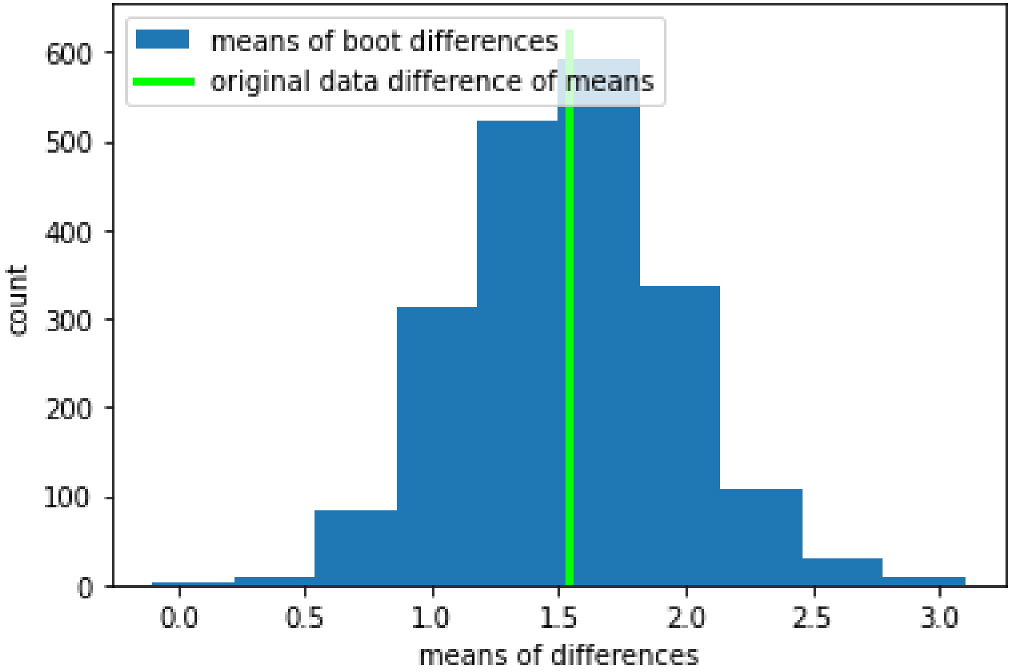
Exercise 2

Matched pairs: same children in both groups



Exercise 3
Matched pairs

Regular bootstrap:
calculate differences
sample from group of differences
calculate means of sampled differences



Visualization

Different ways of plotting data

- Plot all the data points
- Cumulative frequency plots
- Histogram …

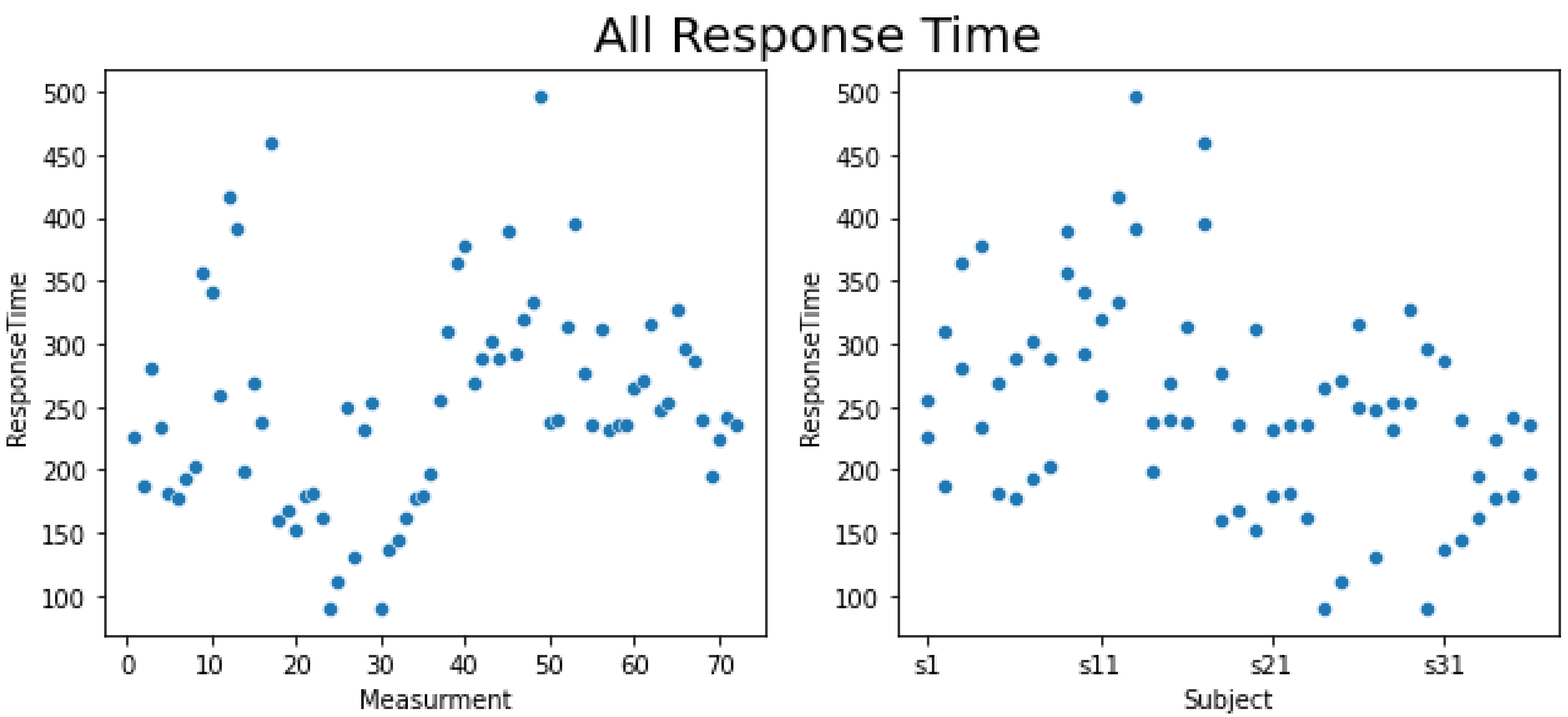
Example dataset

Two groups – musicians and non-musicians – reacted to two types of stimuli – visual and auditory. After seeing or hearing a stimulus they had to press a button. We have the file with the reaction time.

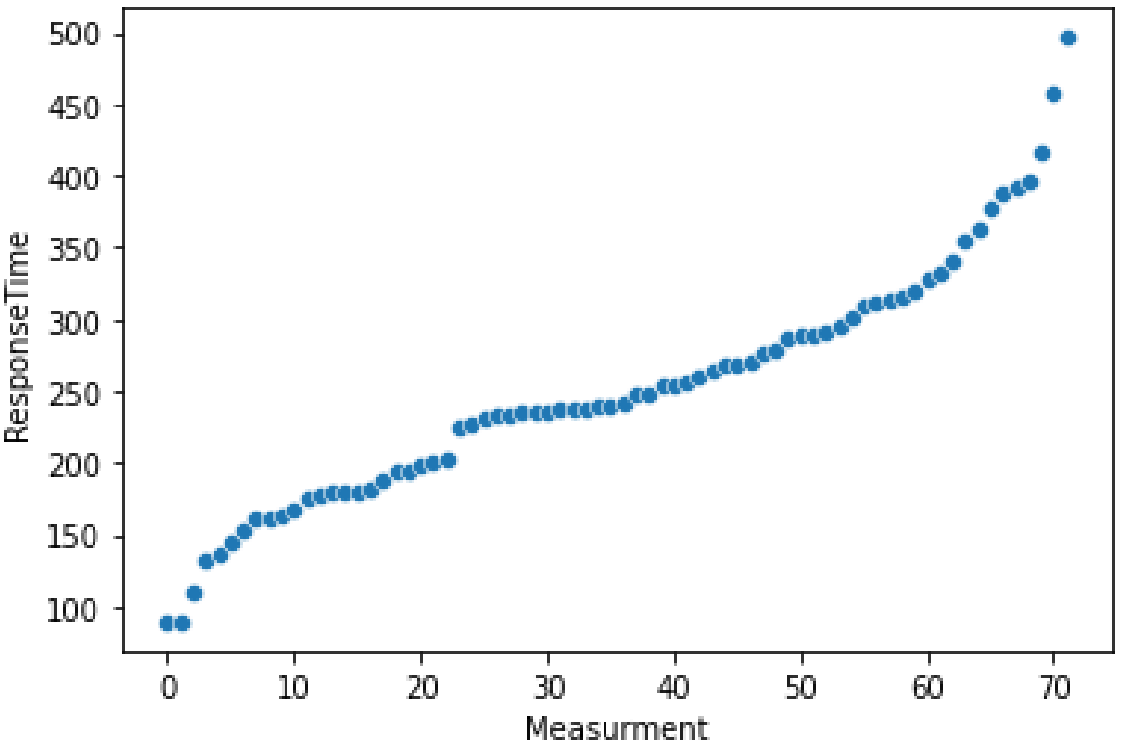
36 subjects: 13 musicians and 23 non-musicians
72 measurements: 36 auditory and 36 visual

	Meas	Subject	ResponseTime	Stimulus	Group
0	1	s1	226.27	auditory	NonMusician
1	2	s2	187.52	auditory	NonMusician
2	3	s3	279.77	auditory	NonMusician
3	4	s4	233.83	auditory	NonMusician
4	5	s5	180.83	auditory	NonMusician
..	...	...	...	...	...
67	68	s32	239.00	visual	Musician
68	69	s33	194.93	visual	Musician
69	70	s34	224.60	visual	Musician
70	71	s35	240.93	visual	Musician
71	72	s36	234.95	visual	Musician

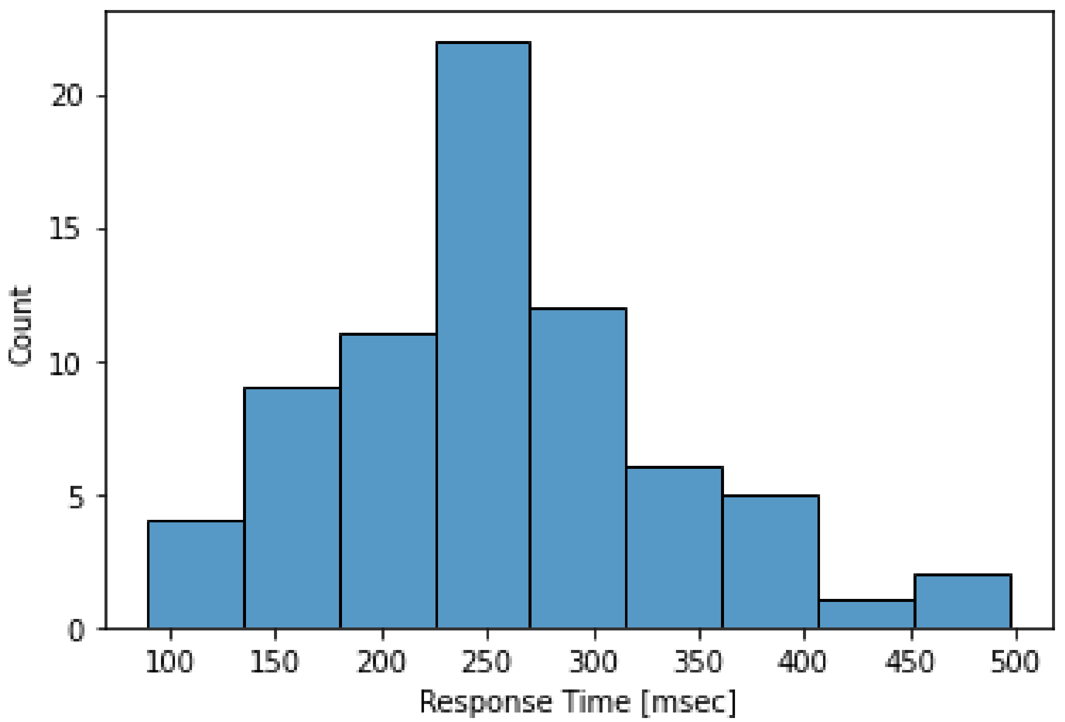
plot all data: scatter



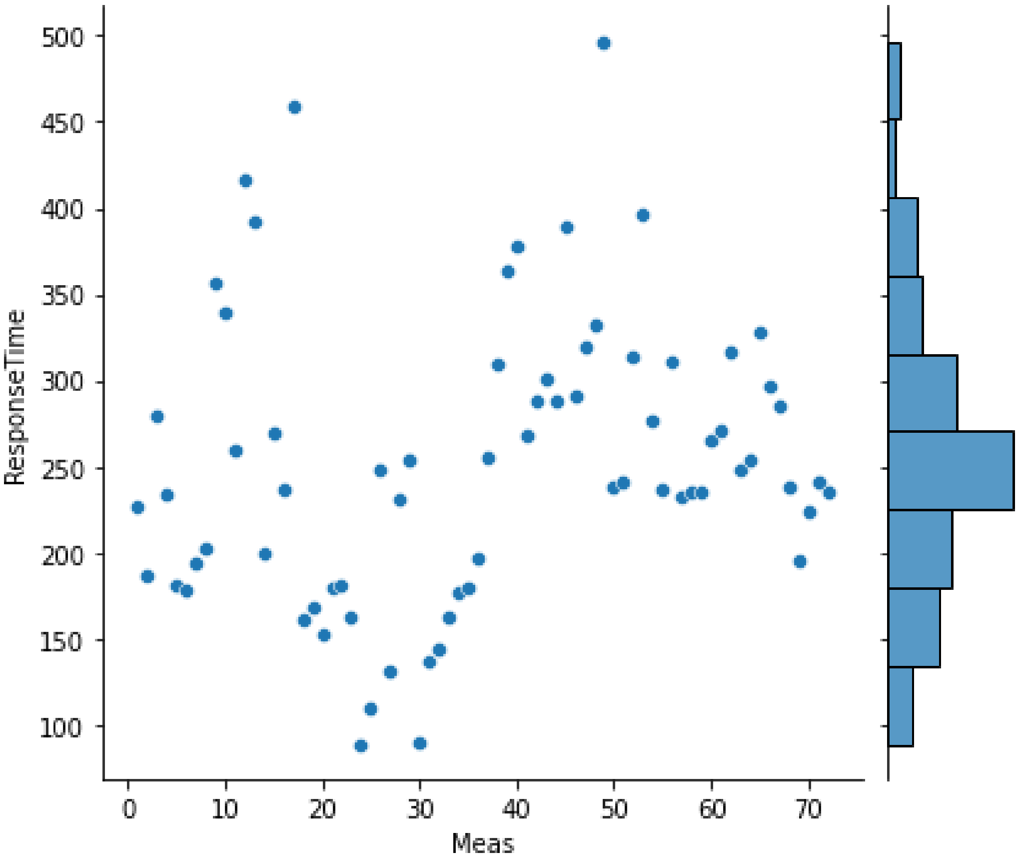
plot all data: cumulative plot



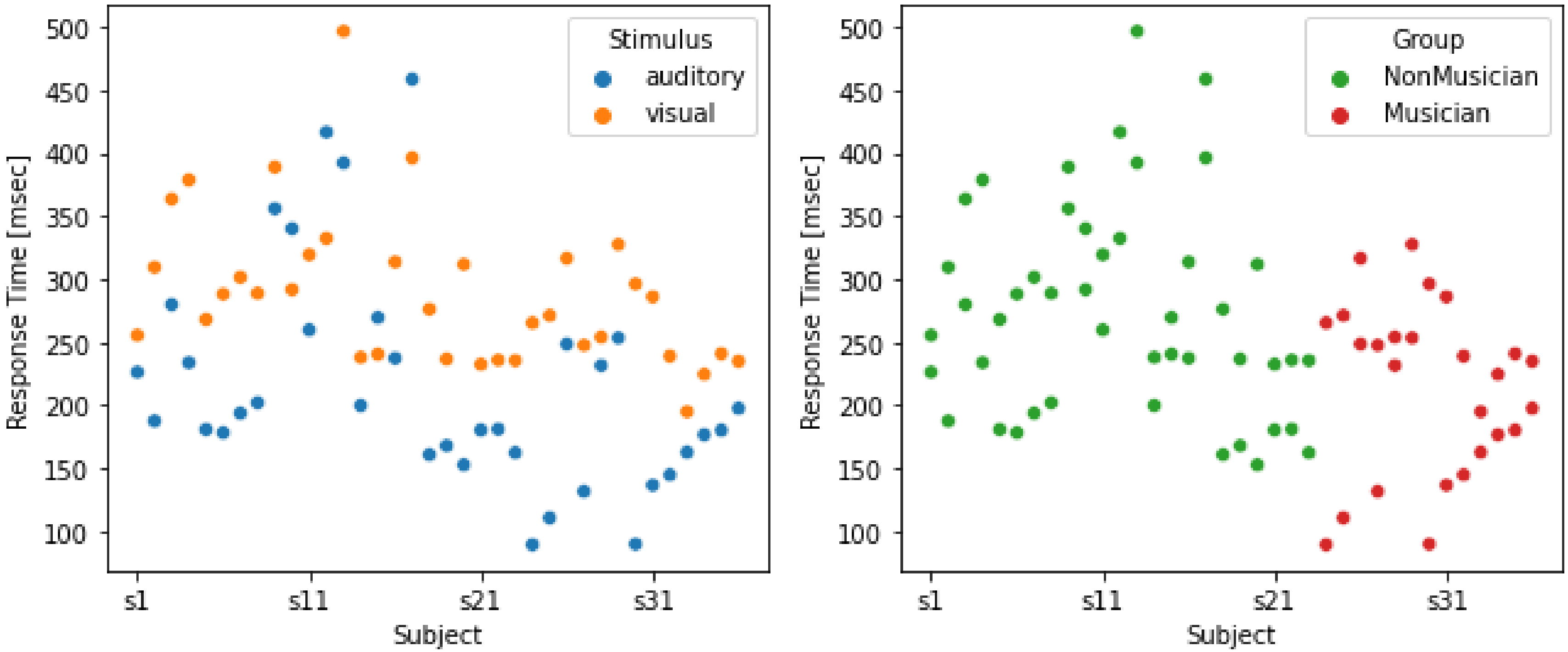
plot all data: histogram



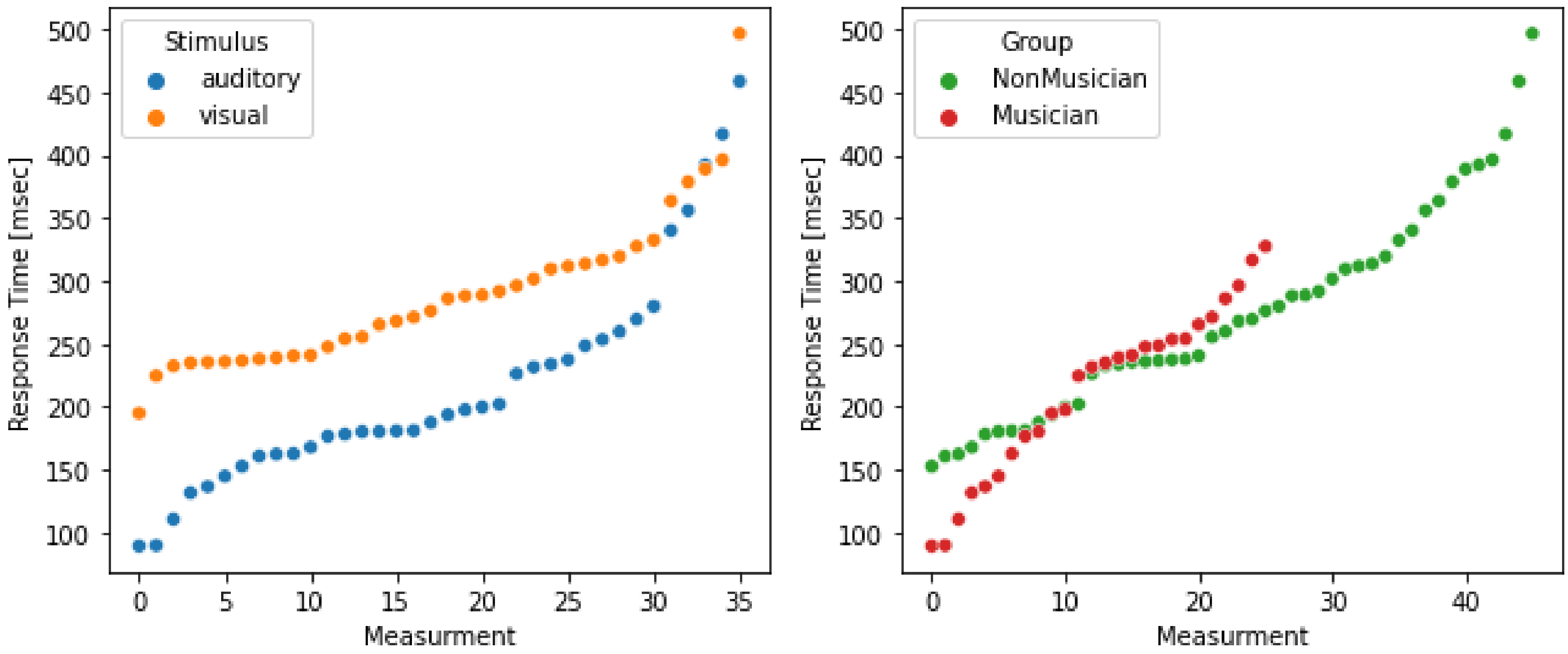
plot all data: combination



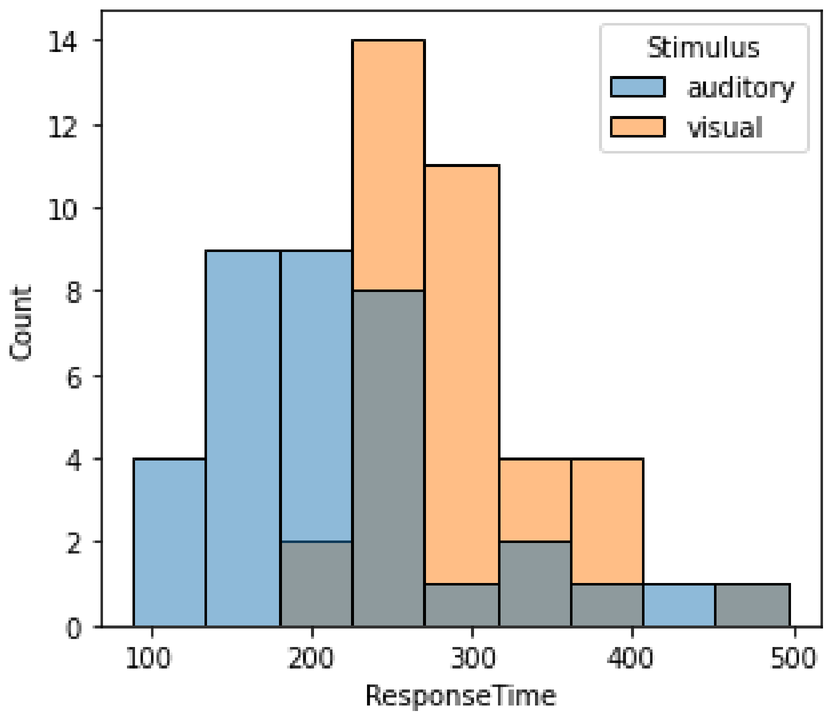
plot data by group: scatter

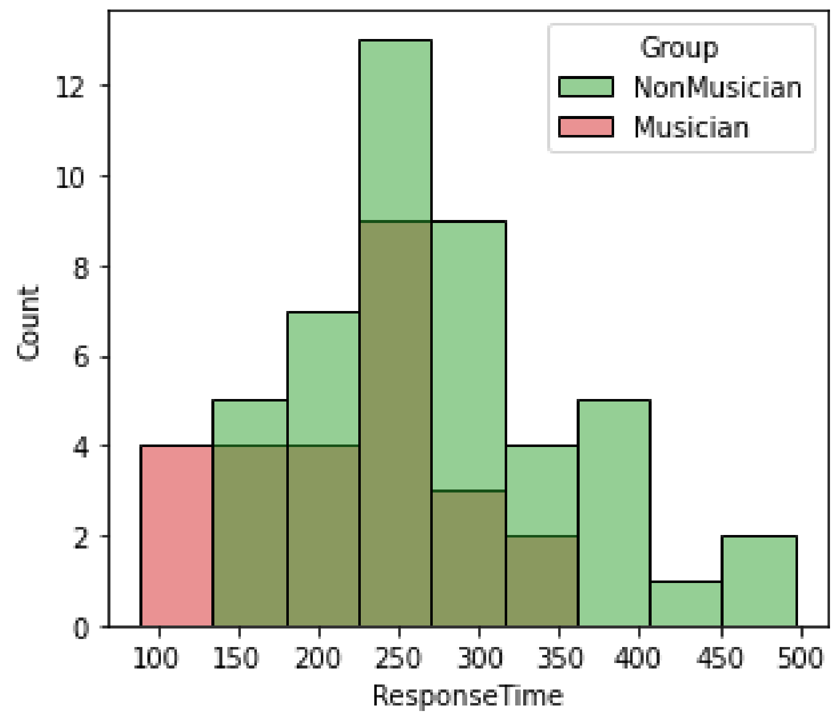


plot data by group: cumulative

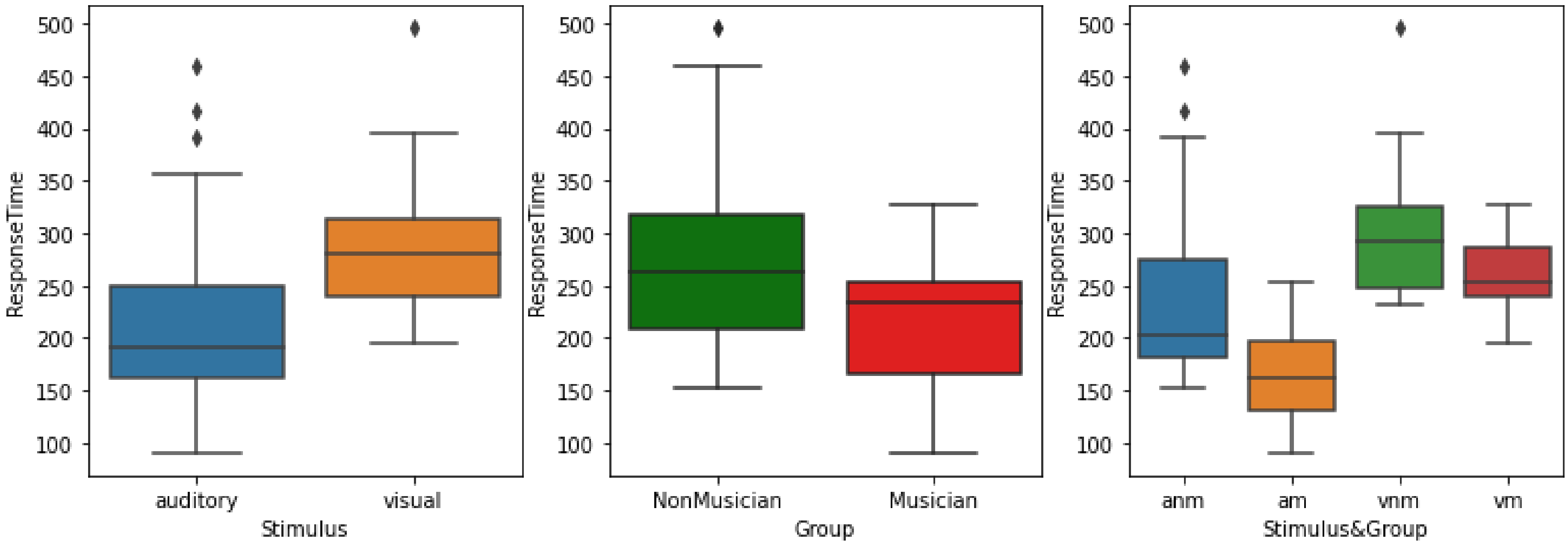


plot data by group: histogram

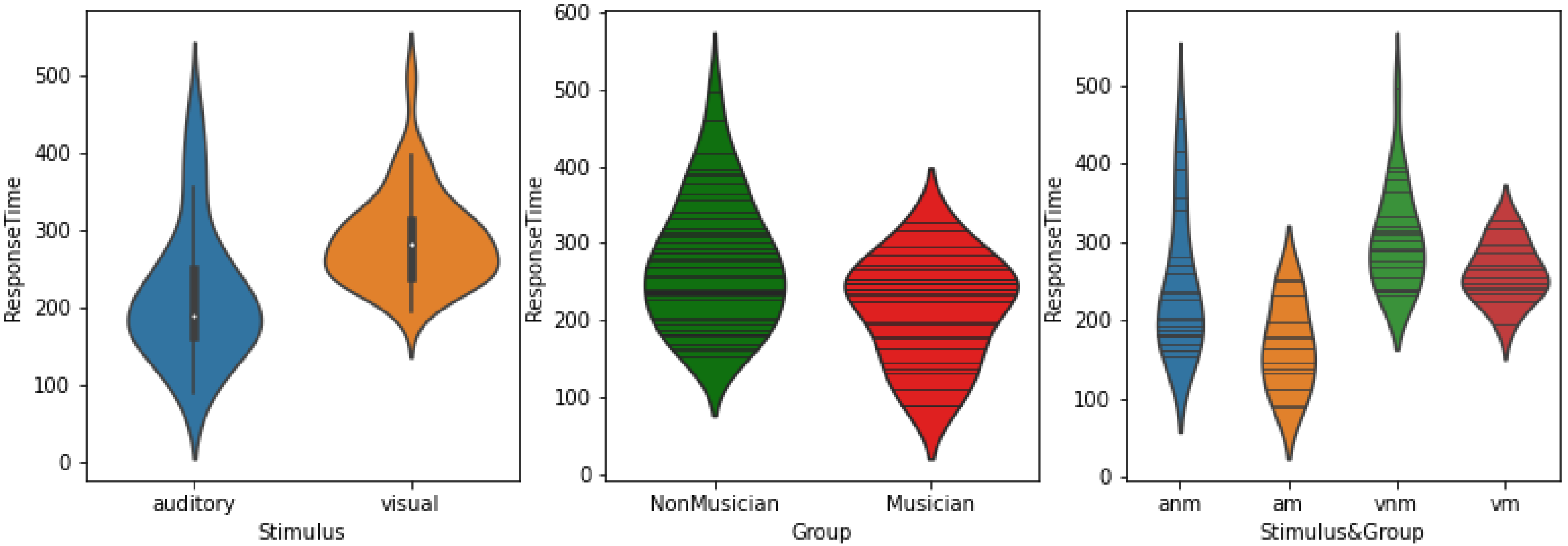




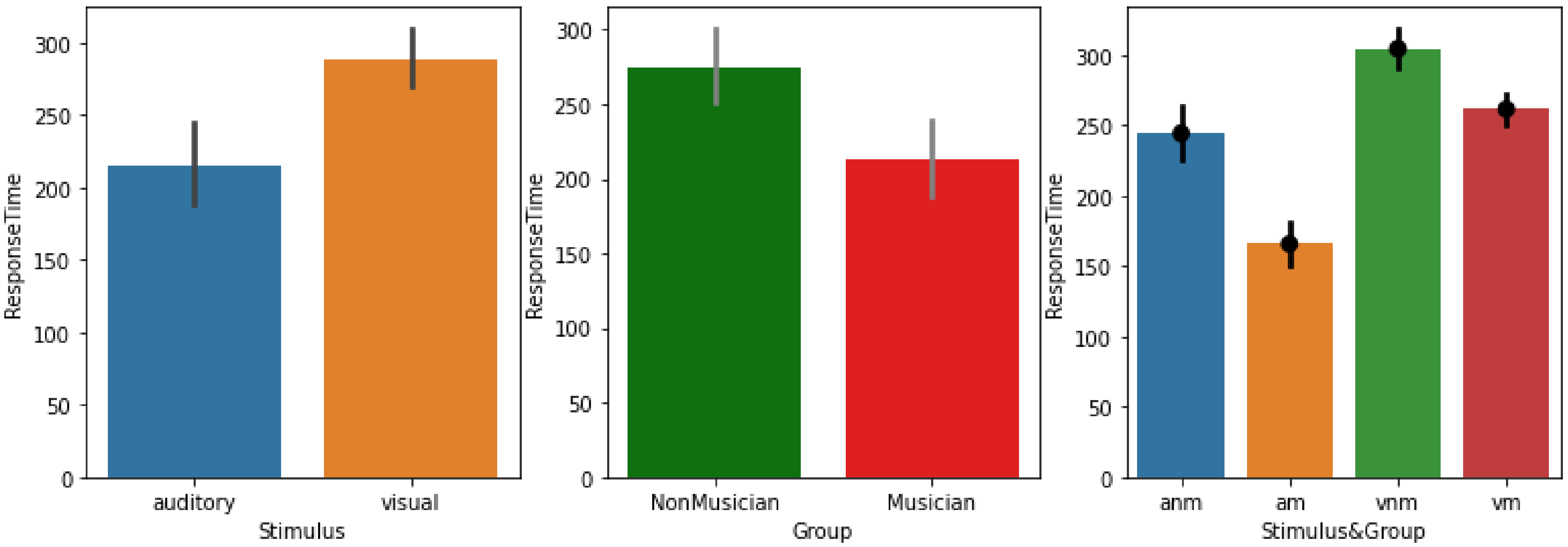
plot data by group: boxplot



plot data by group: violin plot



plot data by group: bar+errorbar



plot data by group: swarmplot