

# **Applied Data Analytics**

## **Descriptive Statistics**

### **Measures of Central Tendency: Cardinal data**

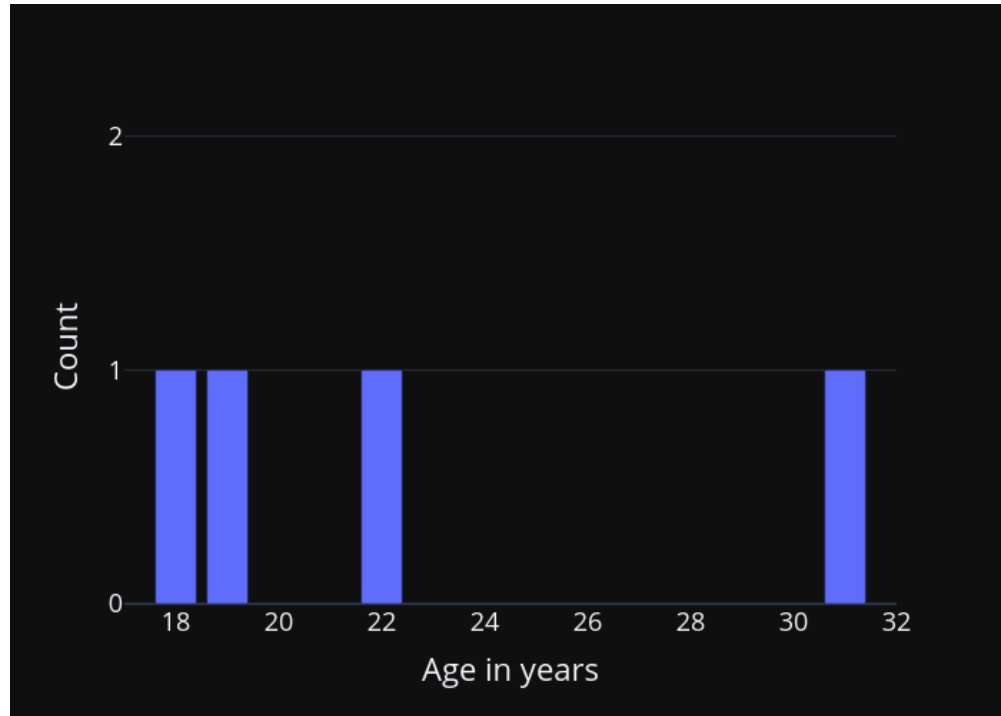
Hans-Martin von Gaudecker and Aapo Stenhammar

# Cardinal data

Example:

- Variable: Age in years
- Possible values: 0, 1, 2, ...
- Observed values: 18, 22, 19, 31

# Distribution



# Median: Definition

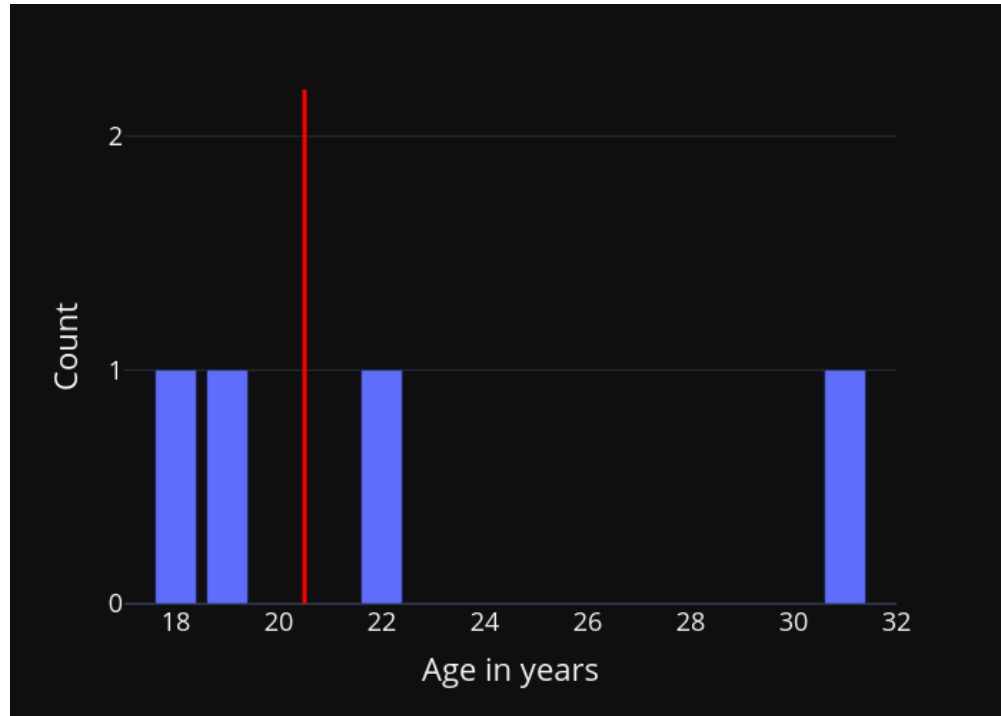
A value such that:

- at least half of the observations are higher or equal than the value
- at least half of the observations are lower or equal than the value

## Median: in practice

- Sort the data and find the *middle value*
- Even  $N$  and distinct values at  $N/2$  and  $N/2 + 1$ : Any value between the two, typically the average.
- Can be computed for any type of *ordered* data (ordinal and cardinal).

# Distribution with median



## Mean: Definition

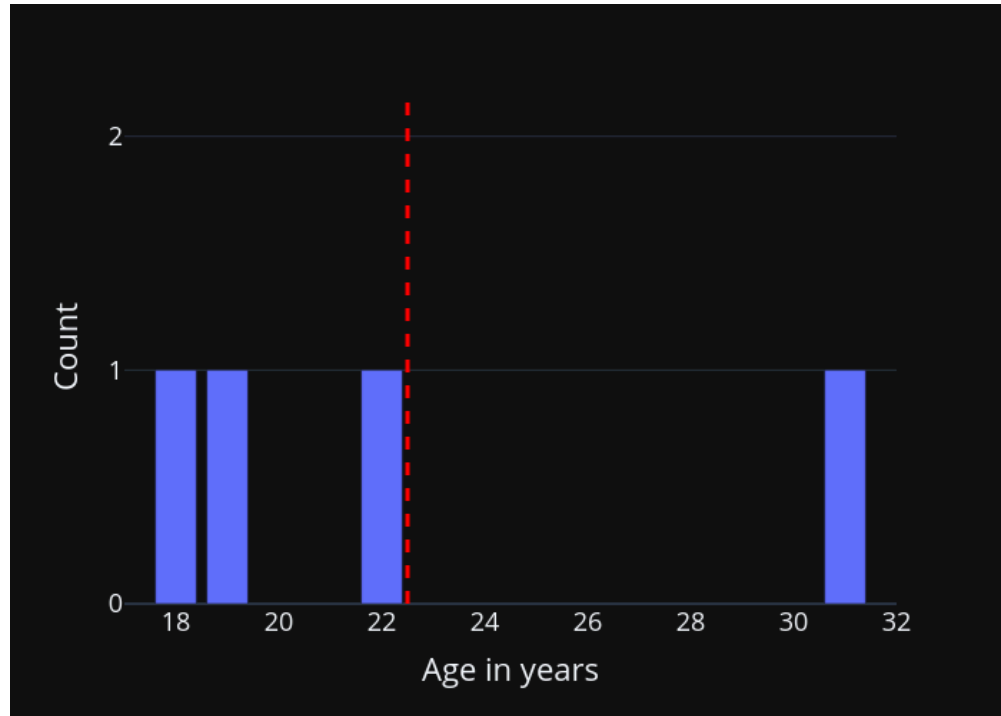
The mean is *the sum of all the values in the sample divided by the total number of values*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

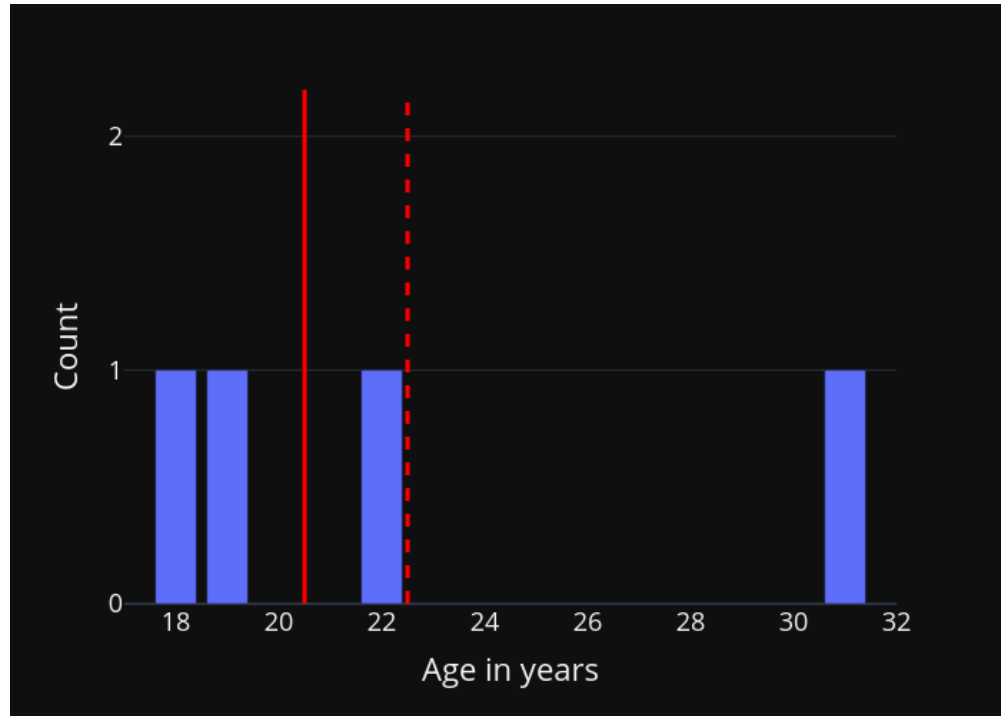
## Mean: in practice

- Sum up all values and divide by the number of observations
- Is influenced by all observations, median only by the middle one(s)

# Distribution with mean



# Distribution with median and mean





# Median and mean in pandas

```
[1] age.median()  
[1] np.float64(20.5)
```

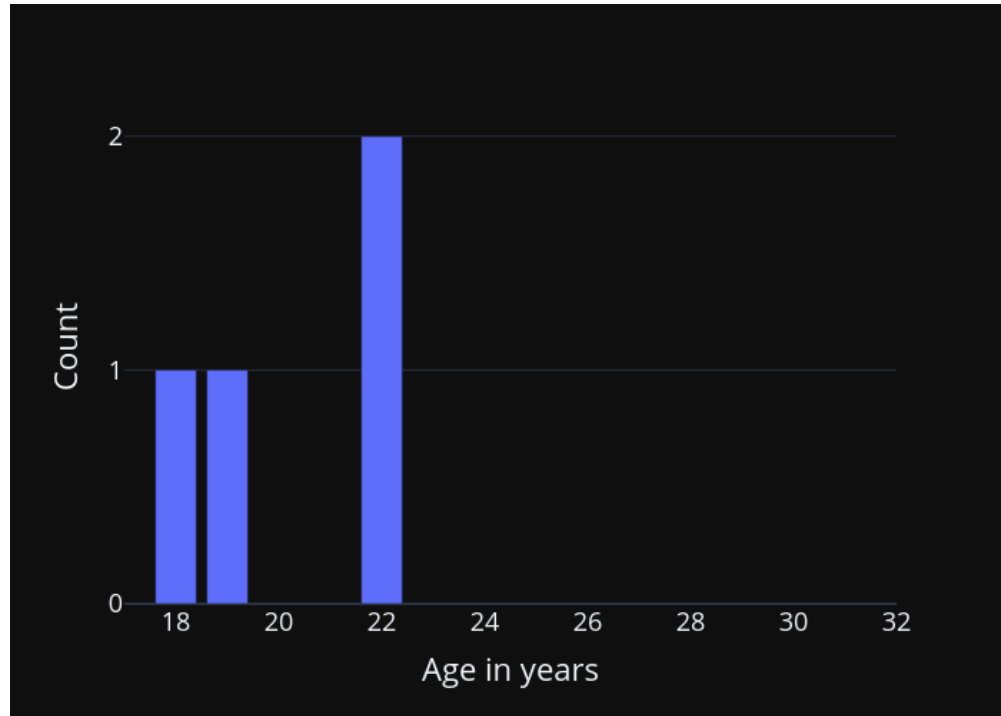
```
[2] age.mean()  
[2] np.float64(22.5)
```

- Just call methods with the respective name
- Median gives the average of the two middle values for even  $N$

## Median vs Mean: sensitivity to outliers.

- The median is less sensitive to outliers than the mean.
- Reason is that it is influenced by all observations, while the median only by the middle one(s).
- Example: Observed values: 18, 22, 19, 22

# Distribution



# Distribution with median and mean

