Applied Data Analytics

Statistics — Basics & location

Measures of Central Tendency: Ordinal data

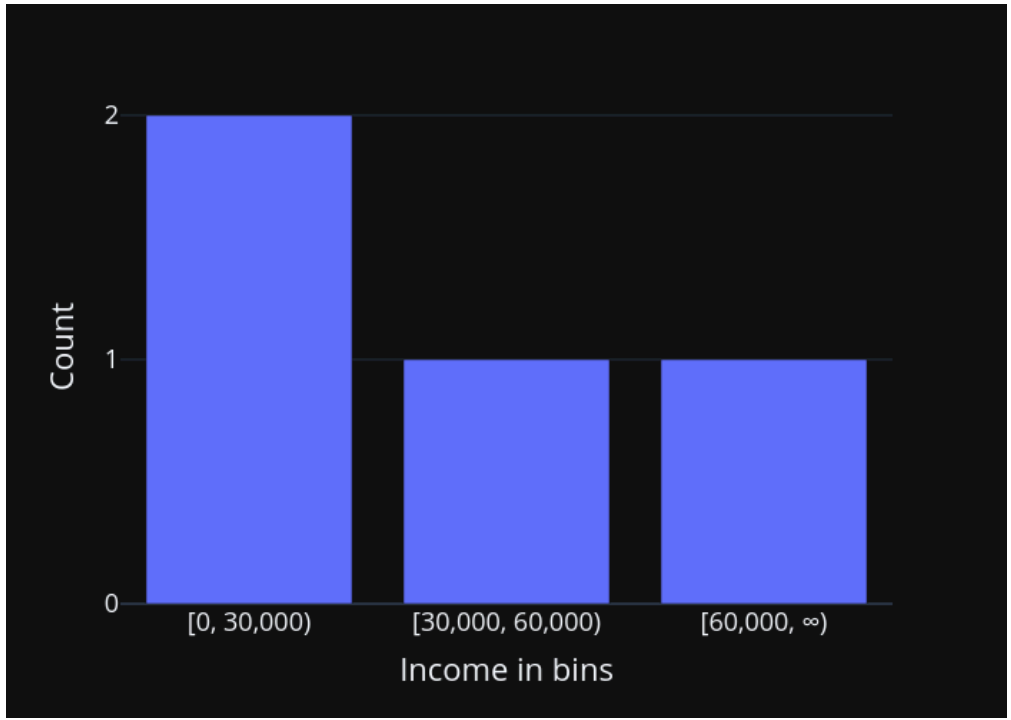
Hans-Martin von Gaudecker and Aapo Stenhammar

Ordinal, non-numeric data

Example:

- Variable: Annual Income in Euros, binned
- Possible values: $[0, 30000)$, $[30000, 60000)$, $[60000, \infty)$
- Observed values:
  - $[0, 30000)$
  - $[60000, \infty)$
  - $[30000, 60000)$
  - $[0, 30000)$

Distribution



Mode: Definition

The mode is the value that appears most frequently in the data.

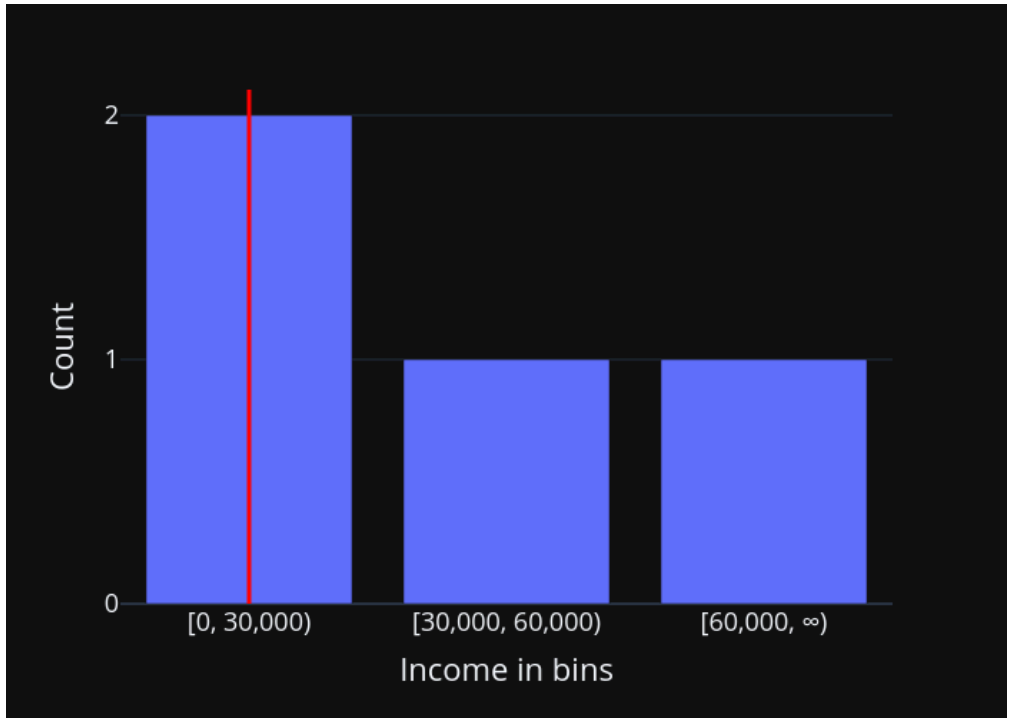
Mode: In practice

- Get a frequency distribution, take the maximum.
- Defined for any type of data (nominal, ordinal, cardinal).

Mode: Corner cases

- If more values appear with the same highest frequency, the data is *multimodal*.
- If no value appears more than once, the data has *no mode*.

Distribution with mode



Median: Definition

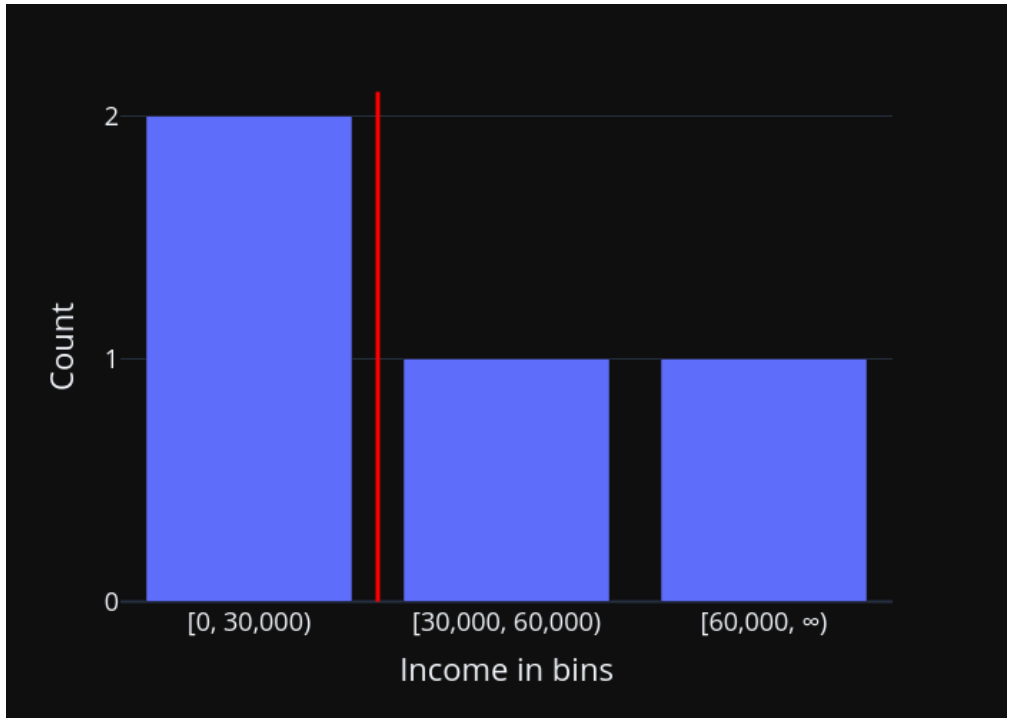
A value such that:

- 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 mode



Mode and median in pandas

```
[1] income.mode()
[1] 0    [0, 30000)
    dtype: category


[2] income.median()
[2] TypeError: 'Categorical' with
    dtype category does not support
    reduction 'median'
```

- Just call methods with the respective name
- Median only works for numerical data
- Reason is that it is not clear what should be returned if the median is not a category in the data.