Applied Data Analytics

Statistics — Basics & location

Histograms

Hans-Martin von Gaudecker and Aapo Stenhammar

Frequency distributions

A frequency distribution is a table that shows the frequency of various outcomes in a sample.

Categorical Data

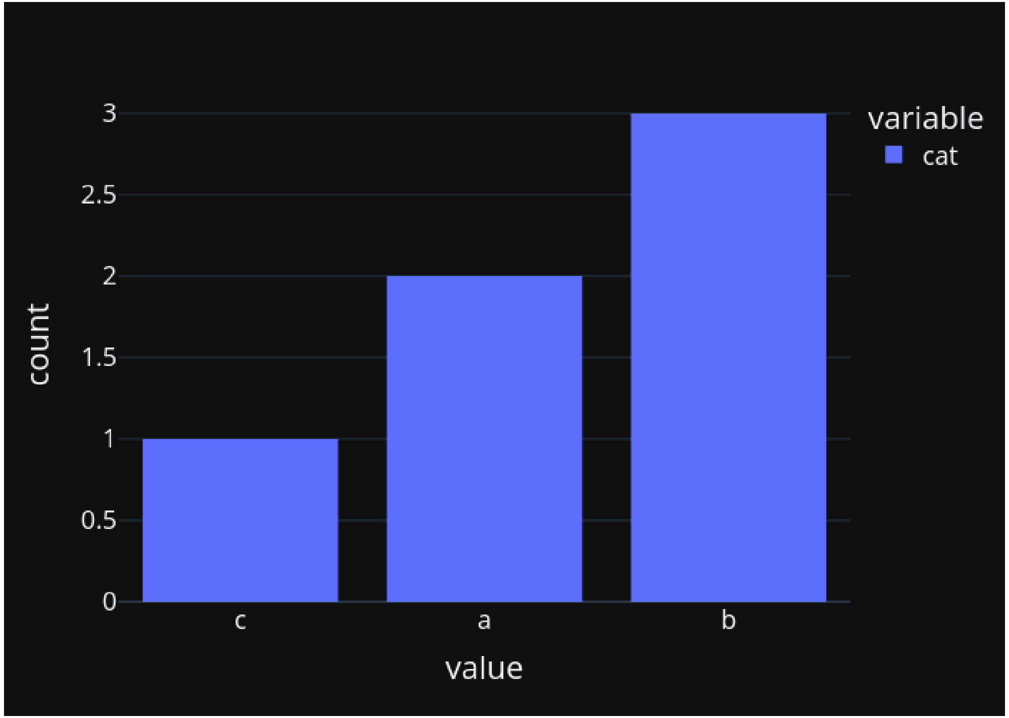
Raw data

	cat
0	c
1	a
2	b
3	a
4	b
5	b

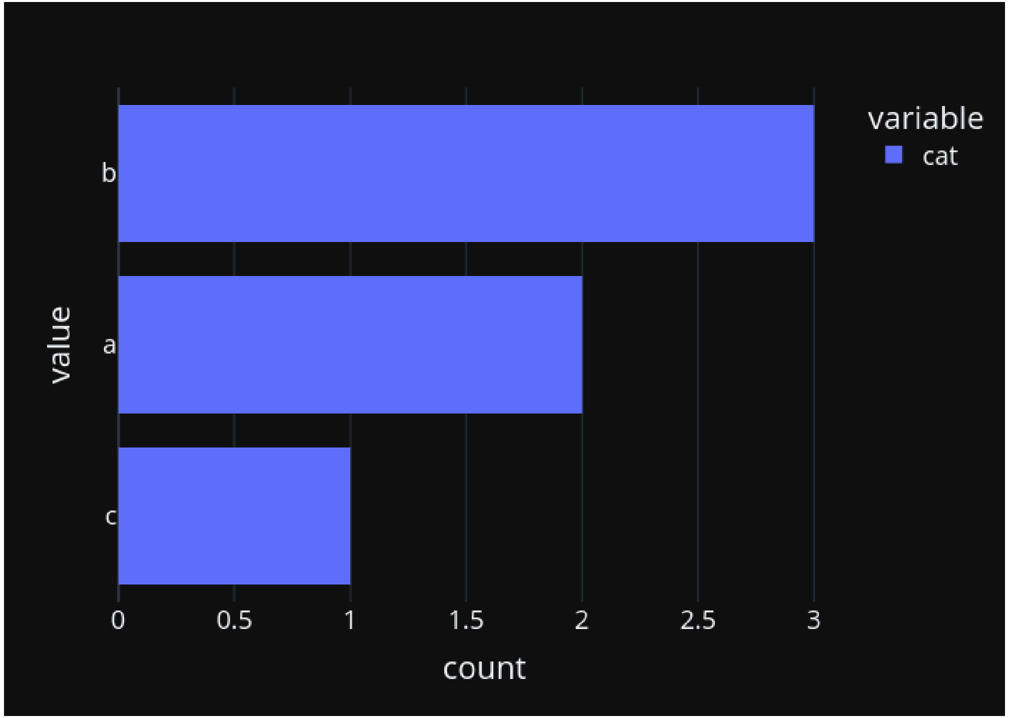
Frequency distribution

cat	count
b	3
a	2
c	1

Histogram



Better (for categorical data)



Continuous data

Raw data

	cont
0	1.57
1	0.09
2	1
3	2.9
4	1.25
5	1
6	0.35
7	2.3
8	2.15

Frequency distribution

cont	count
0.09	1
0.35	1
1	2
1.25	1
1.57	1
2.15	1
2.3	1
2.9	1

Continuous data, binned

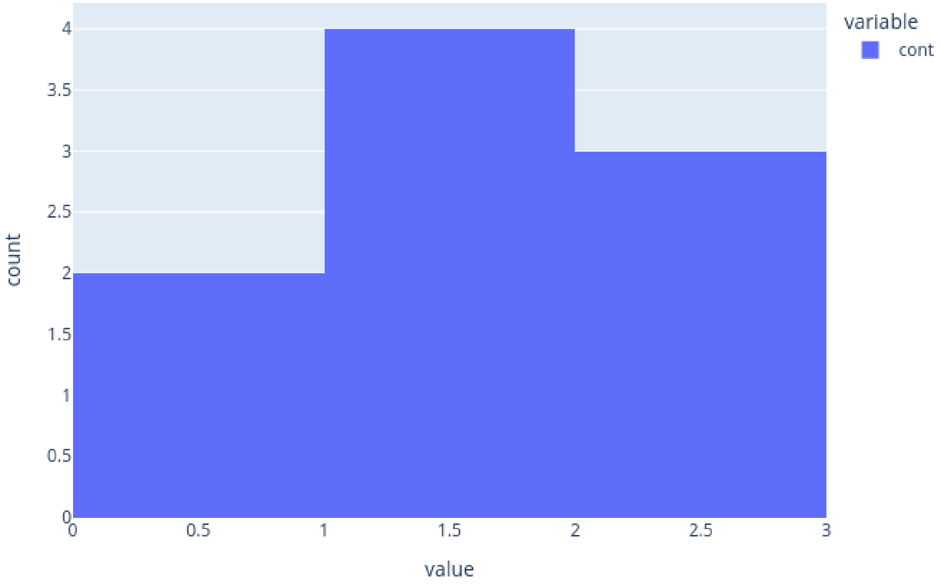
Raw data

	cont
0	1.57
1	0.09
2	1
3	2.9
4	1.25
5	1
6	0.35
7	2.3
8	2.15

Frequency distribution

cont_binned	count
$[0, 1)$	2
$[1, 2)$	4
$[2, 3)$	3

Histogram



Defining a bin

Formally, you just count observations that fulfil a certain condition:

$$\text{count} = \sum_{i=1}^{N} 1\{\text{lb} \leq x_i < \text{ub}\}$$

where $1\{\cdot\}$ is the indicator function, i.e.,

- it is 1 when the condition is fulfilled
- and 0 otherwise

Conditions for a histogram

$$\text{count} = \sum_{i=1}^{N} 1\{\text{lb} \leq x_i < \text{ub}\}$$

- Choose the set of all (lb, ub)-pairs so that each observation is counted exactly once, i.e.,
  - Minimum and maximum of $x_i,\ i \in \{1, ..., N\}$ are included
  - Bins are non-overlapping and there are no gaps
- Equal width bins are crucial for honest communication

Discrete data, i.e., integers

- Plotly defaults to treating them as continuous data
- Prerequisite for that being sensible: Data is ordered and gaps are meaningful (e.g., age in full years)
- If you need to treat them as categorical, you first count values and then make a bar chart from the resulting Series.