Probability Theory - Part 33

Statistics: descriptive statistics | inferential statistics

The rough idea:

sample — finite set of data points

"all the theoretically possible data" / population / real world ...

analyzing
summarizing
visualizing

$\underbrace{\text{analyzing, summarizing, visualizing}}_{\text{descriptive statistics}}$

inferential statistics: finding the underlying probability distribution

prediction

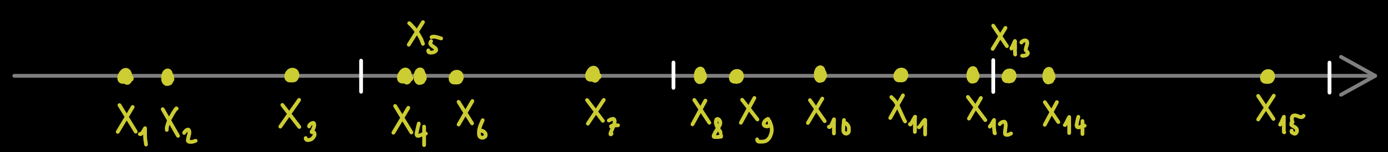
Definition: sample in $\mathbb{R}^d$: $(x_1, x_2, x_3, \dots, x_n)$ with $x_j \in \mathbb{R}^d$

$n \in \mathbb{N}$

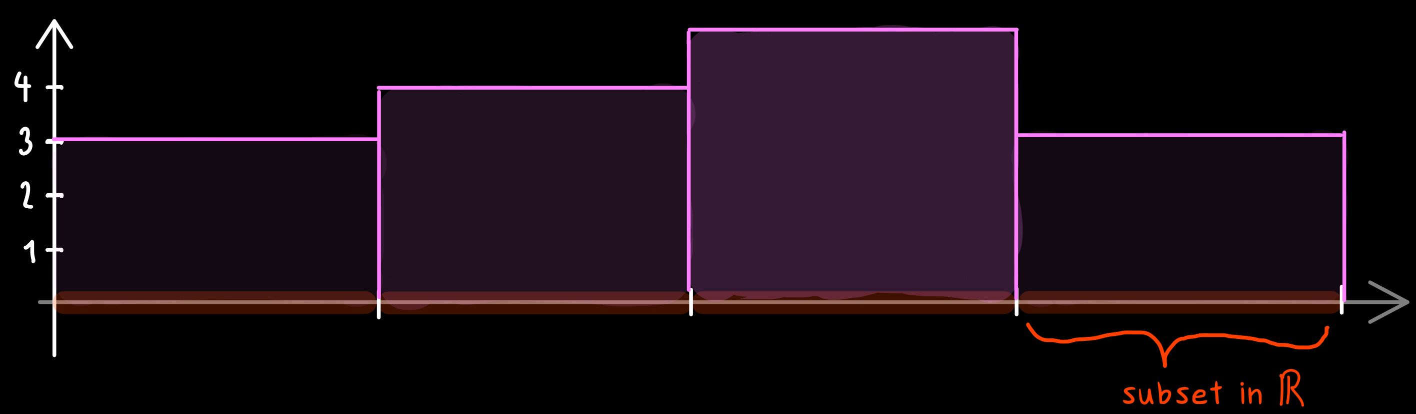
sample in $\mathbb{R}$: $(x_1, x_2, x_3, \dots, x_n)$ with $x_j \in \mathbb{R}$

can be ordered: $x_1 \leq x_2 \leq \dots \leq x_n$

visualization



histogram
(grouping)



Definition: For a given sample in $\mathbb{R}$ $(x_1, x_2, x_3, \dots, x_n)$ and subset $A \subseteq \mathbb{R}$,

we define:

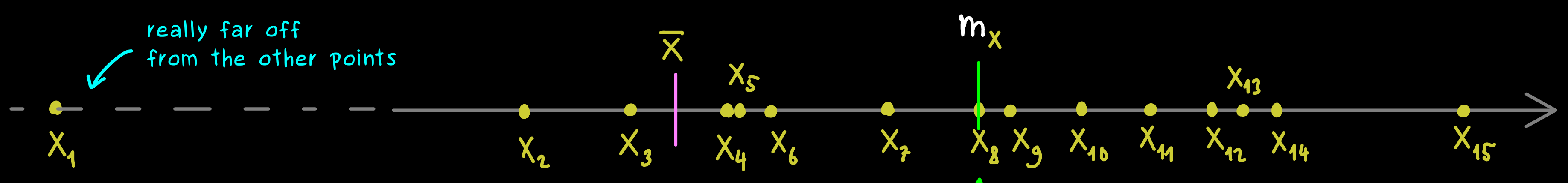
absolute frequency of A $\qquad f_{abs}(A) := \#\{k \in \mathbb{N} \mid x_k \in A\}$

relative frequency of A $\qquad f_{rel}(A) := \dfrac{\#\{k \in \mathbb{N} \mid x_k \in A\}}{n} \in [0,1]$

Definition: For a given sample in $\mathbb{R}$ $x := (x_1, x_2, x_3, \dots, x_n)$

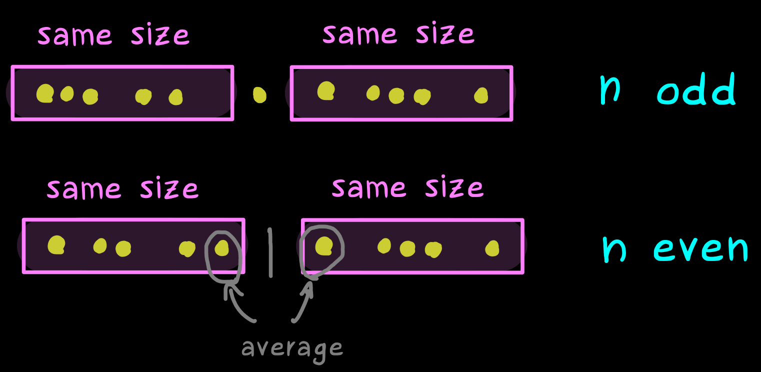
we define the sample mean:

$$\bar{x} := \frac{1}{n}\sum_{k=1}^{n} x_k$$



median: middle point of the sample

$$m_x := \begin{cases} x_{\frac{n+1}{2}}, & n \text{ odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & n \text{ even} \end{cases}$$



unbiased sample variance: $\qquad s_x^2 := \dfrac{1}{n-1}\displaystyle\sum_{k=1}^{n}(x_k - \bar{x})^2$

makes it unbiased