

CS 361

Probability and Statistics for Computer Science

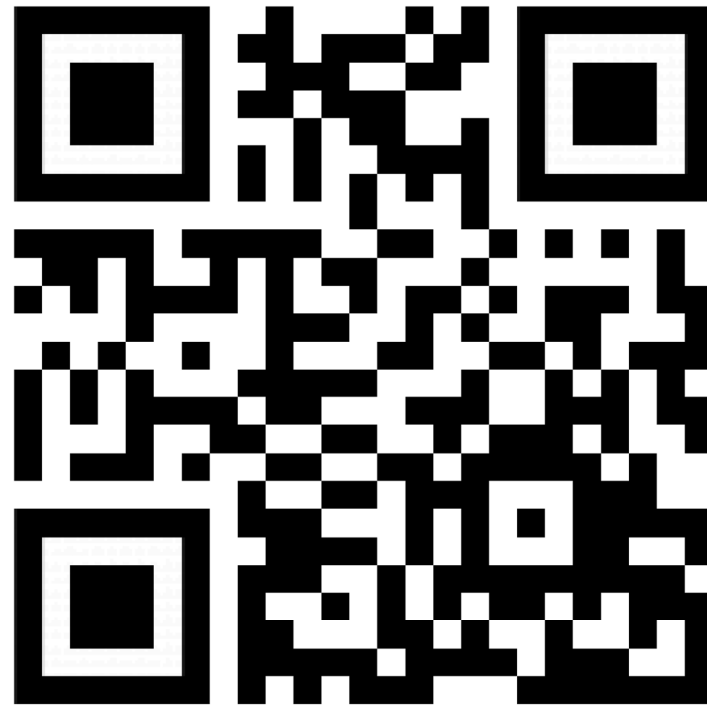
*mathematics
of randomness*

*collection, summary
and analysis of data*

David Varodayan
dpv@illinois.edu

A game of chance (win prizes!)

go.illinois.edu/cs361



About CS 361

courses.engr.illinois.edu/cs361



Datasets

A dataset $\{x\}$ is a set of N items x_1, \dots, x_N , each of which is a tuple

e.g. our dataset consists of 121 4-tuples

Types of data

- Categorical: takes on a small set of prescribed values

e.g. Yes/No

- Ordinal: categorical data where items can be put in order

e.g. number of dice rolls until 6 appears

- Continuous: takes on any numerical value within a range

e.g. amount you could bet

Visualizing data

- Tables
- Bar charts
- Histograms
- Conditional histograms

Examples in today's Jupyter notebook

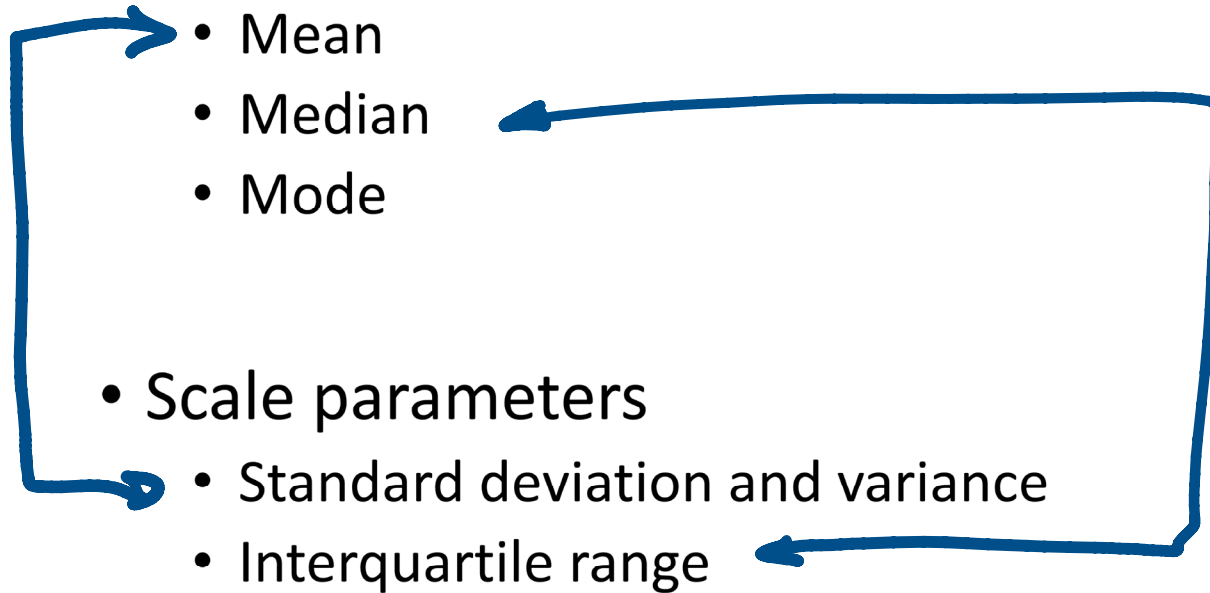
Summarizing data

- Location parameters

- Mean
- Median
- Mode

- Scale parameters

- Standard deviation and variance
- Interquartile range



Mean

$$\text{mean}(\{x_i\}) = \frac{1}{N} \sum_{i=1}^N x_i$$

Properties of the mean

- Scaling data scales the mean

$$\text{mean}(\{kx_i\}) = k\text{mean}(\{x_i\})$$

- Translating data translates the mean

$$\text{mean}(\{x_i + c\}) = \text{mean}(\{x_i\}) + c$$

Less obvious properties of the mean

- The sum of signed distances from the mean is 0

$$\sum_{i=1}^N (x_i - \text{mean}(\{x_i\})) = 0$$

- The mean minimizes the sum of squared distances from the data

$$\arg \min_{\mu} \sum_{i=1}^N (x_i - \mu)^2 = \text{mean}(\{x_i\})$$

Standard deviation

$\text{std}(\{x_i\})$

$$= \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \text{mean}(\{x_i\}))^2}$$

$$= \sqrt{\text{mean} \left(\left\{ (x_i - \text{mean}(\{x_i\}))^2 \right\} \right)}$$

Properties of standard deviation

- Scaling data scales the standard deviation

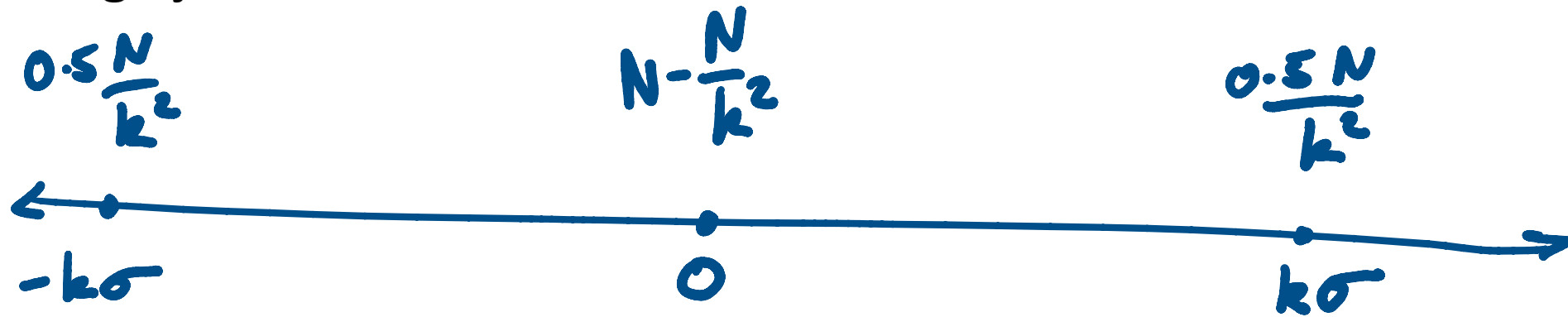
$$\text{std}(\{kx_i\}) = |k|\text{std}(\{x_i\})$$

- Translating the data does **not** change the standard deviation

$$\text{std}(\{x_i + c\}) = \text{std}(\{x_i\})$$

Standard deviation: Chebyshev's inequality

- At most $\frac{N}{k^2}$ items are k standard deviations from the mean
- Rough justification: Assume the data have zero mean ...



$$\text{std} = \sqrt{\frac{1}{N} \left[\left(N - \frac{N}{k^2} \right) 0^2 + \frac{N}{k^2} (k\sigma)^2 \right]} = \sigma$$

Variance = (standard deviation)²

$$\text{var}(\{x_i\}) = \frac{1}{N} \sum_{i=1}^N (x_i - \text{mean}(\{x_i\}))^2$$

- Scaling and translation properties similar to standard deviation

$$\text{var}(\{kx_i\}) = k^2 \text{var}(\{x_i\})$$

$$\text{var}(\{x_i + c\}) = \text{var}(\{x_i\})$$