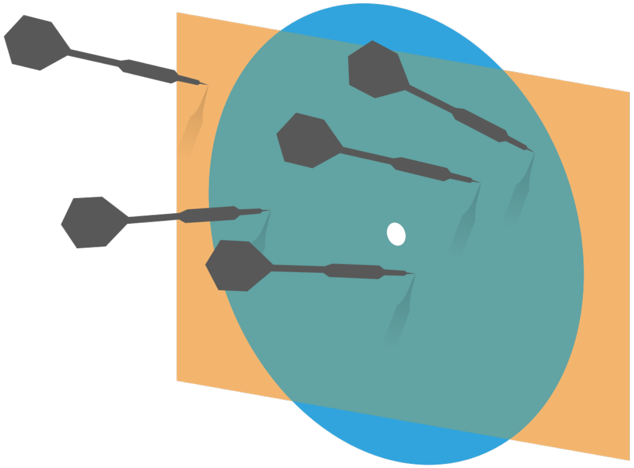
Probability and Statistics for Computer Science



Credit: wikipedia

"The statement that "The average US family has 2.6 children" invites mockery" – Prof. Forsyth reminds us about critical thinking

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 8.27.2021

Last lecture

- Welcome/Orientation
- Big picture of the contents
- Lecture 1 - Data Visualization & Summary (I)
- **Orientation quiz due today**

Warm up question:

- What kind of data is a letter grade?
- What do you ask for usually about the stats of an exam with numerical scores?

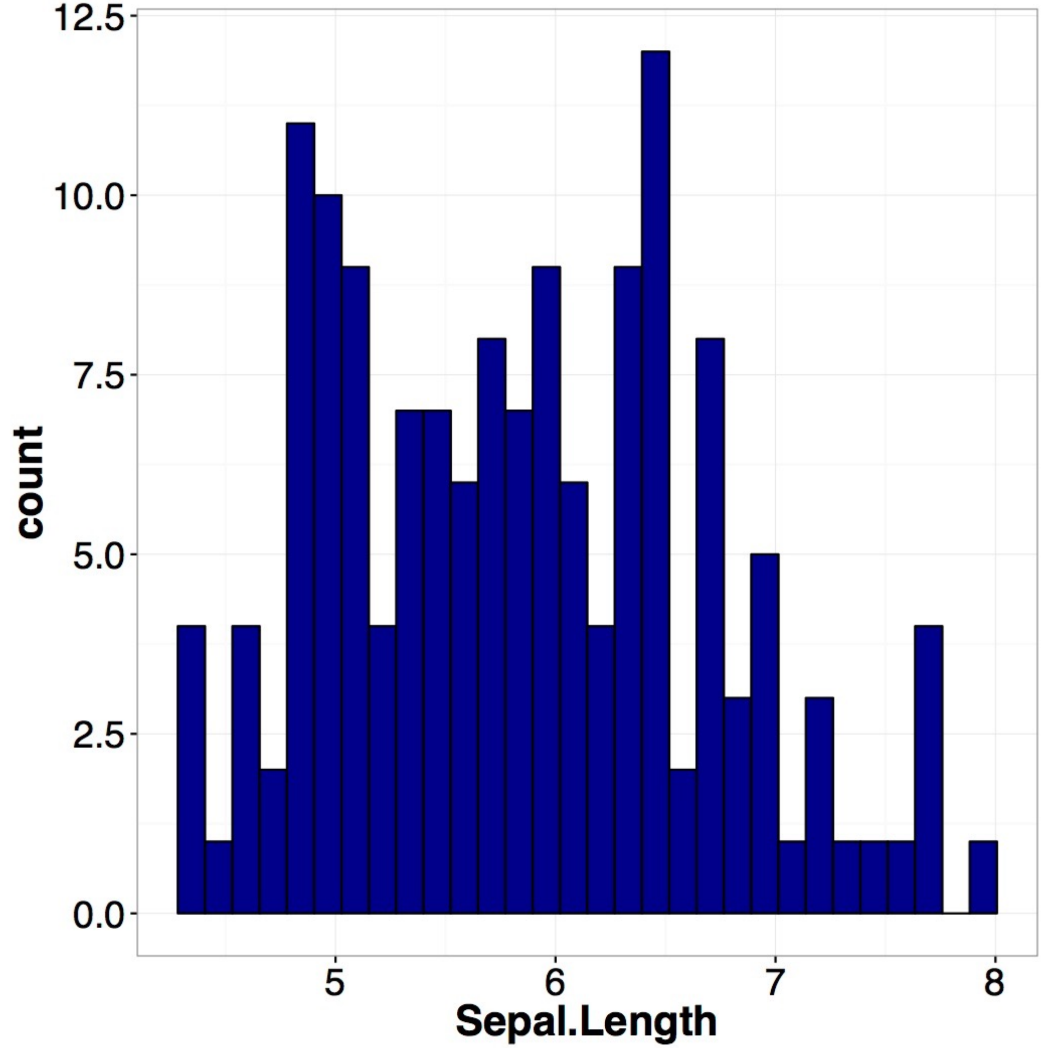
Objectives

- **Histograms**
- **Grasp** Summary Statistics

Visualizing Data with Histogram

- Histogram

A set of bars that are organized by bins that contains numerical data (discrete or continuous)



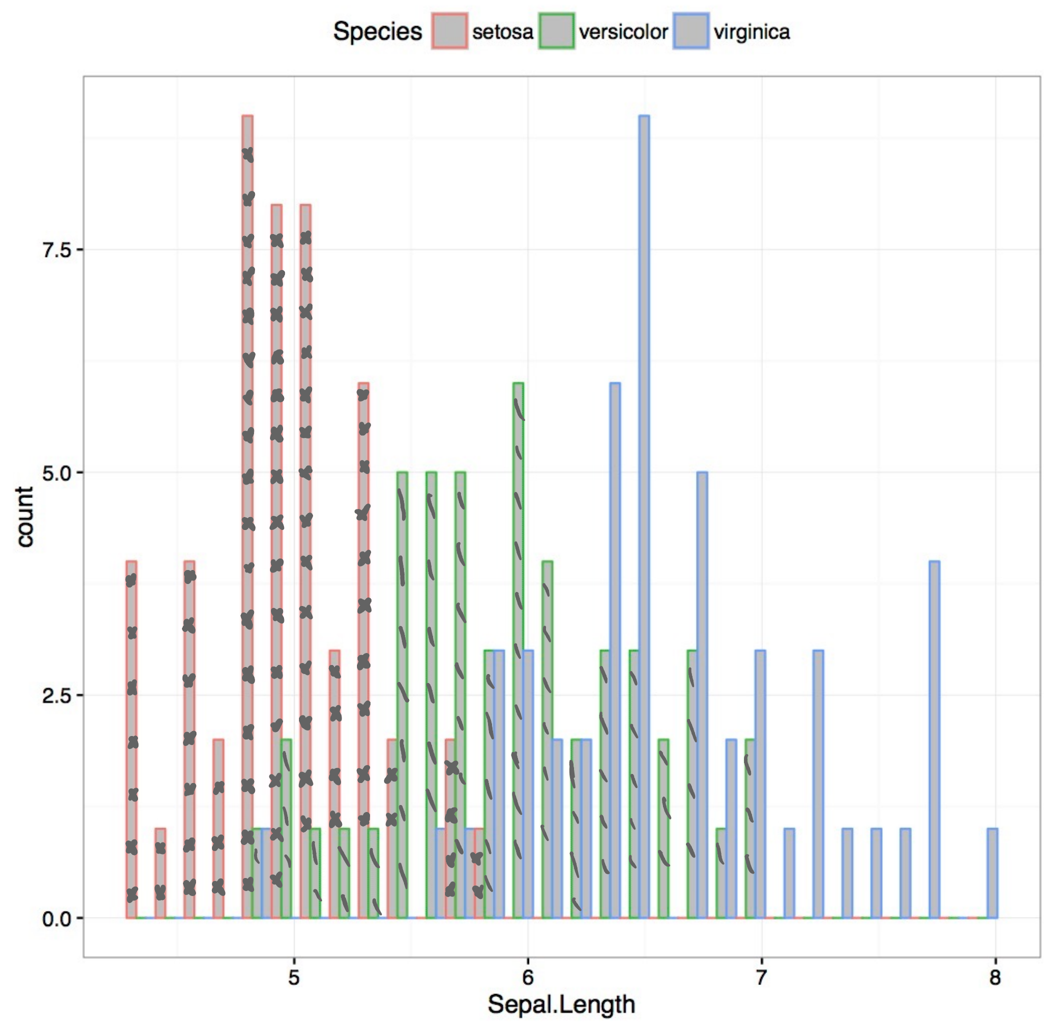
Data: “iris”

Visualizing Data with Histogram (II)

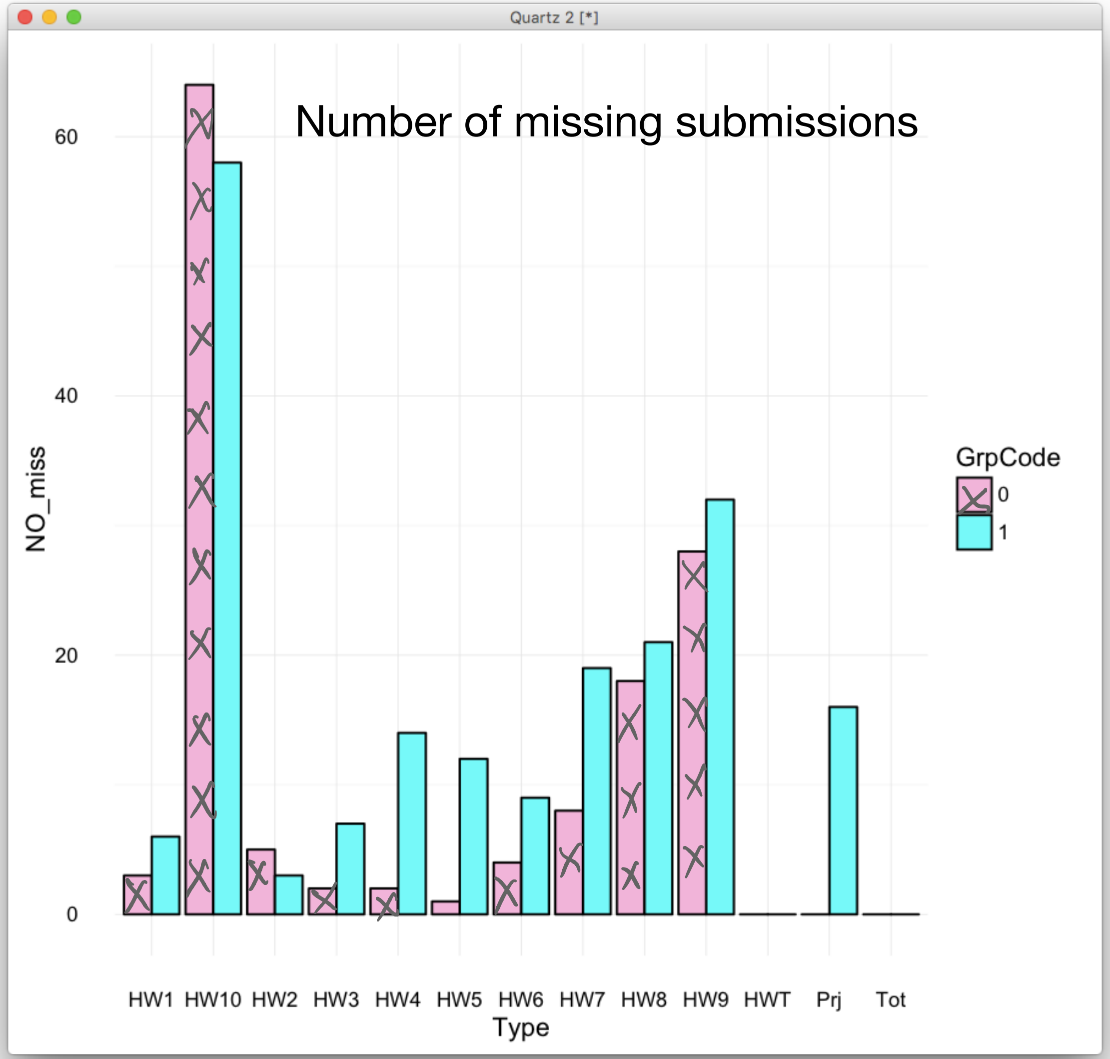
- Conditional histogram

Histogram generated by subsets of the data

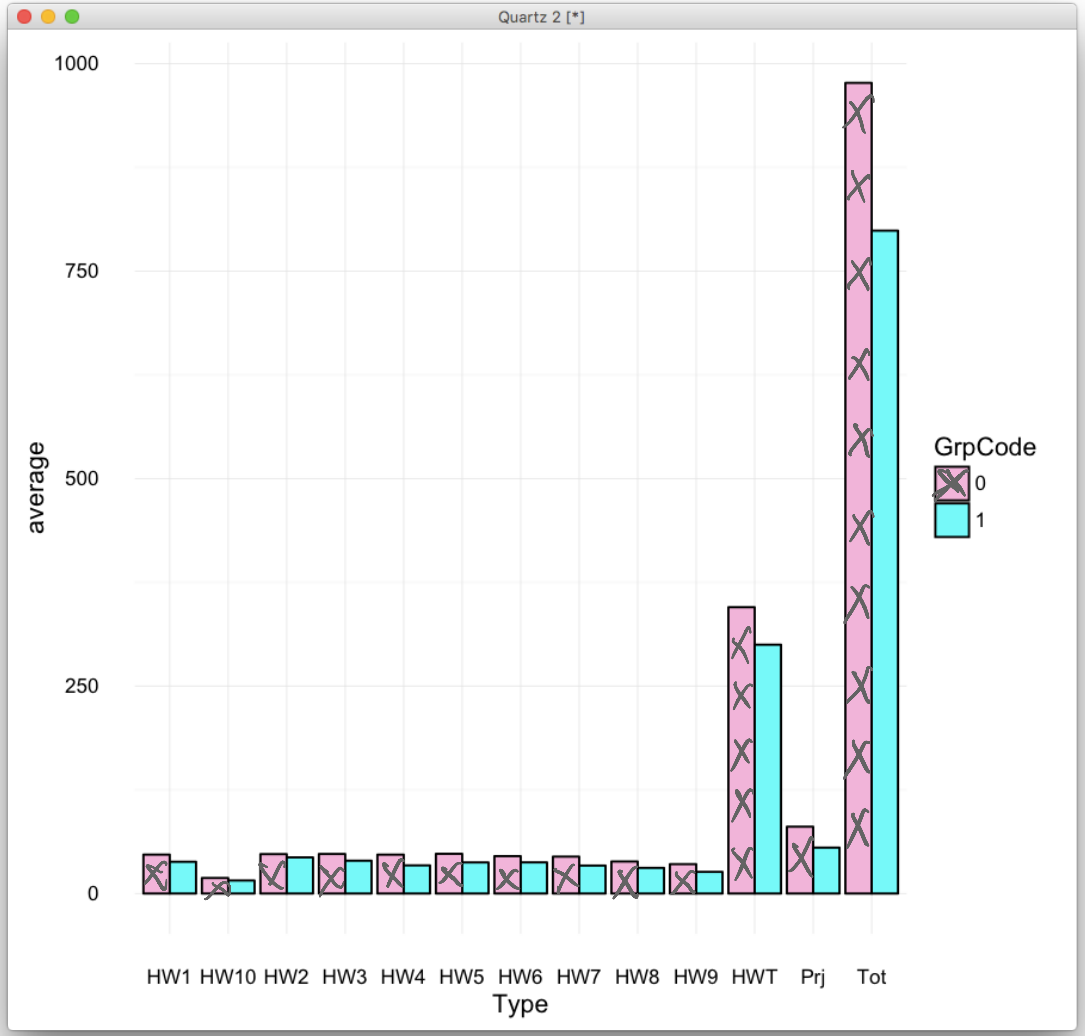
Data: “iris”



Which group has the higher total scores?



Which group has the higher total scores?



Summarizing 1D continuous data

For a data set {x} or annotated as $\{x_i\}$, we summarize with:

N items

- Location Parameters

Mean (μ), Median, Mode

- Scale parameters

Standard deviation (σ), Variance (σ^2)

Interquartile range

Summarizing 1D continuous data

- Mean

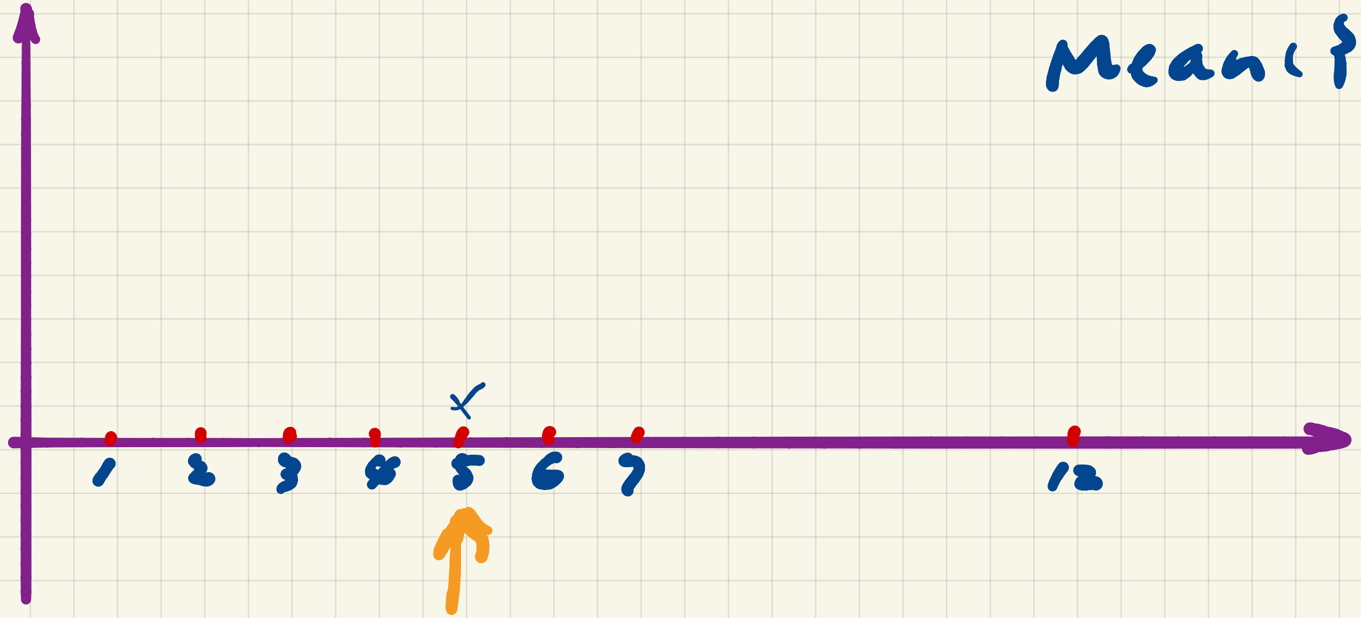
$$mean(\{x_i\}) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It's the centroid of the data geometrically,
by identifying the data set at that point, you find the center of balance.

$\{x_i\} \quad i \in [1, 8]$

$\{x_i\} = 1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6, \quad 7, \quad 12$

$\mathrm{Mean}(\{x_i\}) = 5$



Properties of the mean

- Scaling data scales the mean

$mean(\{k x_i + c\})$

$= k \cdot mean(\{x_i\})$

$$mean(\{k \cdot x_i\}) = k \cdot mean(\{x_i\})$$

$+ c$

- Translating the data translates the mean

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

Less obvious properties of the mean

- The signed distances from the mean

  sum to 0

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

- The mean minimizes the sum of the squared distance from any real value

$$\underset{\mu}{argmin} \sum_{i=1}^{N}(x_i - \mu)^2 = mean(\{x_i\})$$

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

$$\text{LHS} = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \underbrace{\text{mean}(\{x_i\})}_{\to\ \text{Const}}$$

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

$$= \sum_{i=1}^{N} x_i - \cancel{N} \cdot \frac{\sum_{i=1}^{N} x_i}{\cancel{N}}$$

$$= \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

Proof: $\underset{\mu}{\text{Argmin}}\left(\sum_{i=1}^{N}(x_i-\mu)^2\right) = \text{mean}(\{x_i\})$

$\underset{\mu}{\text{Argmin}}$: Argument μ that minimizes the function that follows

LHS $= \hat{\mu}$ -> the special μ that minimizes $f(\mu)=\sum_{i=1}^{N}(x_i-\mu)^2$

To find $\hat{\mu}$, set $\frac{df(\mu)}{d\mu} = 0$ & Solve it

One way is to use the Chain rule

$$f(\mu) = \sum_{i=1}^{N} h(\mu) = \sum_{i=1}^{N} g^2(\mu) \qquad g = x_i - \mu$$

$$\frac{df}{d\mu} = \frac{d\sum h}{d\mu} = \sum \frac{dh}{d\mu} = \sum_{i=1}^{N} \frac{dh}{dg}\cdot\frac{dg}{d\mu}$$

Proof: $\underset{\mu}{\operatorname{Argmin}} \left(\sum_{i=1}^{N} (x_i - \mu)^2 \right) = \operatorname{mean}(\{x_i\})$

$$\frac{d f(\mu)}{d\mu} = \sum \frac{dh}{dg} \frac{dg}{d\mu} = \sum 2g \cdot (-1) = 0$$

$h = g^2$

$g = x_i - \mu$

$$\Rightarrow \sum_{i=1}^{N} g = 0$$

$$\Rightarrow \sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\sum_{i=1}^{N} x_i - N \cdot \mu = 0$$

$$\therefore \hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$= \operatorname{mean}(\{x_i\})$$

$$\frac{d^2 f(\mu)}{d^2 \mu} ?$$

Q1:

- What is the answer for

$$mean(\{mean(\{x_i\})\})\ ?$$

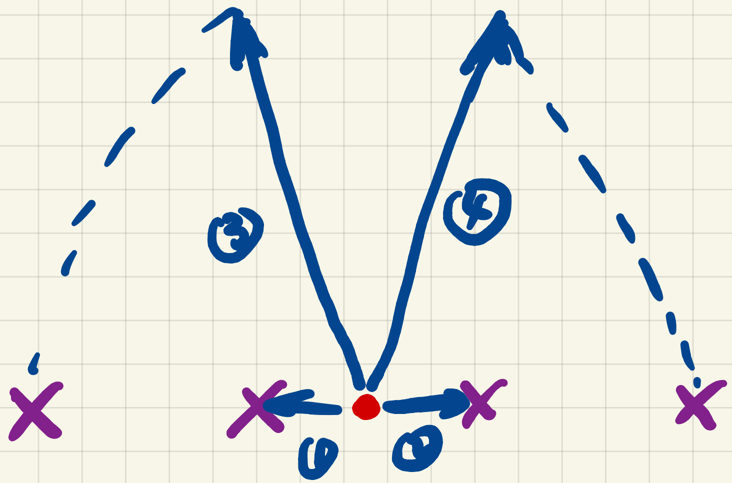
$\boxed{\text{A.}}$ $mean(\{x_i\})$ B. unsure C. 0

Standard Deviation (σ)

- The standard deviation

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

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



How much the data spreads out wrt mean

$$std = \sqrt{\frac{1}{4}\sum_{i=1}^{4} d_i^2}$$

Q2. Can a standard deviation of a dataset be -1?

A. YES

B. NO

Properties of the standard deviation

- Scaling data scales the standard deviation

$$std(\{k \cdot x_i\}) = |k| \cdot std(\{x_i\})$$

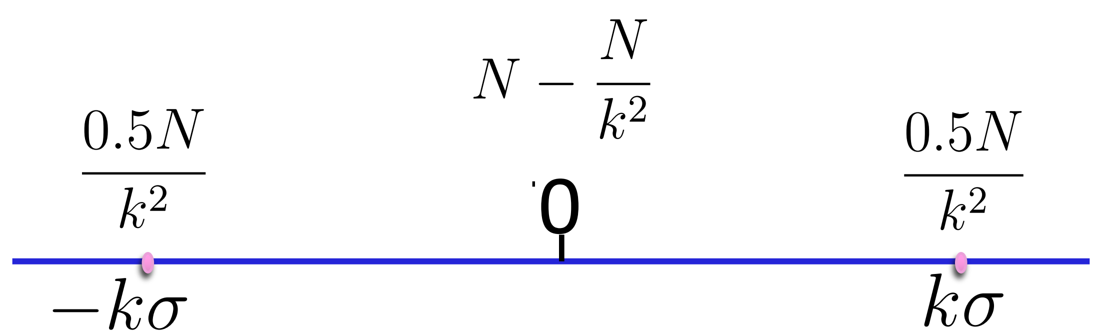
- Translating the data does **NOT** change the standard deviation

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

Standard deviation: Chebyshev's inequality (1st look)

- At most $\frac{N}{k^2}$ items are k standard deviations (σ) away from the mean

- Rough justification: Assume mean =0



$\{0, 0, \cdots 0, k\sigma, k\sigma, \cdots k\sigma, -k\sigma, -k\sigma, \cdots -k\sigma\}$

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

Variance (σ^2)

- Variance = (standard deviation)^2

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

- Scaling and translating similar to standard deviation

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

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

Q3: Standard deviation

- What is the value of

  $std(\{mean(\{x_i\})\})$?

  A. 0 B. 1 C. unsure

Standard Coordinates/normalized data

- The *mean* tells where the data set is and the *standard deviation* tells how spread out it is. If we are interested only in comparing the shape, we could

  define:

  ! every

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

- We say $\{\widehat{x_i}\}$ is in standard coordinates

Q4: Mean of standard coordinates

- mean($\{\widehat{x_i}\}$) is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

Q5: Standard deviation (σ) of standard coordinates

- Std($\{\widehat{x_i}\}$) is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

Q6: Variance of standard coordinates

- Variance of $\{\widehat{x_i}\}$ is:

  A. 1 B. 0 C. unsure

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

Q7: Estimate the range of data in standard coordinates

- Estimate as close as possible, 90% data is within:

  A. [-10, 10]

  B. [-100, 100]

  C. [-1, 1]

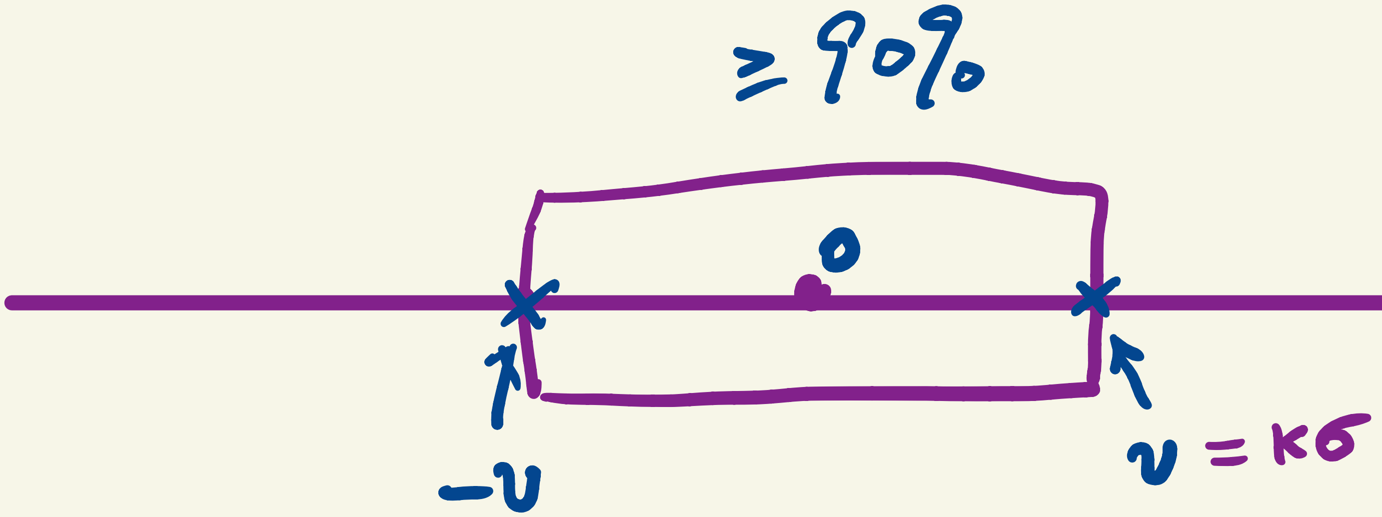
  D. [-4, 4]

  E. others

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})}$$

$$\frac{\frac{N}{k^2}}{N} = \frac{1}{k^2} \le 10\%$$

$\{\hat{x}_i\}$



$k\sigma = k$

$\because \sigma(\{\hat{x}_i\}) = 1$

Standard Coordinates/normalized data to $\mu=0$, $\sigma=1$, $\sigma^2=1$

- Data in standard coordinates always has

  mean = 0; standard deviation =1;

  variance = 1.

- Such data is unit-less, plots based on this sometimes are more comparable

- We see such normalization very often in statistics

Median

- We first sort the data set $\{x_i\}$
- Then *if* the number of items N is odd

  median = middle item's value

  if the number of items N is even

  median = mean of middle 2 items' values

Properties of Median

- Scaling data scales the median

$$median(\{k \cdot x_i\}) = k \cdot median(\{x_i\})$$

- Translating data translates the median

$$median(\{x_i + c\}) = median(\{x_i\}) + c$$

Percentile

- k^{th} percentile is the value relative to which k% of the data items have smaller or equal numbers
- Median is roughly the 50^{th} percentile

$\{1, 2, 3, 4, 5, 6, 7, 12\}$

75th percentile = ? 6

$\neq$ 75%

$\neq$ 0.75

Interquartile range

- iqr = (75th percentile) - (25th percentile)
- Scaling data scales the interquartile range

$$iqr(\{k \cdot x_i\}) = |k| \cdot iqr(\{x_i\})$$

- Translating data does **NOT** change the interquartile range

$$iqr(\{x_i + c\}) = iqr(\{x_i\})$$

Assignments

- **HW1** due Mon. Sept. 6.
- **Quiz 1 (open 4:30pm 9/1 next Wed until Sat.9/4)**
- Reading upto Chapter 2.1
- Next time: more summary statistics and correlation coefficient

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See You!