Working with data in your research and paper

Ioannis Konstantinidis

Sr. Researcher, Dept. of Computer Science

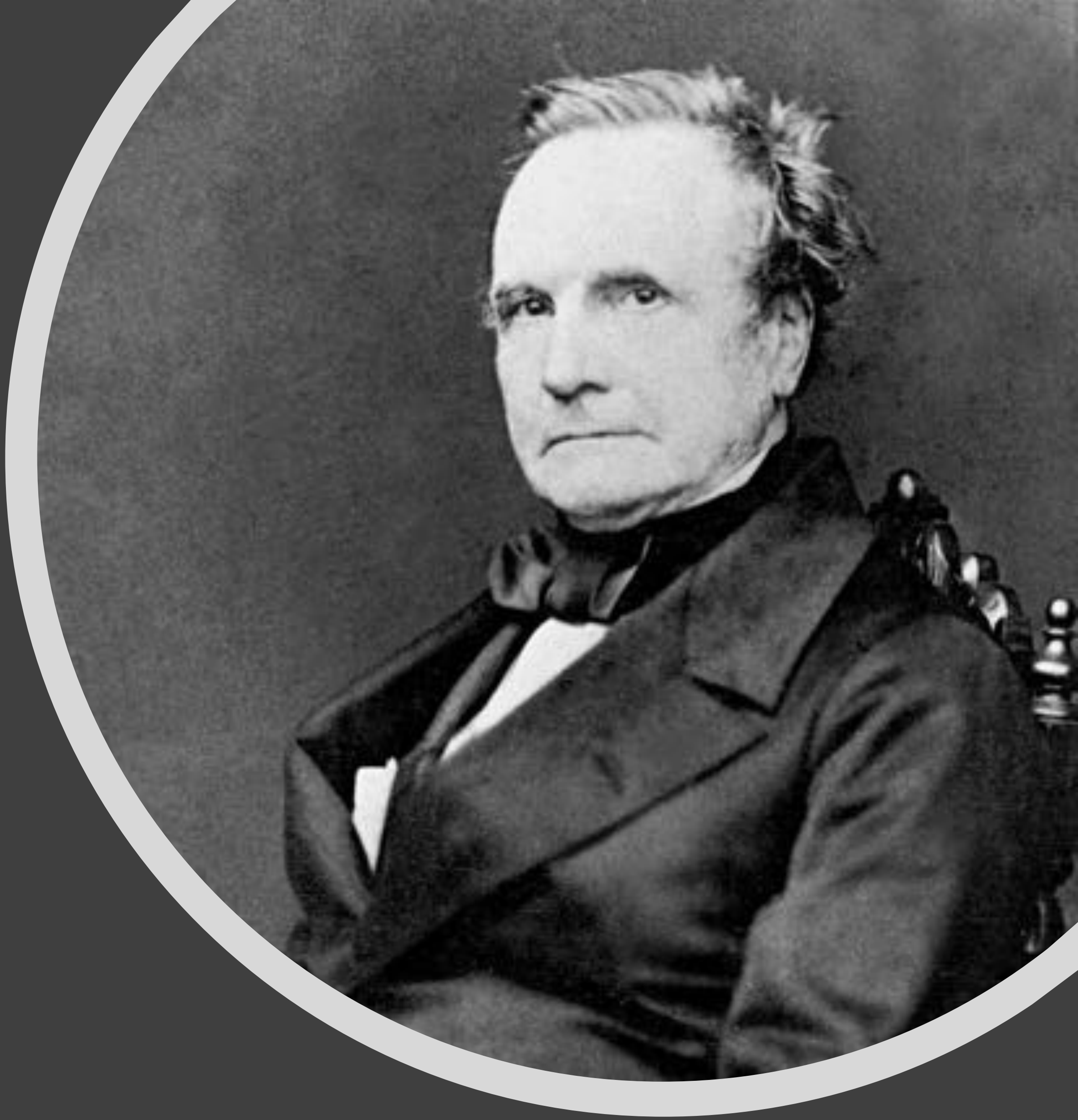
ikonstantinidis@uh.edu

Nutrition Facts
Serving Size
Servings Per Container

These slides were manufactured on equipment that processes words. May contain typos, mistakes, or omissions.

On two occasions I have been asked,—”Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?”
... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.

Charles Babbage (1791-1871) *Passages from the Life of a Philosopher*, ch. 5 “Difference Engine No. 1” (1864)



Does

- the statistical summary say what you *think* it says?
- the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?

STATISTICAL SUMMARIES

Congratulations!

Your dataset summaries look right

But does your dataset contain “wrong figures”?

Does

- ➢ the statistical summary say what you *think* it says?
- the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?

If your weight is **average**, then

A. You are as likely to run into someone that weighs more than you as you are to run into someone that weighs less than you

B. If everyone else’s weight changed to match yours exactly, elevator capacity signs could stay the same; but if everyone’s weight changed to be double your weight, then elevator capacities would need to be cut in half

C. None of the above

If your weight is **average**, then

A. Median

VS.

B. Mean

Text-based summary (by threshold)

Centrality
What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN

The median describes RELATIVE POSITION for a SINGLE individual within an ENSEMBLE of peers

Text-based summary (by threshold)

Centrality
What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN

The median describes RELATIVE POSITION for a SINGLE individual within an ENSEMBLE of peers

We need to reorder the column of observations to compute (they must be in ascending / descending order) – standard LeetCode question!

Text-based summary (in aggregate)

Centrality
How does the sum total of all **values** compare[1]? MEAN

The mean compares CUMULATIVE VALUES for a POOLED ENSEMBLE of peers to a STANDARDIZED MEASURE (sum/#)

[1] to the number of observations

Text-based summary (in aggregate)

Centrality
How does the sum total of all **values** compare[1]? MEAN

Simple to compute, even on paper – no need to reorder the column of observations

The mean compares CUMULATIVE VALUES for a POOLED ENSEMBLE of peers to a STANDARDIZED MEASURE (sum/#)

[1] to the number of observations

MEAN as a stand-in for MEDIAN

If the histogram is symmetric,

i.e., for each value above the median,

there is a value at equal distance below the median

and vice versa

then all these differences will cancel each other out when we compute the sum total of all the values,

so the MEAN will be equal to the MEDIAN

Cautions

If the histogram is not symmetric (we call that skew)

then the MEDIAN and MEAN might be very different from each other

Cautions

If the histogram is not symmetric (we call that skew)

then the MEDIAN and MEAN might be very different from each other

Why does this matter?

MEAN is the flip-side of the MEDIAN

The mean is the POV of the house

- Q: How much profit did the house *realize* (*per gambler)?*
- A: The mean is equal to the profit per gambler

Note: This is not saying how many people profited/lost

MEAN is the flip-side of the MEDIAN

The mean is the POV of the house

Q: How much profit did the house *realize* (*per gambler)?*

A: The mean is equal to the profit per gambler

Note: This is not saying how many people profited/lost

The median is the POV of the gambler

Q: How many gamblers in a group *realized a* profit?

A: If median > 0, then more than half profited; If median < 0, then less than half did

Note: This is not saying how much the profit/loss would be per gambler

If your weight is average, then

A. You are as likely to run into someone that weighs more than you as you are to run into someone that weighs less than you

B. If everyone else’s weight changed to match yours exactly, elevator capacity signs could stay the same; but if everyone’s weight changed to be double your weight, then elevator capacities would need to be cut in half

C. Clothes fitted in your size are the most popular size option

D. All of the above

E. None of the above

Text-based summaries: three ways

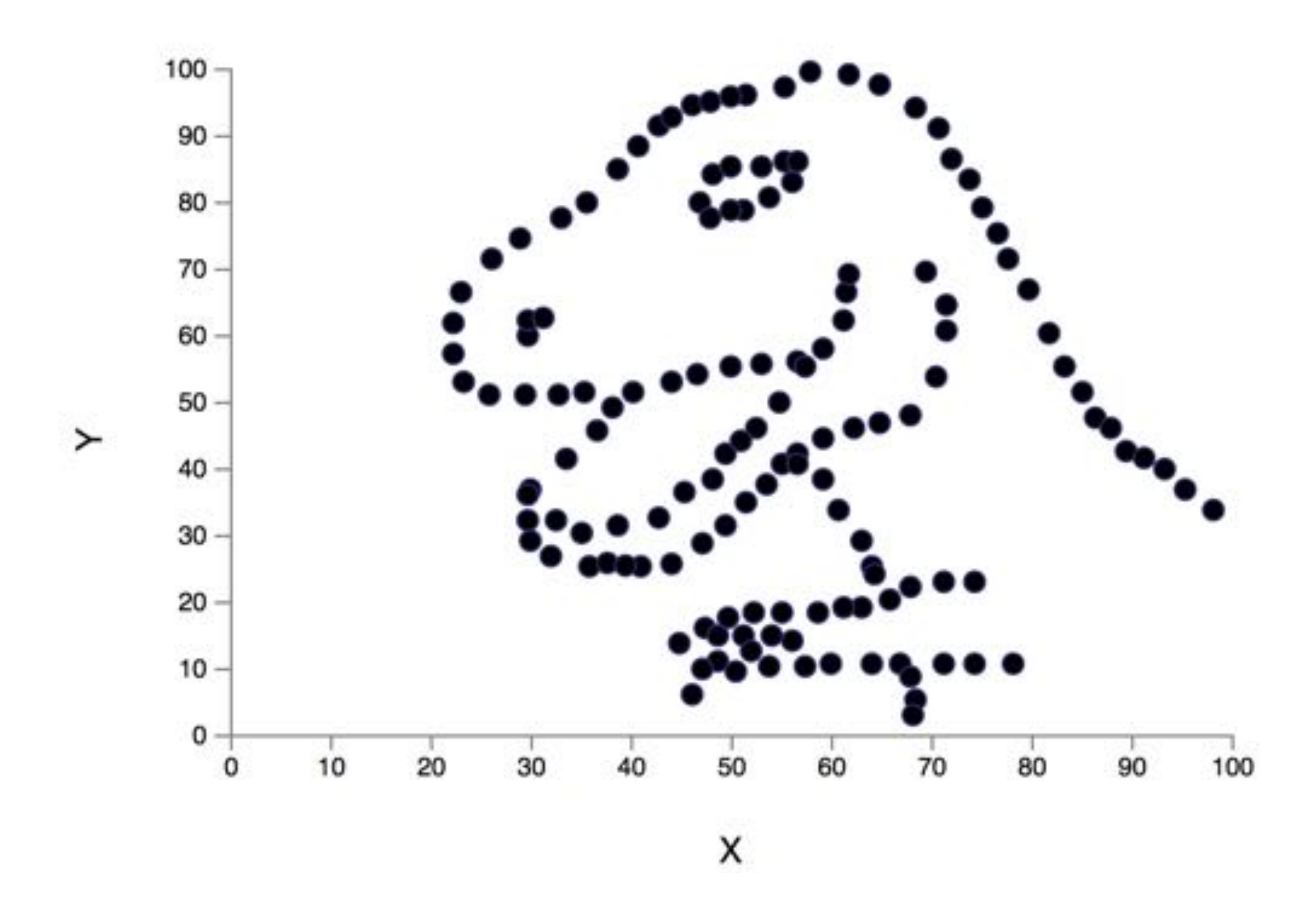
	Centrality	Dispersion
vote	What **value** is the most popular? MODE	How **many values** are very popular? Modality
threshold	What **value** splits the observations in half? (half the values are above, the other half are below) MEDIAN	What **band of values** splits the observations in half? (half the values are inside, the other half are outside) IQR
aggregate	How does the sum total of all **values** compare[1]? MEAN	How does the sum total of all **deviations**[2] compare[1]? Variance = $(\text{standard deviation})^2$

[1] to the number of observations, i.e., sum/#

[2] squared distances from the mean, i.e., $(\text{value-MEAN})^2$

Does

- ✓the statistical summary say what you *think* it says?
- ➢the statistical summary give the *full* picture?
- the statistical test ask the *right* question?
- the statistical test say what you *think* it says?



https://www.autodesk.com/research/publications/same-stats-different-graphs

The Datasaurus

STATISTICAL TESTS: meaningful differences

Congratulations! Your experiment found a difference in performance

STATISTICAL TESTS: meaningful differences

Congratulations! Your experiment found a difference in performance

But should you be measuring this difference to begin with?