SERIOUS SQL LIVE WEEK 3: 4TH DEC

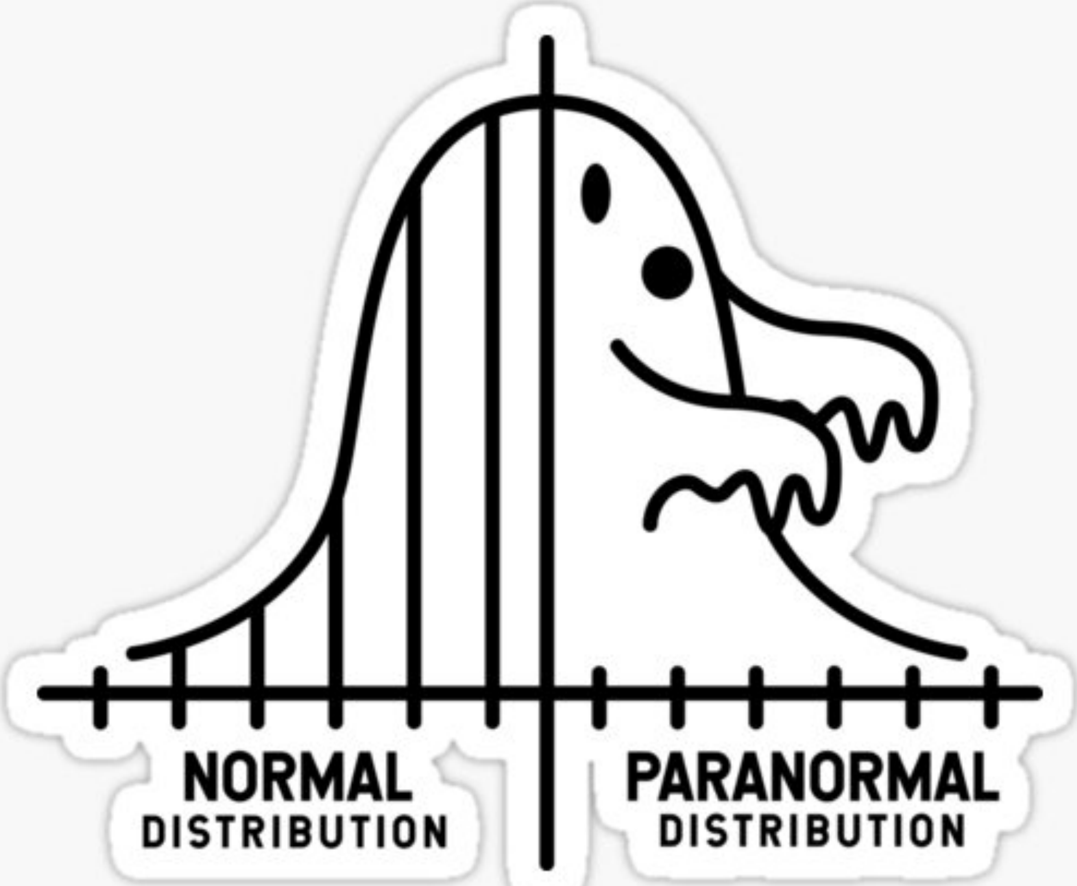
BY DANNY MA

DATA WITH DANNY

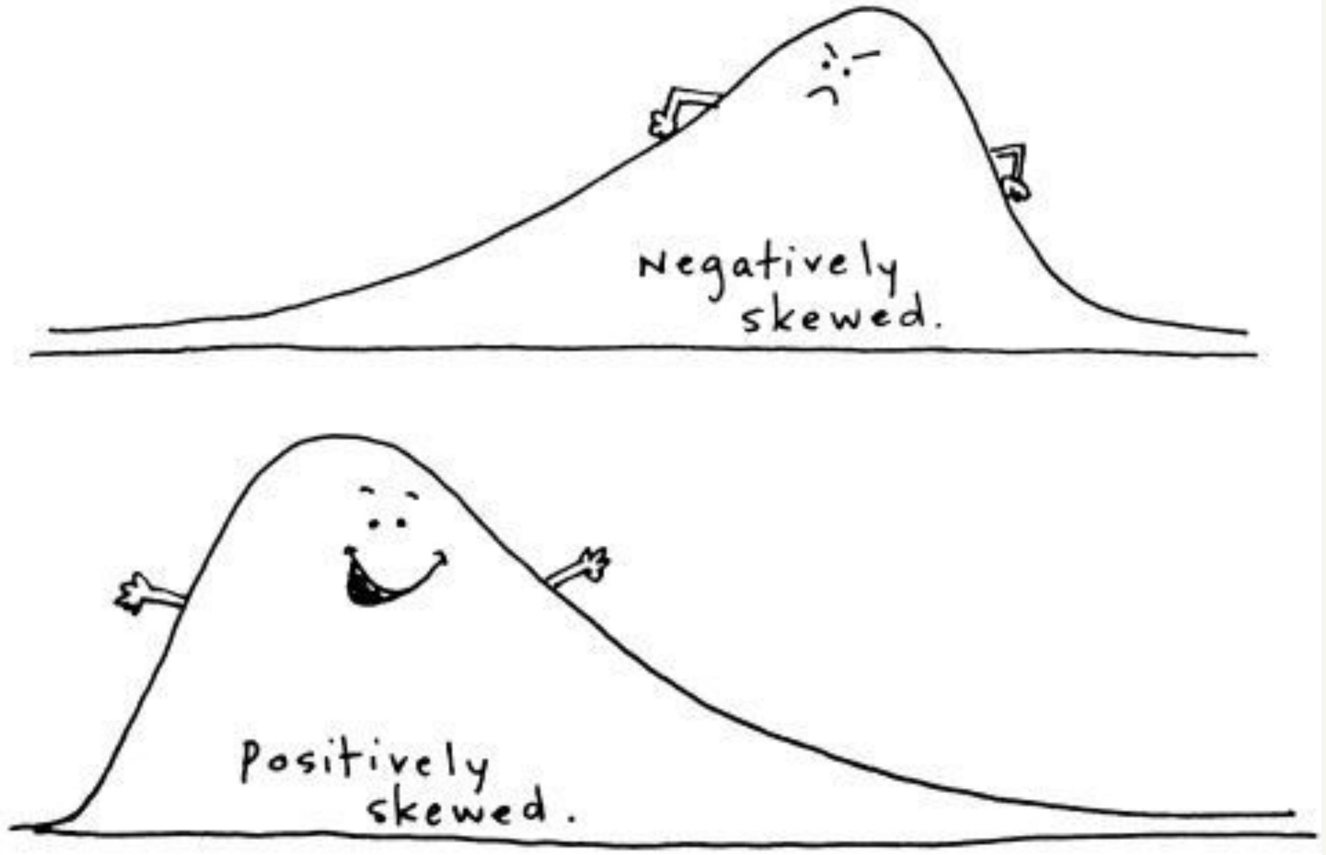
AGENDA:

- Intro (5 mins)
- Summary Statistics (55 mins)

SUMMARY STATISTICS



"Data don't make any sense,
we will have to resort to statistics."



STATISTICS 101: CENTRAL TENDANCY

- Mean/Average (arithmetic mean) ✓
- Median (50th percentile)
- Mode (most frequent value)

MEAN / AVERAGE

$\{X_1, X_2, X_3 \cdots X_N)$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

$\rightarrow x_1 + x_2 + x_3 + \cdots + x_N$

mu

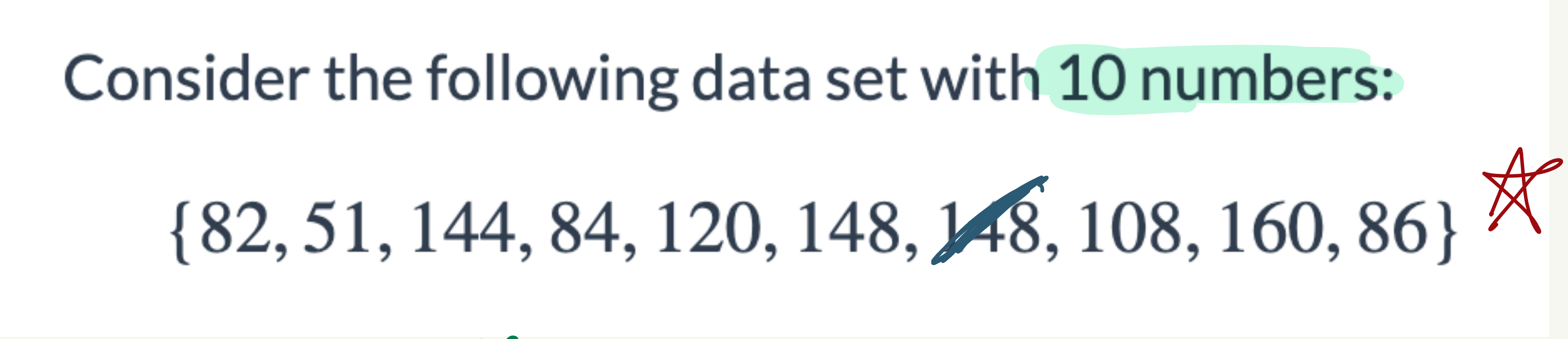
average

APPLIED STATS

blood glucose
pressure
weight

- What is the average measure_value?
- Does this look right?
- Let's look at the average "inputs"
- What about for each measure?

AVG, MEDIAN & MODE



Consider the following data set with 10 numbers:

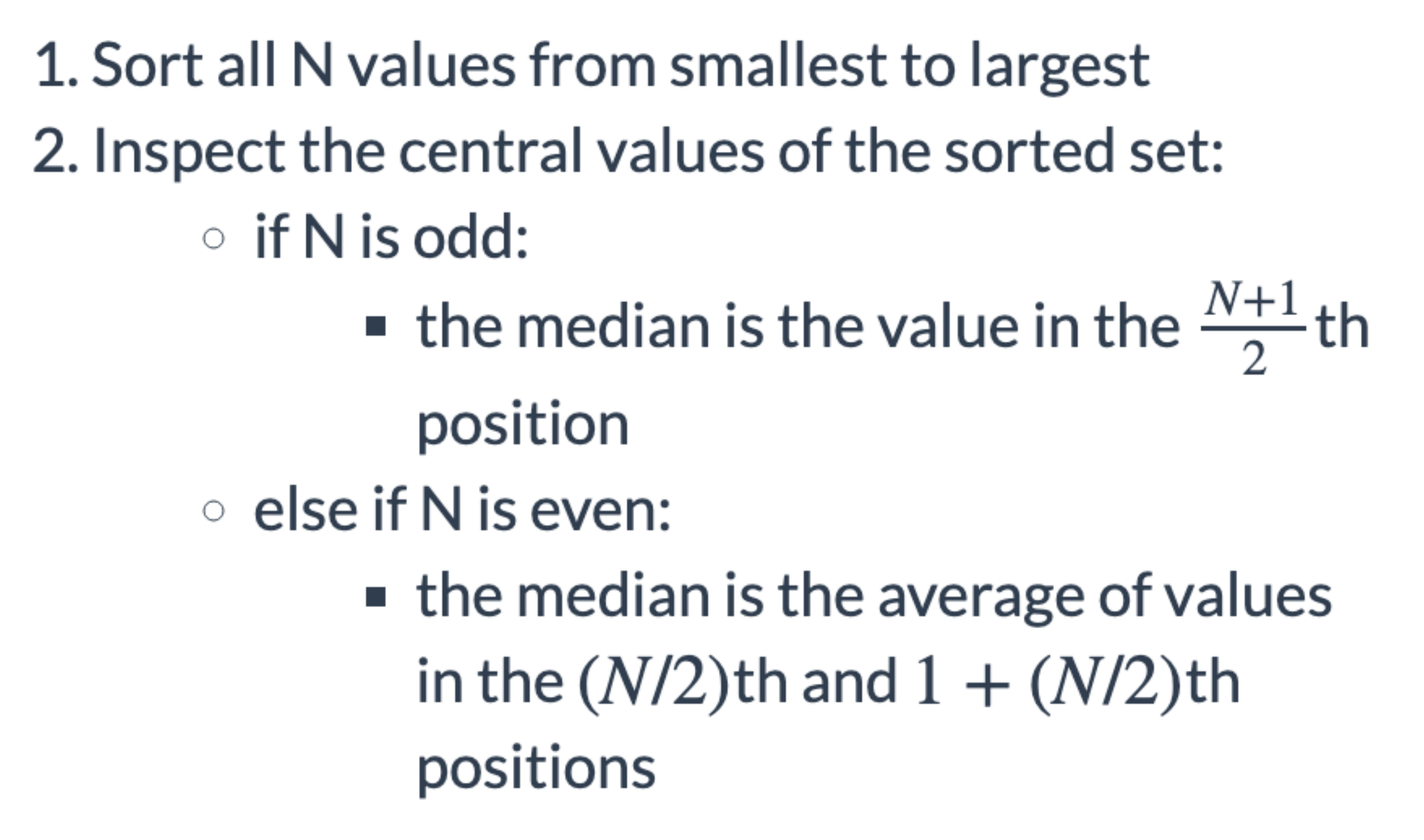
{82, 51, 144, 84, 120, 148, ~~148~~, 108, 160, 86} ☆

$$\text{AVG} = \frac{82+51+\cdots+86}{10} = 113.1$$

$$51, 82, 84, 86, \boxed{108, 120}, 144, \boxed{148, 148}, 160$$

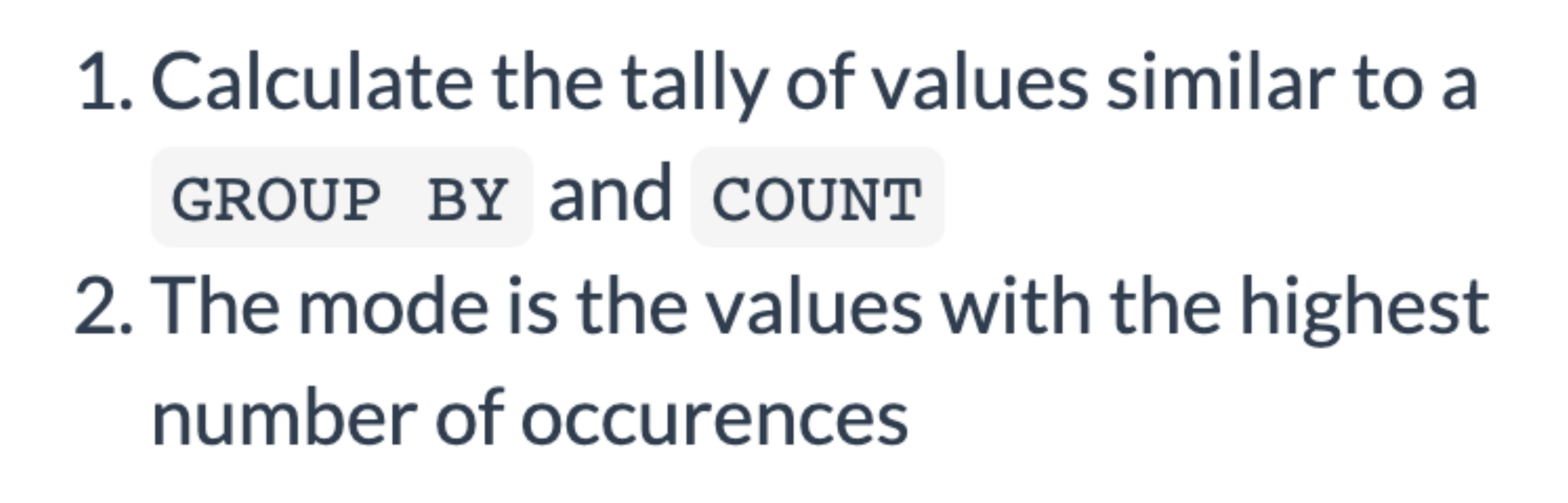
$$\frac{108+120}{2} = 114$$

MEDIAN ALGORITHM



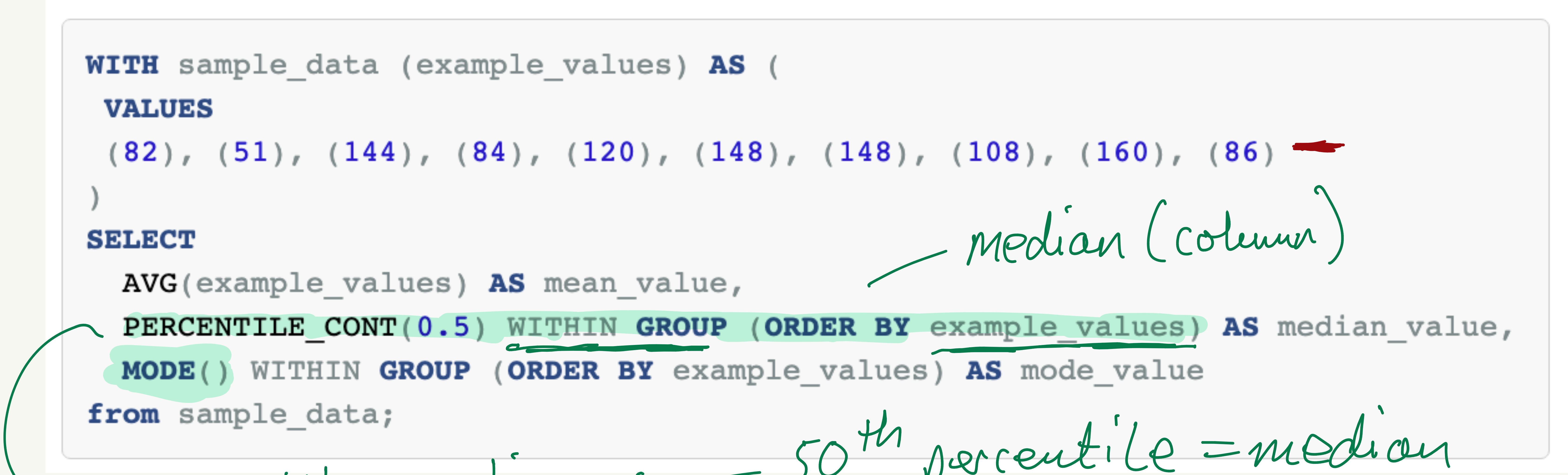
1. Sort all N values from smallest to largest
2. Inspect the central values of the sorted set:
    - if N is odd:
        - the median is the value in the $\frac{N+1}{2}$th position
    - else if N is even:
        - the median is the average of values in the $(N/2)$th and $1 + (N/2)$th positions

MODE ALGORITHM



1. Calculate the tally of values similar to a `GROUP BY` and `COUNT`
2. The mode is the values with the highest number of occurences

SQL IMPLEMENTATION



```
WITH sample_data (example_values) AS (
 VALUES
 (82), (51), (144), (84), (120), (148), (148), (108), (160), (86)
)
SELECT
  AVG(example_values) AS mean_value,
  PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY example_values) AS median_value,
  MODE() WITHIN GROUP (ORDER BY example_values) AS mode_value
from sample_data;
```

median (column)

percentile_continuous — 50th percentile = median
middle value

SPREAD STATISTICS

- MIN, MAX, range (MIN - MAX)
- Variance & Standard Deviation

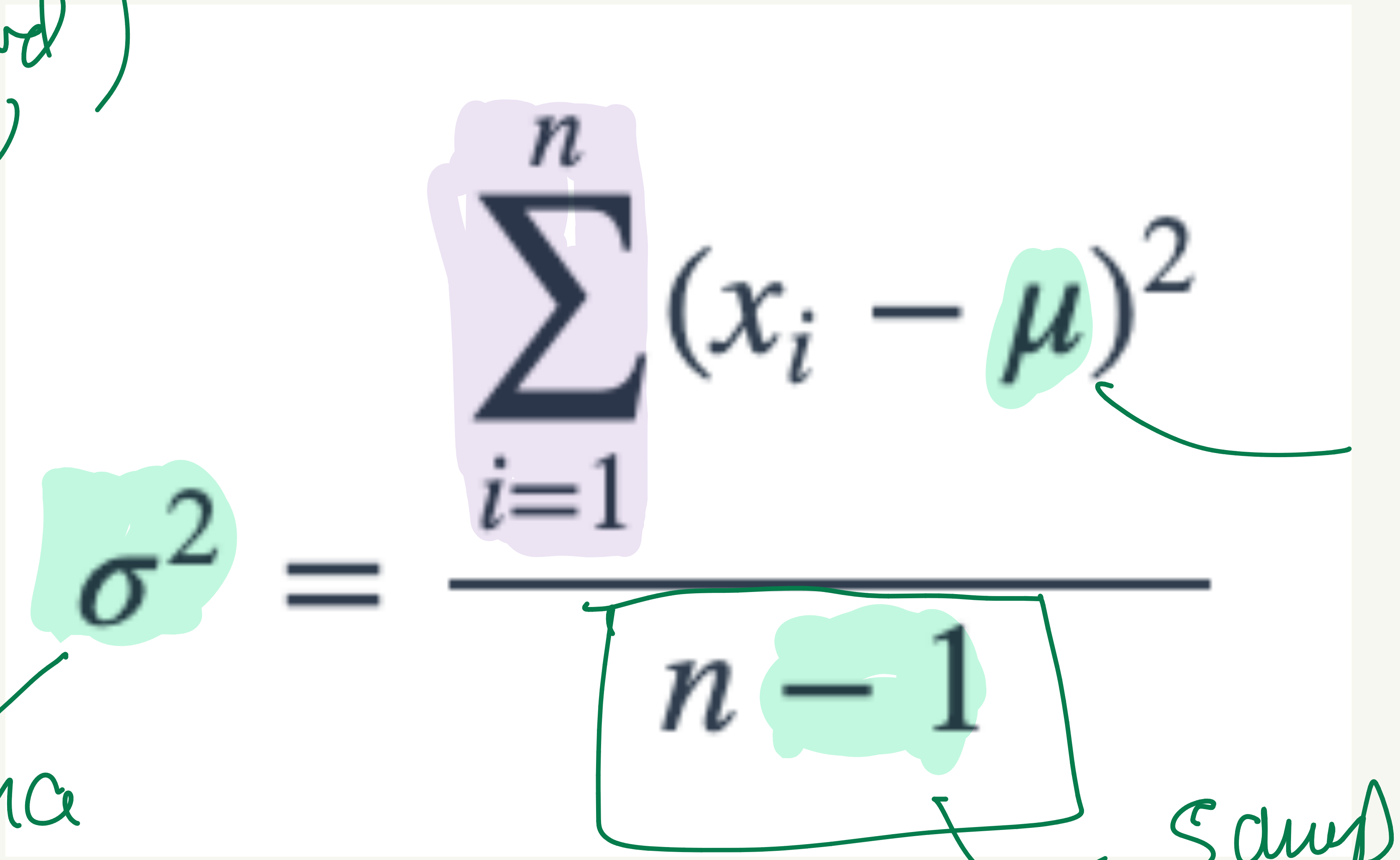
MIN AND MAX WEIGHTS

- What is the max and min weights?
- What is the range?
- Do you think this is "normal"?

VARIANCE & STDDEV

variance = $\left(\text{standard dev}\right)^2$

Standard Deviation



$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\boxed{n-1}}$$

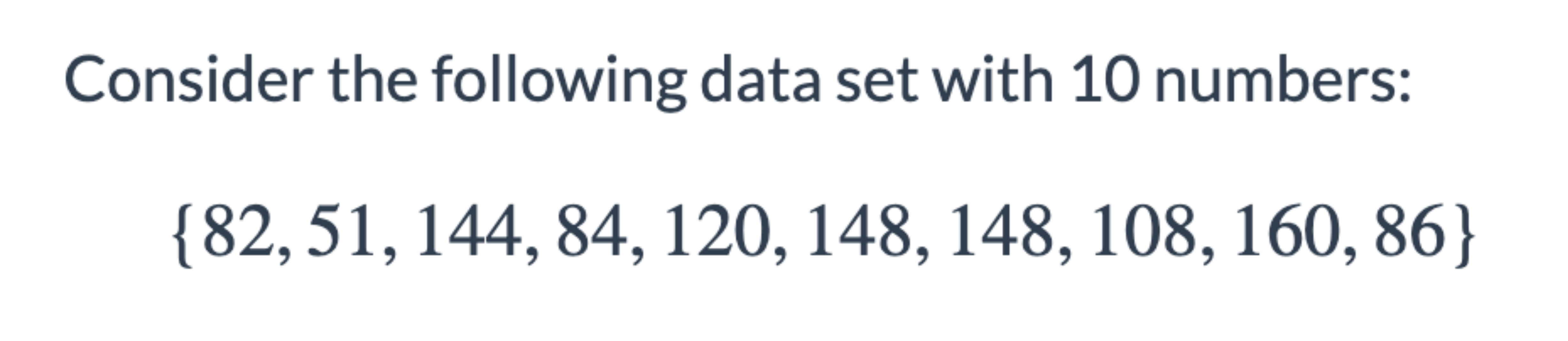
mean

sigma

Standard deviation

sample variance

VARIANCE ALGORITHM



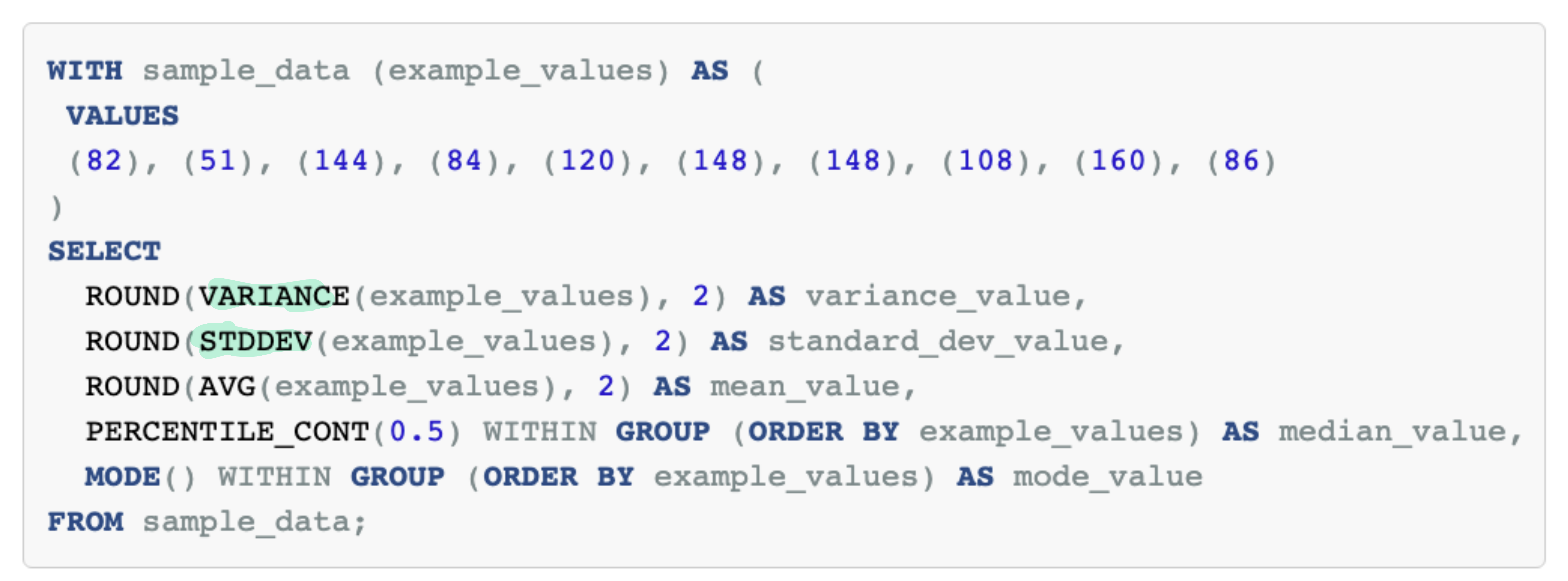
Consider the following data set with 10 numbers:

$$\{82, 51, 144, 84, 120, 148, 148, 108, 160, 86\}$$

$\mu = 113 \cdot 1$

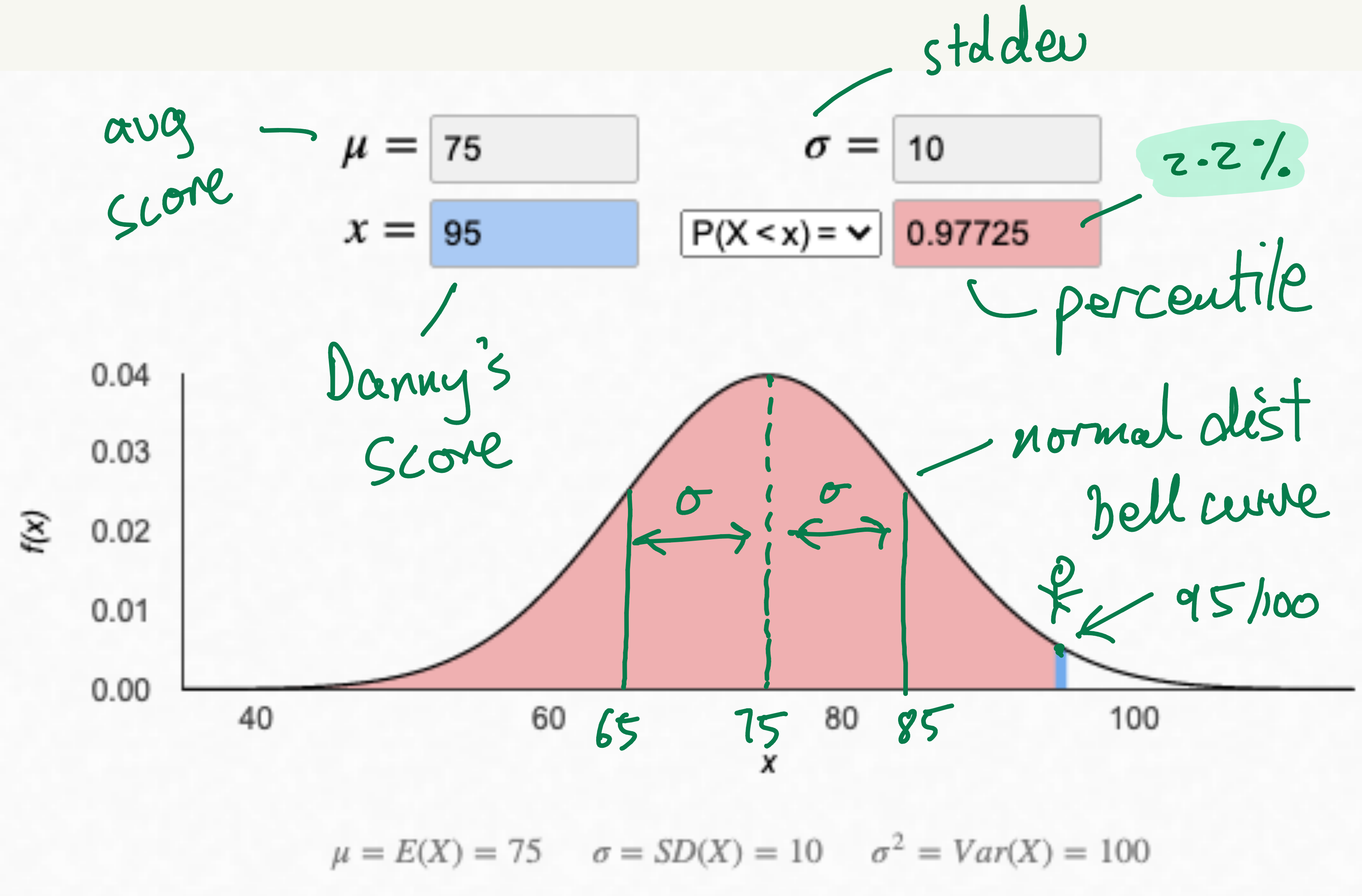
$$\sigma^2 = \frac{(82-113\cdot1)^2 + (51-113\cdot1)^2 + \cdots + (86-113\cdot1)^2}{10-1}$$

ALL THE STATISTICS



```
WITH sample_data (example_values) AS (
 VALUES
 (82), (51), (144), (84), (120), (148), (148), (108), (160), (86)
)
SELECT
  ROUND(VARIANCE(example_values), 2) AS variance_value,
  ROUND(STDDEV(example_values), 2) AS standard_dev_value,
  ROUND(AVG(example_values), 2) AS mean_value,
  PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY example_values) AS median_value,
  MODE() WITHIN GROUP (ORDER BY example_values) AS mode_value
FROM sample_data;
```

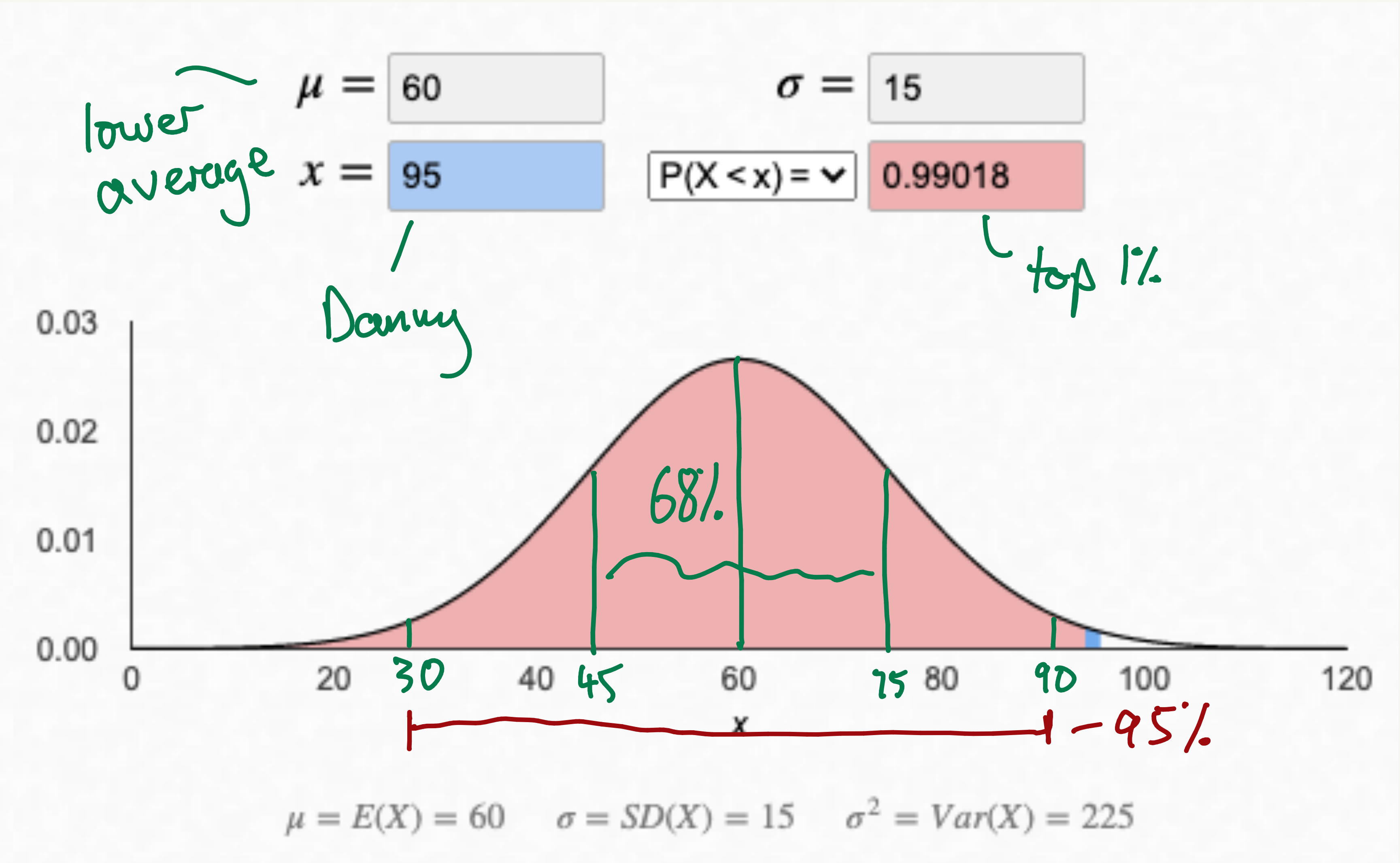
SPREAD & PERCENTILES



$\mu =$ 75 $\sigma =$ 10

$x =$ 95 P(X < x) = 0.97725

$\mu = E(X) = 75$ $\sigma = SD(X) = 10$ $\sigma^2 = Var(X) = 100$

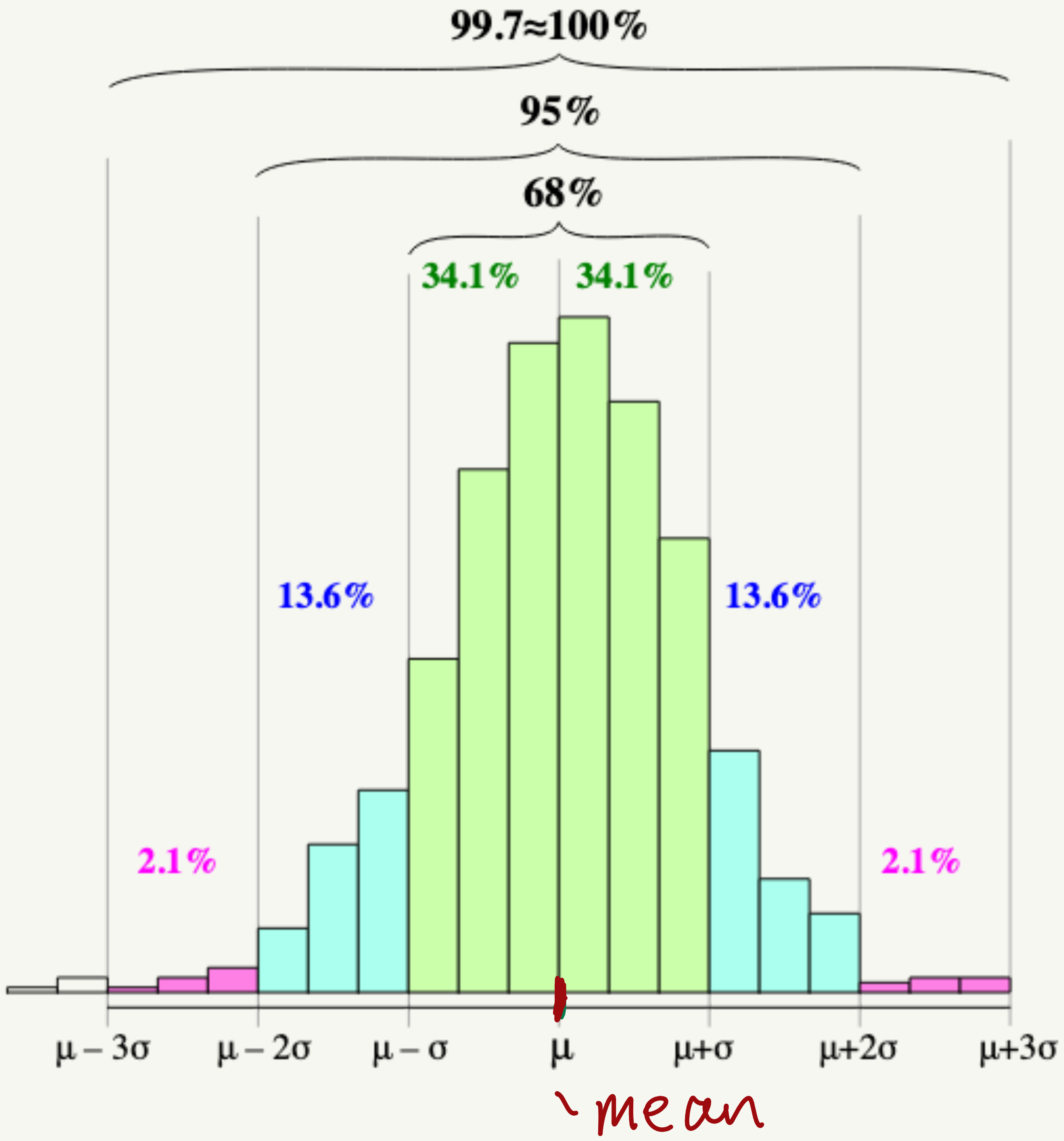


$\mu =$ 60 $\sigma =$ 15

$x =$ 95 P(X < x) = 0.99018

$\mu = E(X) = 60$ $\sigma = SD(X) = 15$ $\sigma^2 = Var(X) = 225$

THE EMPIRICAL RULE



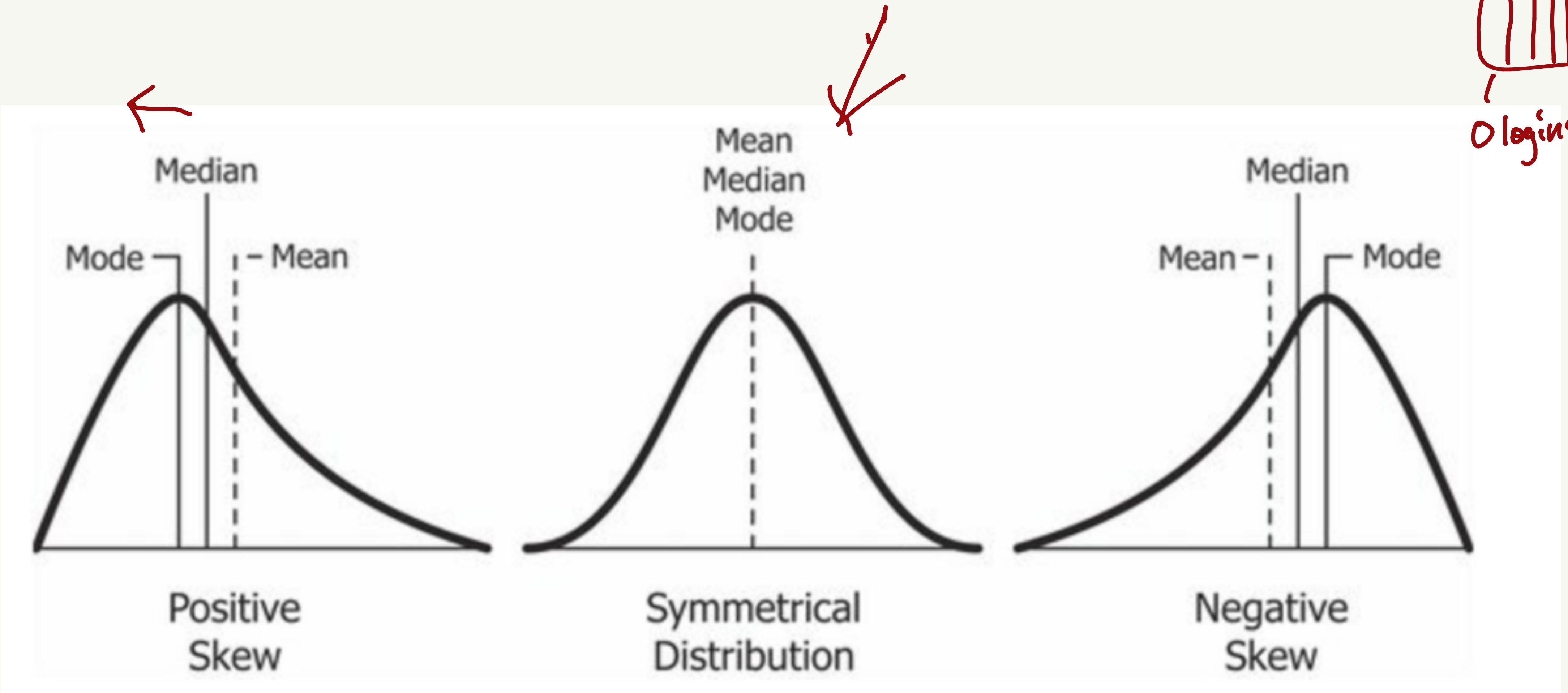
68% lies between

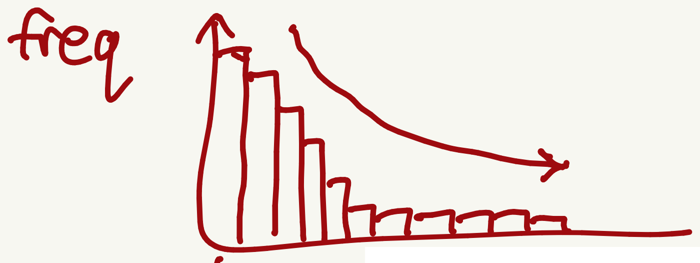
$\mu \pm 1 \times SD$

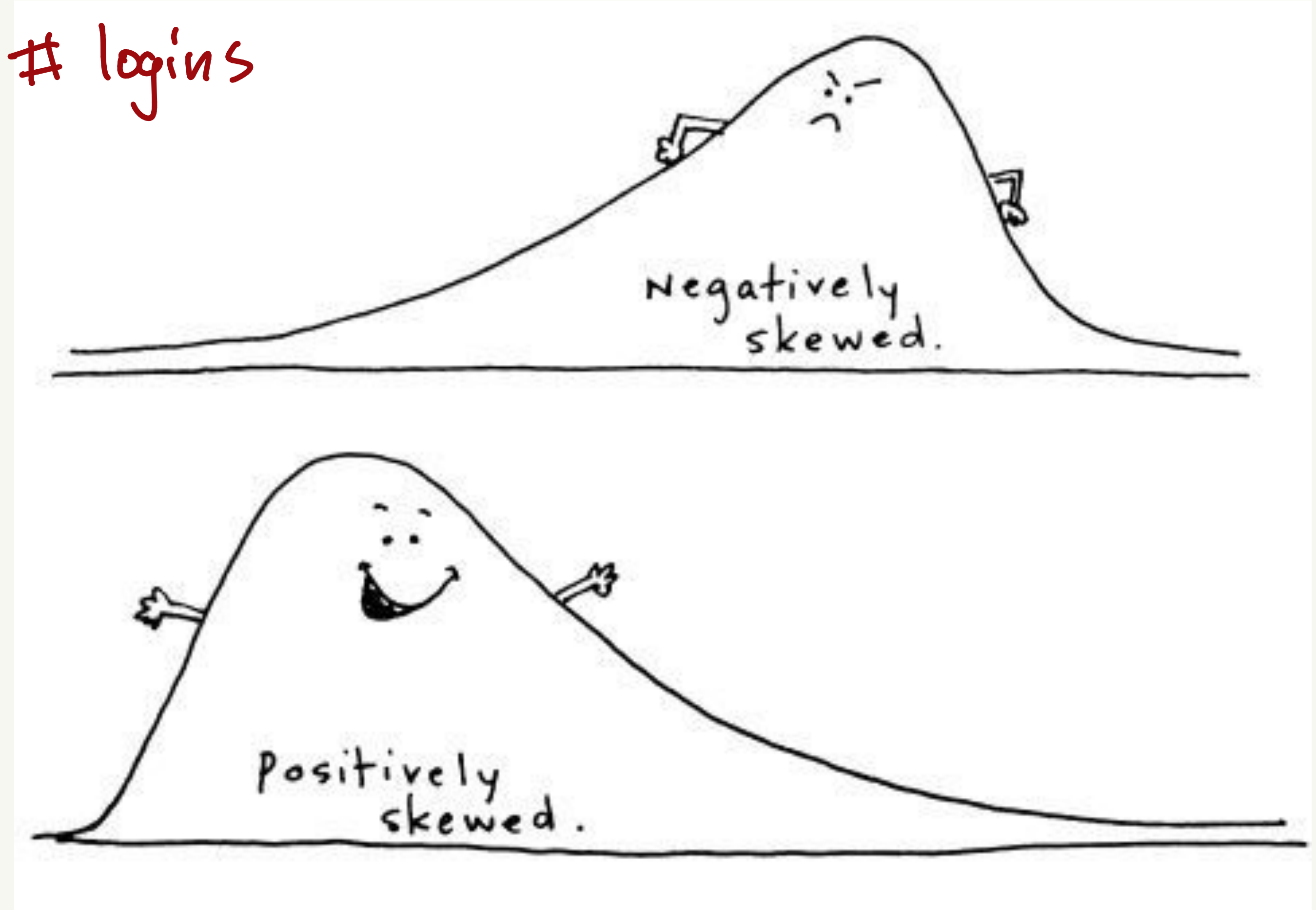
95% Lie between

$\mu \pm 2 \times SD$

REAL DISTRIBUTIONS







REAL WEIGHT STATISTCS

- Average weight
- 50th percentile median weight
- Most frequent mode weight
- Min, max and range of weights
- Variance and standard deviation

SUMMARY STATISTICS

- Central stats: mean, median, mode
- Spread stats: min, max, range
- Variance and Standard Deviation
- Percentiles
- Confidence intervals
- Skewed Distributions