GENERAL ASSEMBLY

STATS 101: FUNDAMENTALS (+HOW TO LIE WITH STATISTICS)

Joseph Nelson, Data Science Immersive

AGENDA

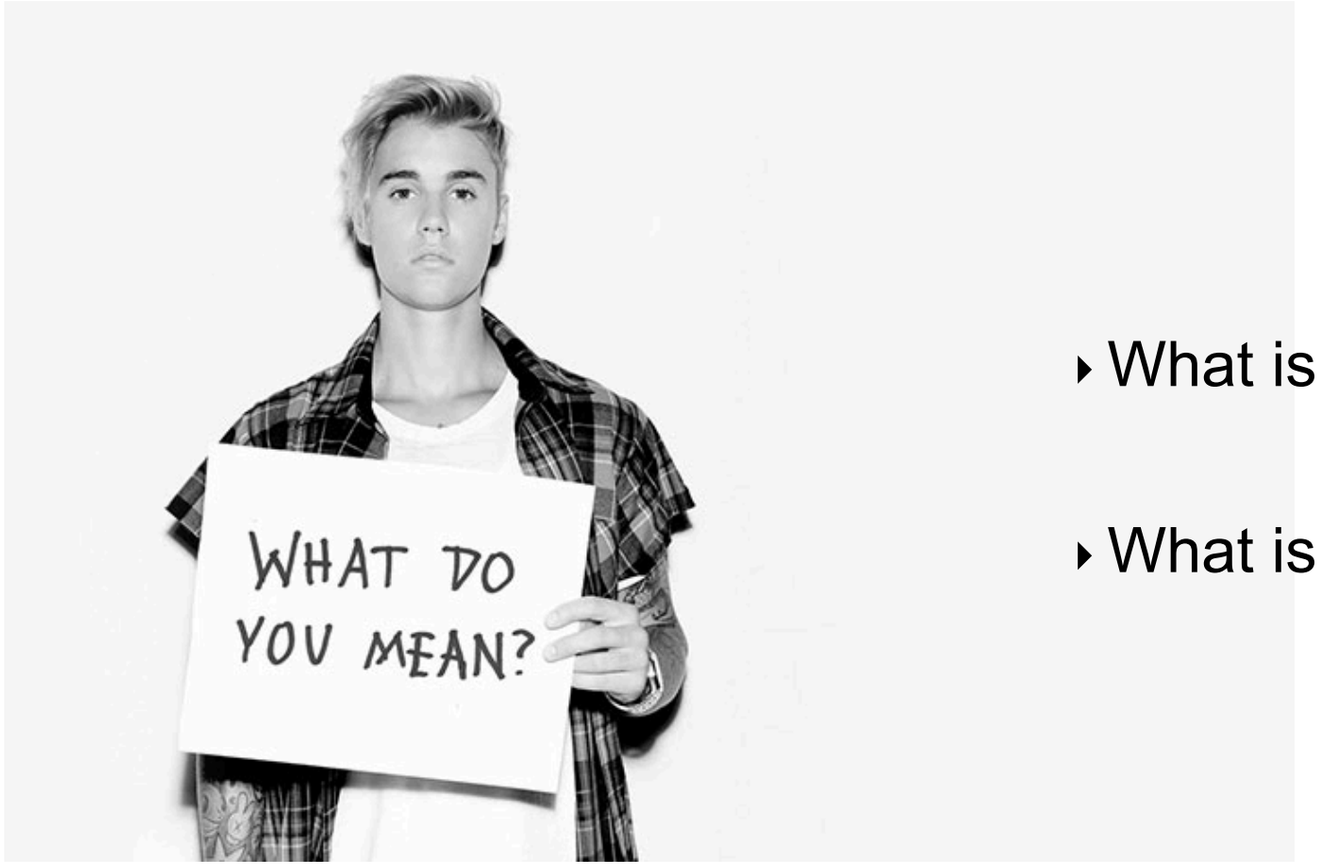
- Basic Descriptive Statistics
- Quick Box Plots (Snuck in here)
- Bias vs Variance: A Preface
- Standard Deviation and Variance
- Normality
- Central Limit Theorem
- Dummy Variables

- We’ll see coding implementations along the way!

BASIC DESCRIPTIVE STATISTICS

- Mean
- Median
- Mode
- Max
- Min
- Quartile
- Inter-quartile Range
- Variance
- Standard Deviation
- Correlation

MEAN



- What is the mean?
- What is another name for the mean?

MEAN

- What is the mean?
- The mean of a set of values is the sum of the values divided by the number of values. It is also called the average.
- It is also known as the average.
- Example: Find the mean of 19, 13, 15, 25, and 18

MEDIAN

- What is the median?
- How do you find the median?

MEDIAN

- What is the median?
- How do you find the median?
- Bonus: Why might the median be advantageous instead of the mean? When does this condition NOT hold?

MEDIAN

- The median refers to the midpoint in a series of numbers.

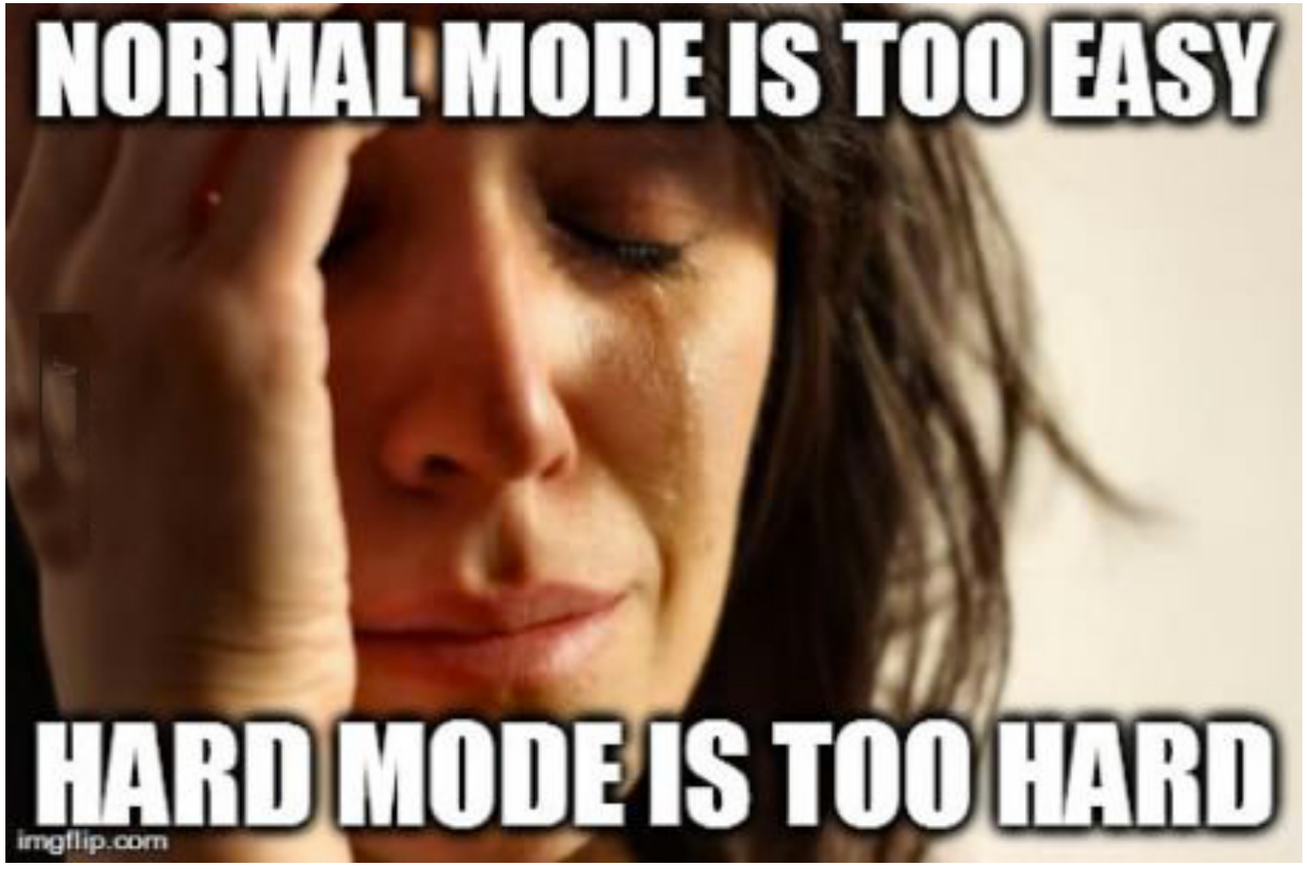
- To find the median, arrange the numbers in order from smallest to largest. If there is an odd number of values, the middle value is the median. If there is an even number of values, the average of the two middle values is the median.

MEDIAN

- The median refers to the midpoint in a series of numbers.
- Example #1: Find the median of 19, 29, 36, 15, and 20
- Example #2: Find the median of 67, 28, 92, 37, 81, 75
- Bonus: Median may be more useful than average in a highly skewed population.

MODE

- What is the mode?
- What is the mode in the following: 1, 2, 3, 4, 5



MODE

- What is the mode?
- The mode of a set of values is the value that occurs most often.
- A set of values may have more than one mode or no mode.

CHECK FOR UNDERSTANDING

- For the following groups of numbers, calculate the mean, median and mode by hand:

- A. 18, 24, 17, 21, 24, 16, 29, 18

- B. 75, 87, 49, 68, 75, 84, 98, 92

- C. 55, 47, 38, 66, 56, 64, 44, 39

CHECK FOR UNDERSTANDING

- Answers:

- A. Mean = 20.875 Median = 19.5 Mode = 18, 24 Max = 29 Min = 16

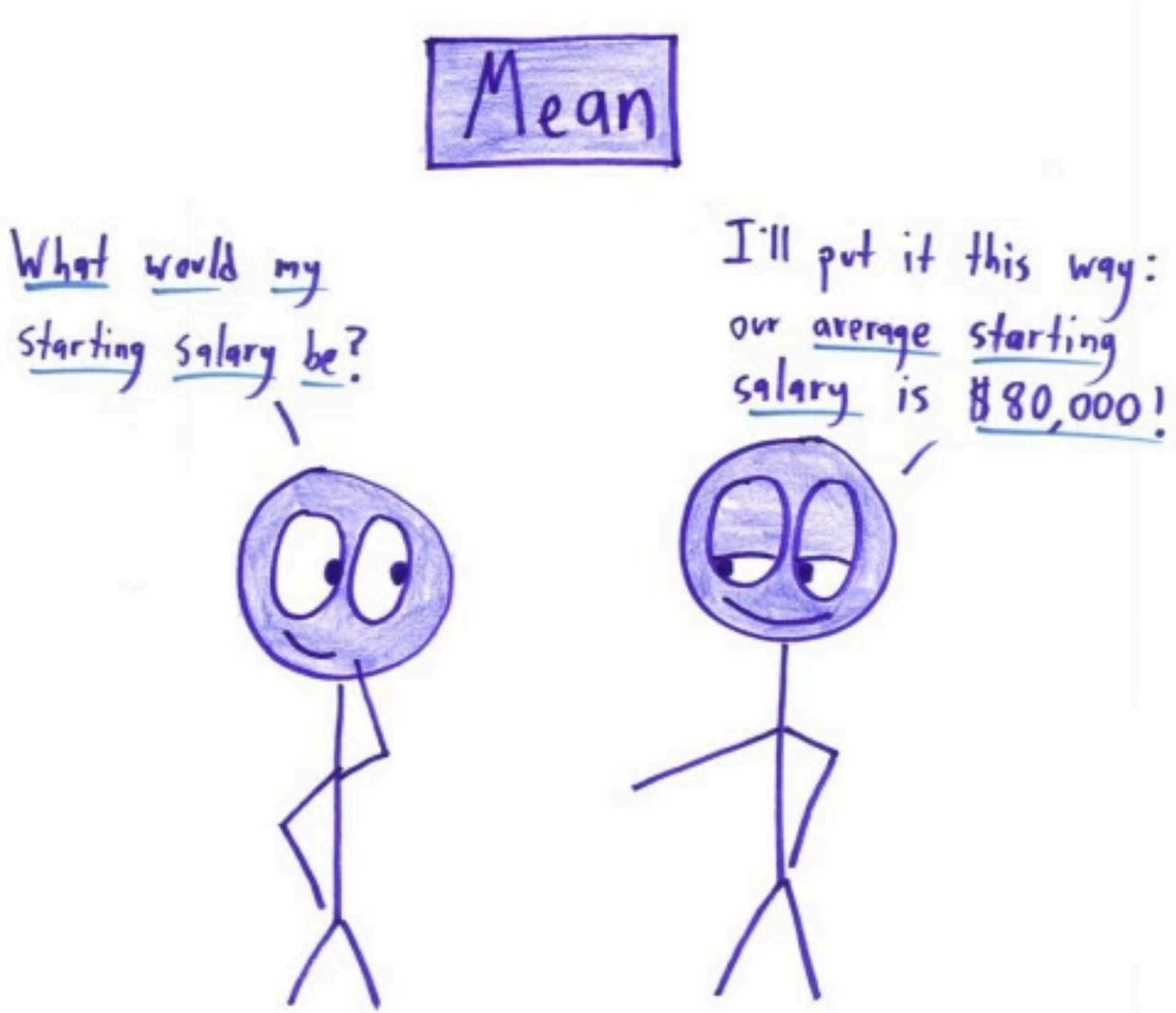
- B. Mean = 78.5 Median = 79.5 Mode = 75 Max = 98 Min = 49

- C. Mean = 51.125 Median = 51 Mode = none Max = 66 Min = 38

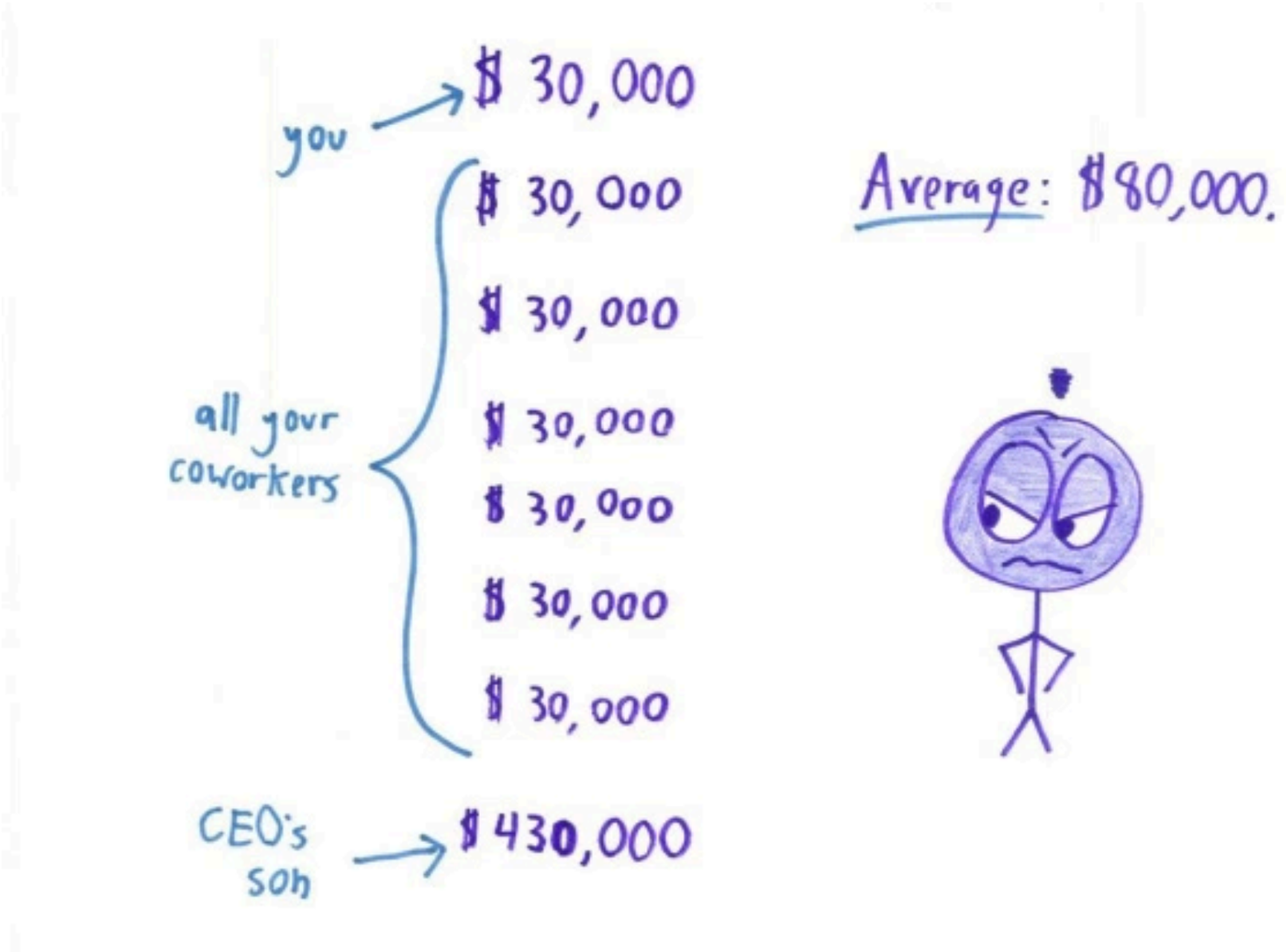
HOW TO LIE WITH STATISTICS

- For each picture:
- 1) What could go wrong
- 2) How to fix it

HOW TO LIE WITH STATISTICS

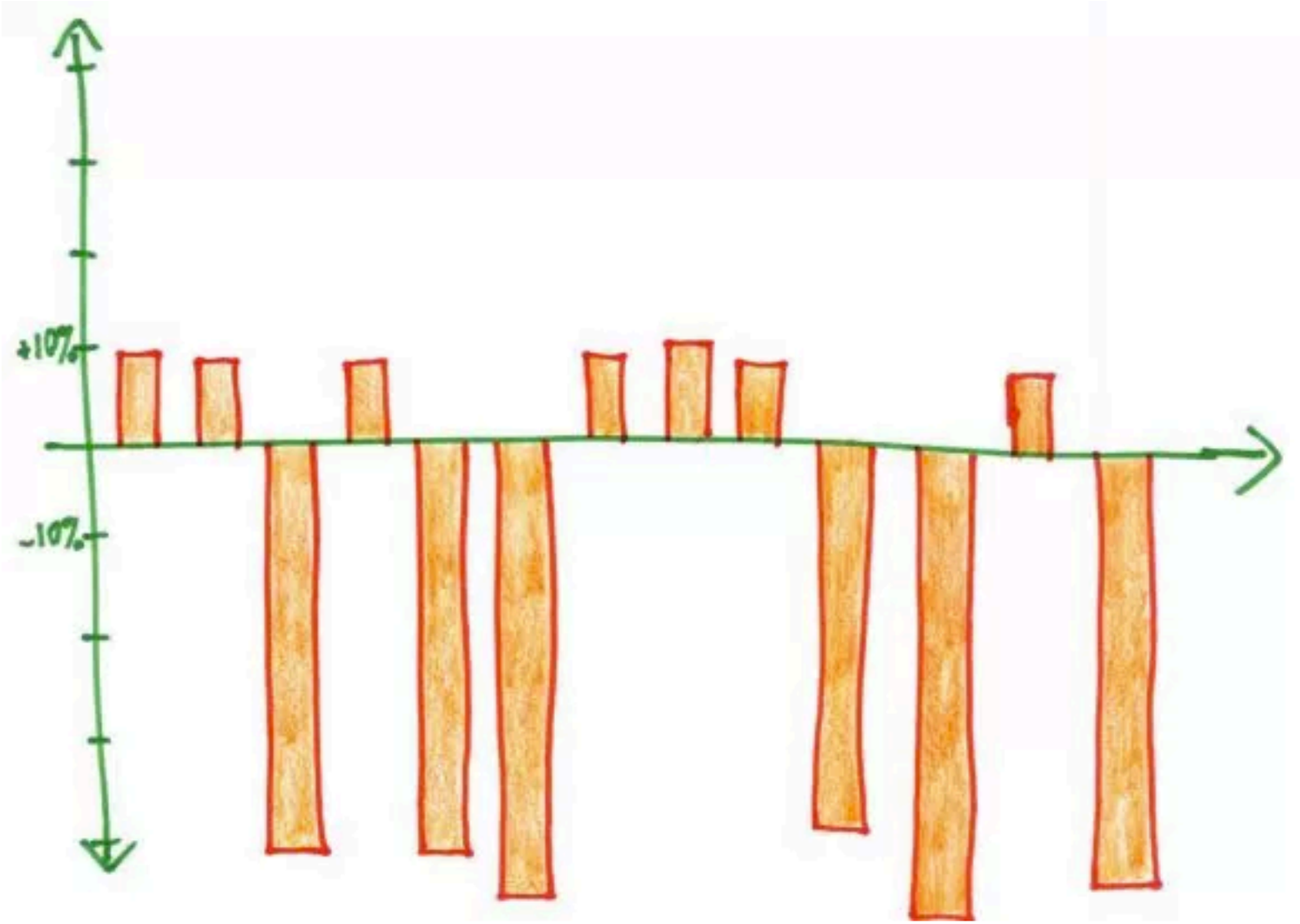


HOW TO LIE WITH STATISTICS

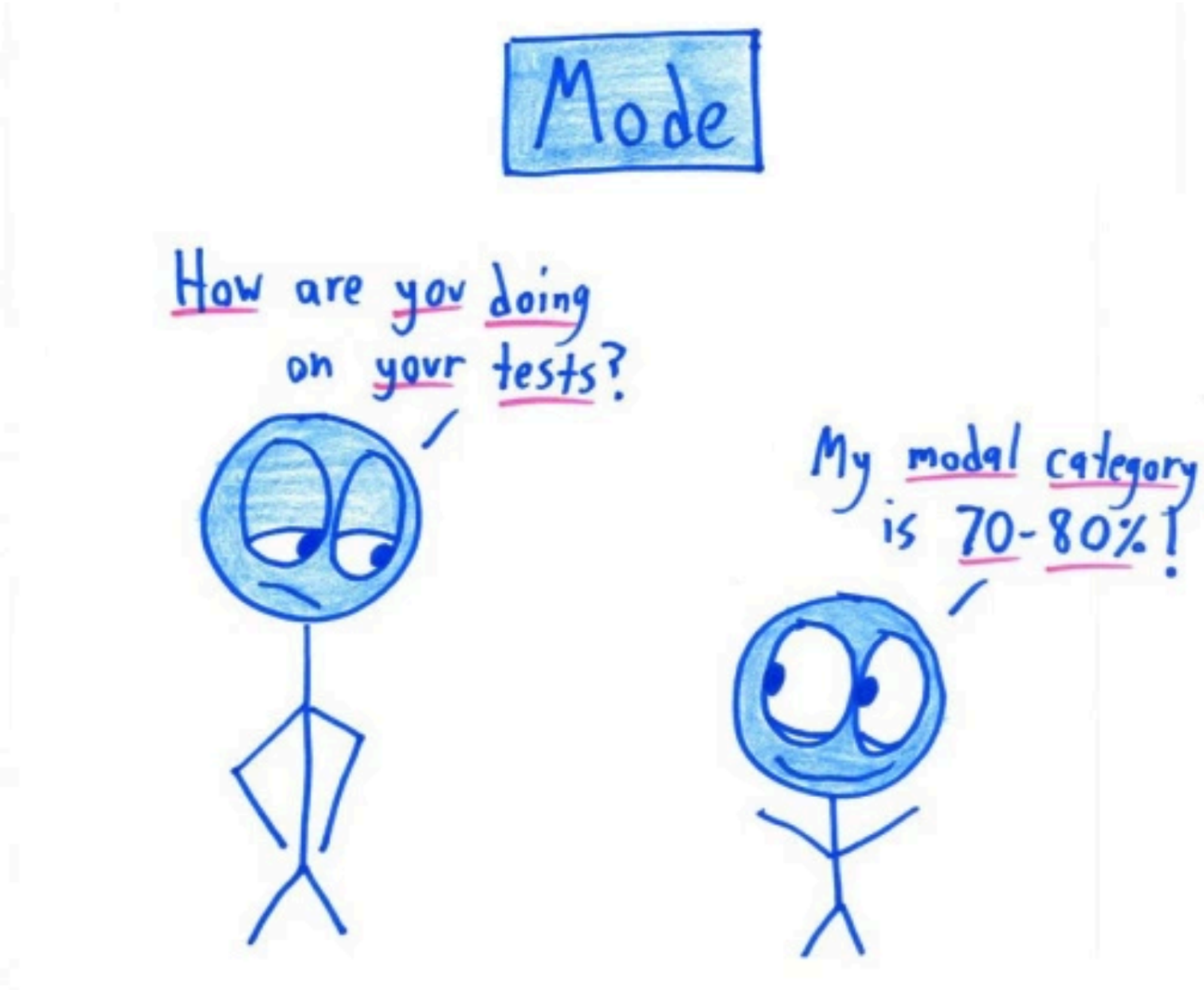


HOW TO LIE WITH STATISTICS

HOW TO LIE WITH STATISTICS



HOW TO LIE WITH STATISTICS

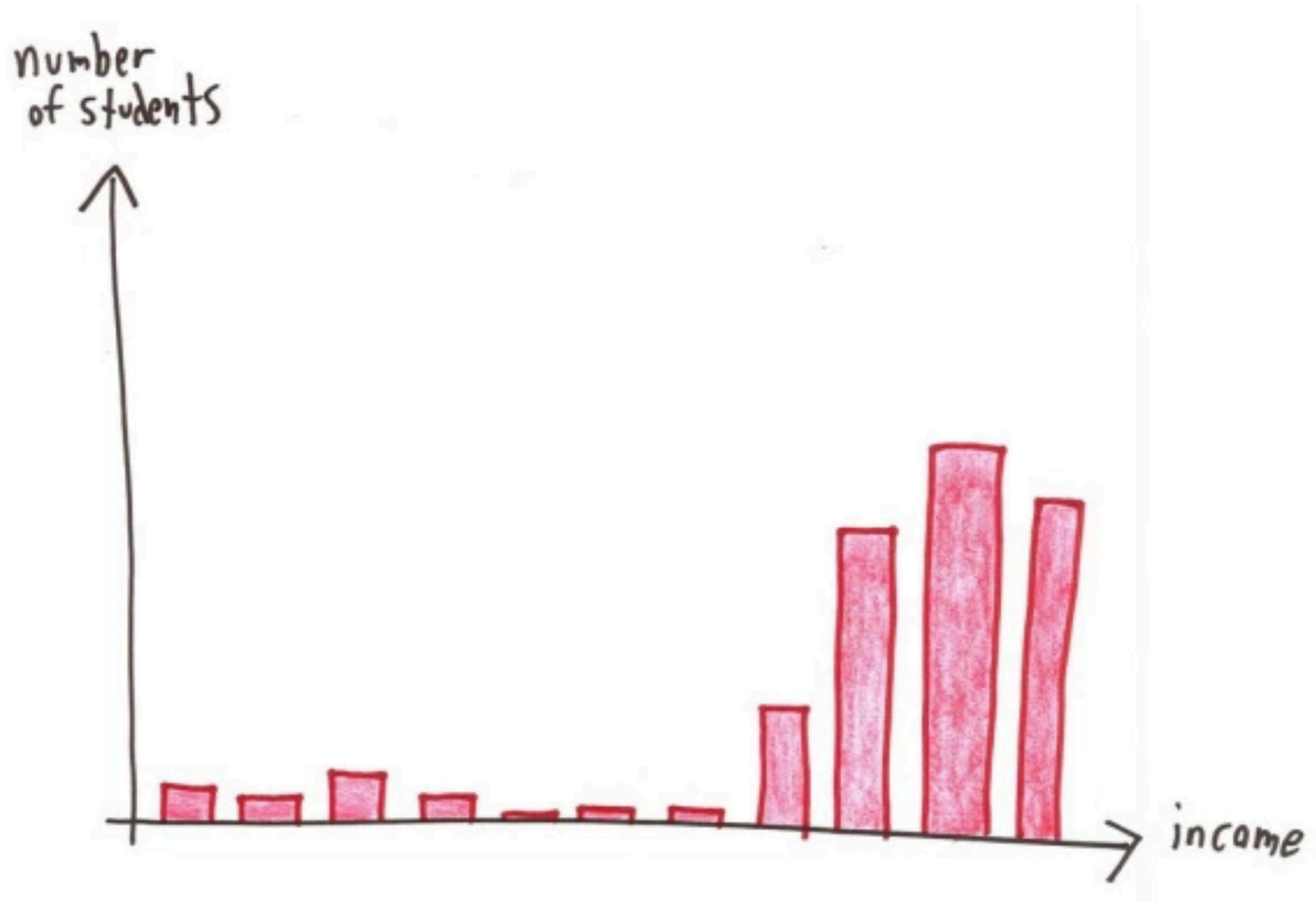


HOW TO LIE WITH STATISTICS

Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

HOW TO LIE WITH STATISTICS

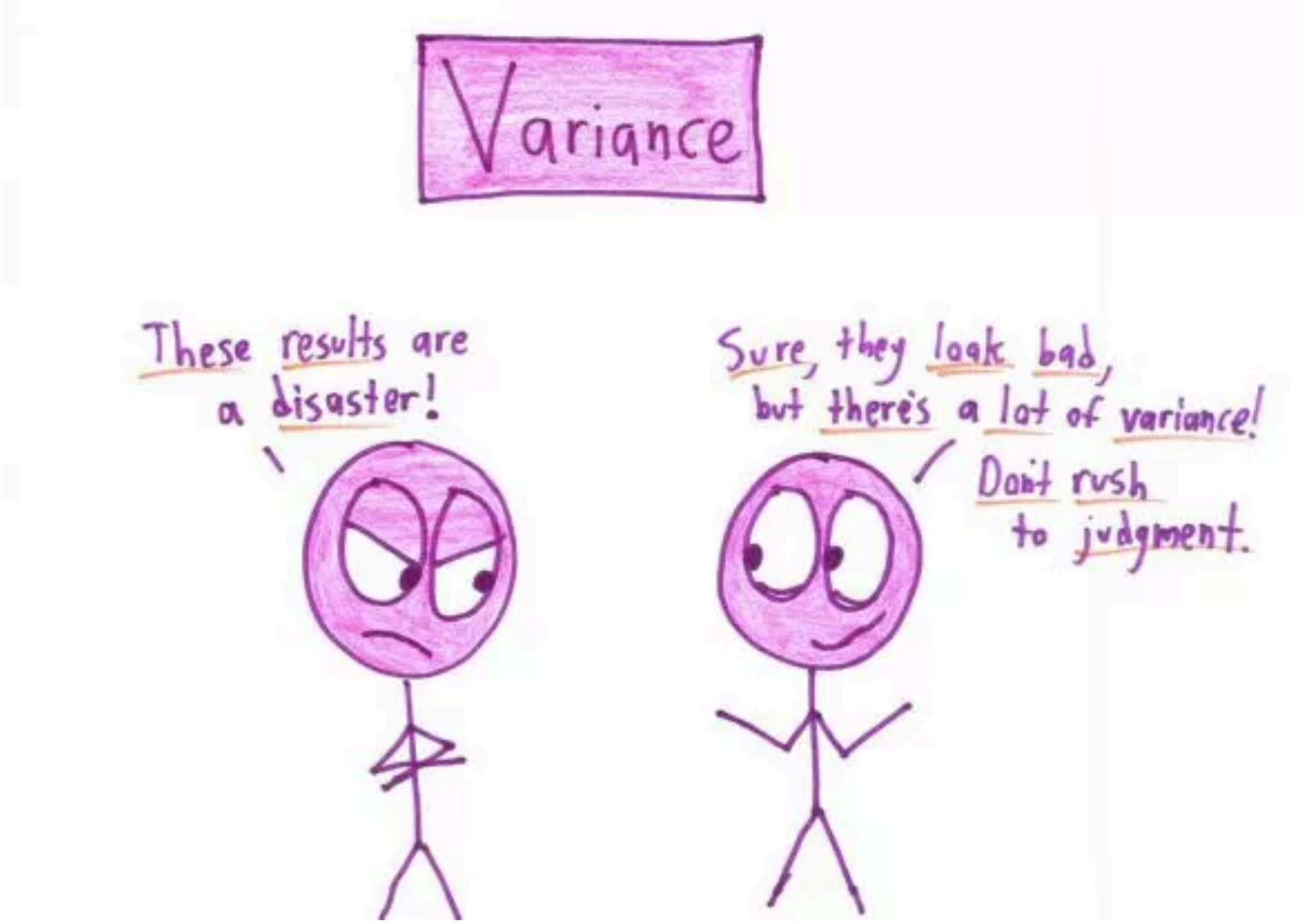
HOW TO LIE WITH STATISTICS



HOW TO LIE WITH STATISTICS

HOW TO LIE WITH STATISTICS

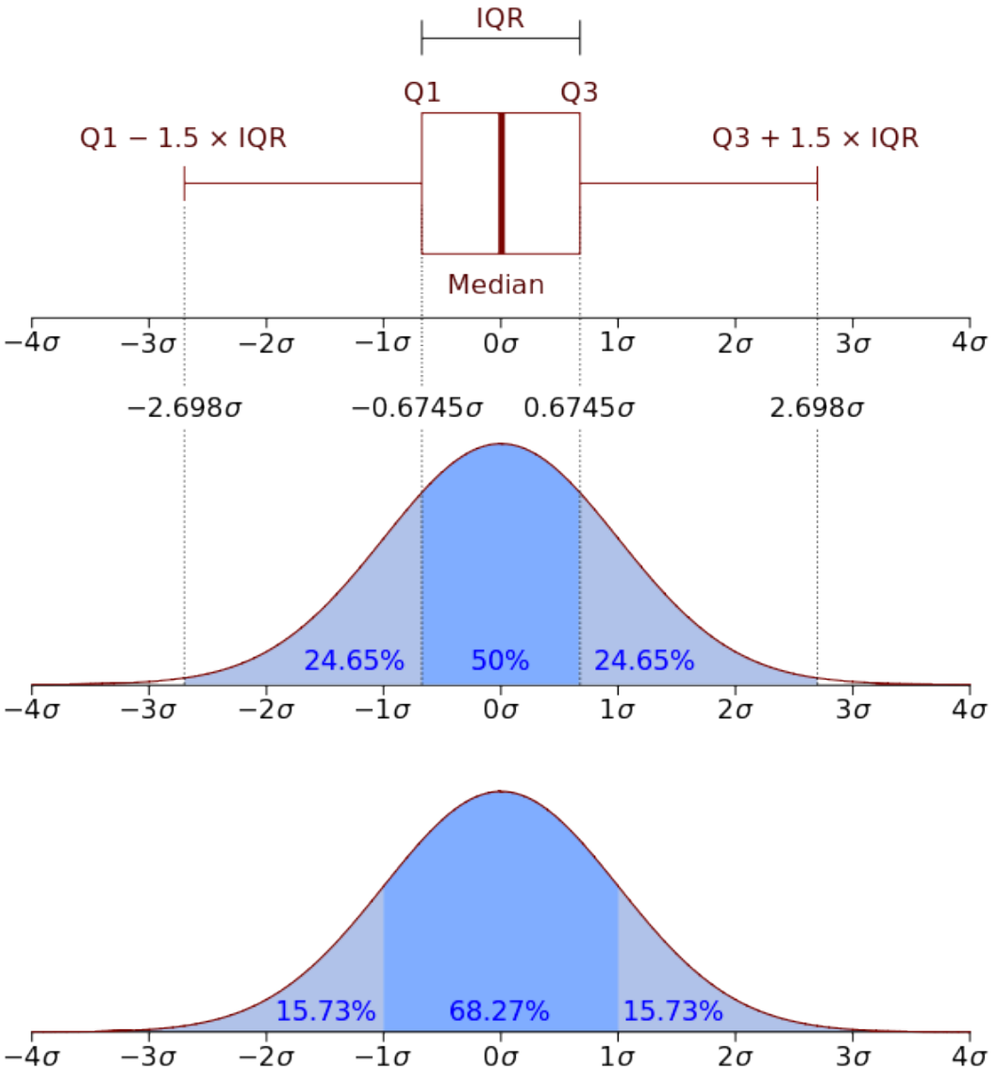
HOW TO LIE WITH STATISTICS



HOW TO LIE WITH STATISTICS

QUARTILES AND THE INTER QUARTILE RANGE

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.
- The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. It is the “middle 50” of your data. Also called the H-spread. IQR = Q3-Q1
- Outliers: Q1 - 1.5(IQR), Q3 + 1.5(IQR)



BIAS VS VARIANCE

- **Error due to Bias:** Error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Imagine you could repeat the whole model building process more than once: each time you gather new data and run a new analysis, thereby creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions. Bias measures how far off in general these models' predictions are from the correct value.

BIAS VS VARIANCE

- **Error due to Bias:** Error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Imagine you could repeat the whole model building process more than once: each time you gather new data and run a new analysis, thereby creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions. Bias measures how far off in general these models' predictions are from the correct value.
- **Error due to Variance**: The error due to variance is taken as the variability of a model prediction for a given data point. Again, imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.

BIAS VS VARIANCE

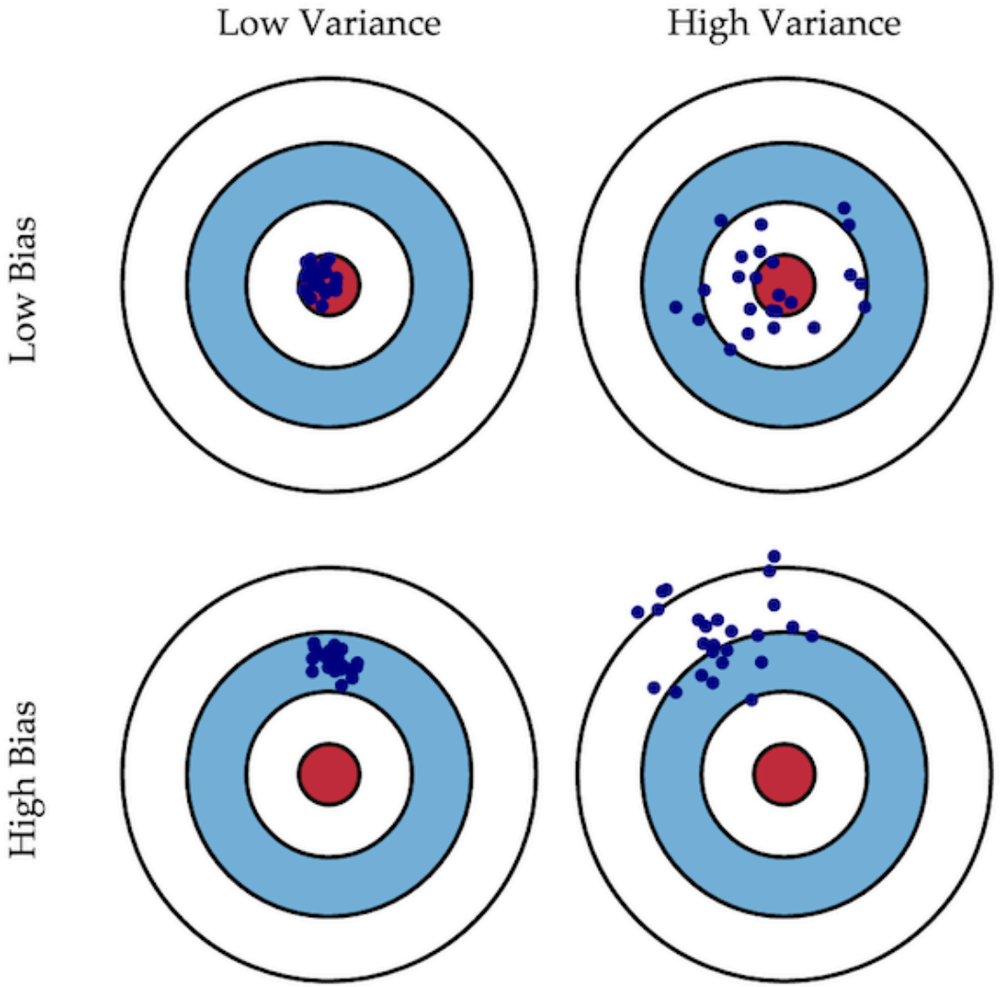


Fig. 1 Graphical illustration of bias and variance.

STANDARD DEVIATION

- What is variance?

- What is standard deviation?

STANDARD DEVIATION

- What is variance?

- What is standard deviation?

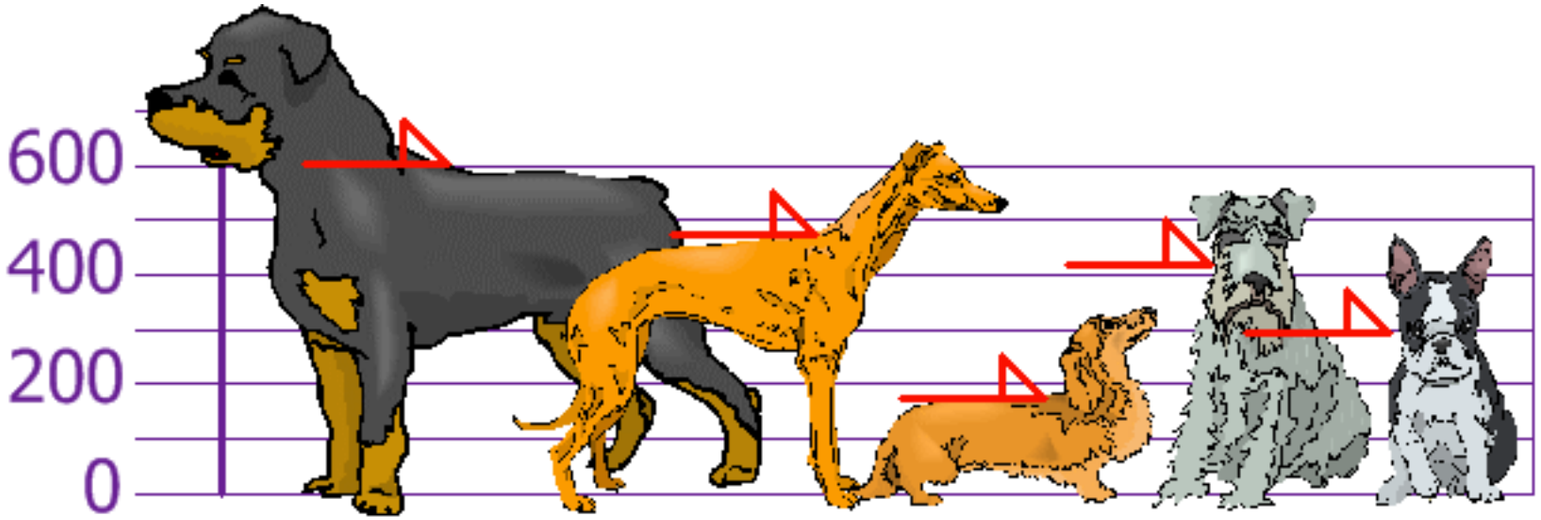
STANDARD DEVIATION

- What is variance?
- Variance is the squared difference between a given observation and the sample mean.

- What is standard deviation?
- Standard deviation is the square root of variance.

STANDARD DEVIATION

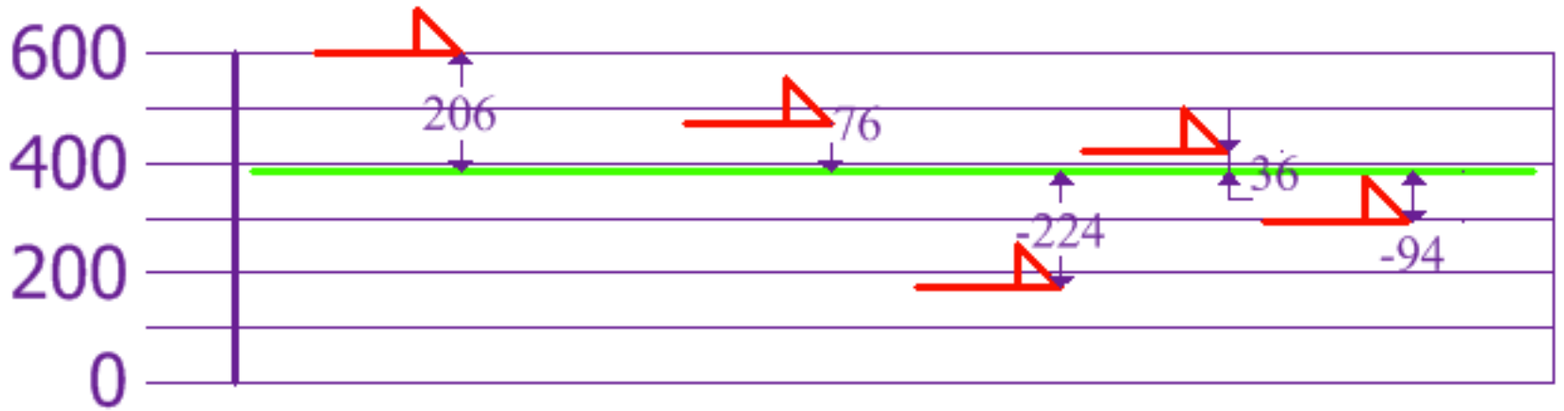
- You’re finding the variance and standard deviation of a bunch of dogs. First, you measure the height of dogs in millimeters



- You find the mean to be 394.

STANDARD DEVIATION

- You then find the difference between each dog's height and the mean



STANDARD DEVIATION

- To calculate variance, square each difference we calculated, and average the result for our sample:

$$\begin{aligned}\text{Variance: } \sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\ &= \frac{108,520}{5} = 21,704\end{aligned}$$

- What is the standard deviation?

STANDARD DEVIATION

- To calculate variance, square each difference we calculated, and average the result for our sample:

$$\begin{aligned}\text{Variance: } \sigma^2 &= \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} \\ &= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5} \\ &= \frac{108,520}{5} = 21,704\end{aligned}$$

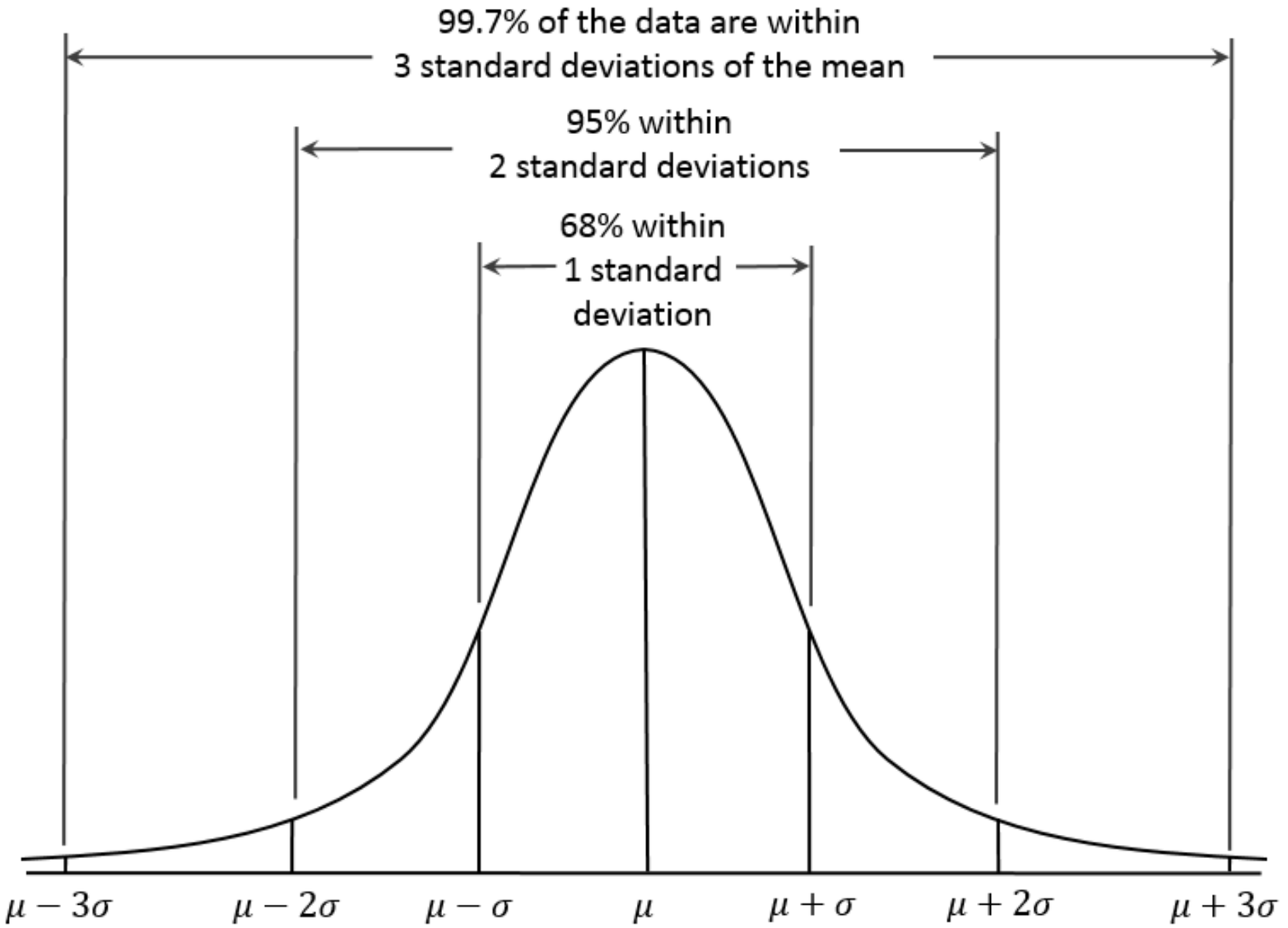
- What is the standard deviation?
- The square root of variance! Roughly equal to 147 here.

THE NORMAL DISTRIBUTION

- A normal distribution is a key assumption to many models we will later be using. But what is normal?

- The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve.

THE NORMAL DISTRIBUTION



THE NORMAL DISTRIBUTION

- We have two key metrics to describe normal distributions!

- **Skewness**
- In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.

THE NORMAL DISTRIBUTION

- We have two key metrics to describe normal distributions!

- **Skewness**
- In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.

- **Kurtosis**
- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.

CENTRAL LIMIT THEOREM

- The central limit theorem is a fundamental tool in statistics. **It says, with some assumptions, that sampling distributions are normal with a specific mean and variance.** It's a vital tool in data science when working with large data sets. Often a random sample (or many random samples) can tell us crucial information about a much larger dataset.

- It says that, as the size n of a sample increases, that:
- the mean of the sample $\bar{x}$ converges to the mean of the true distribution, and
- the standard deviation s of the sample is the same as the true standard deviation σ