Introduction to Statistics

Riley Harper

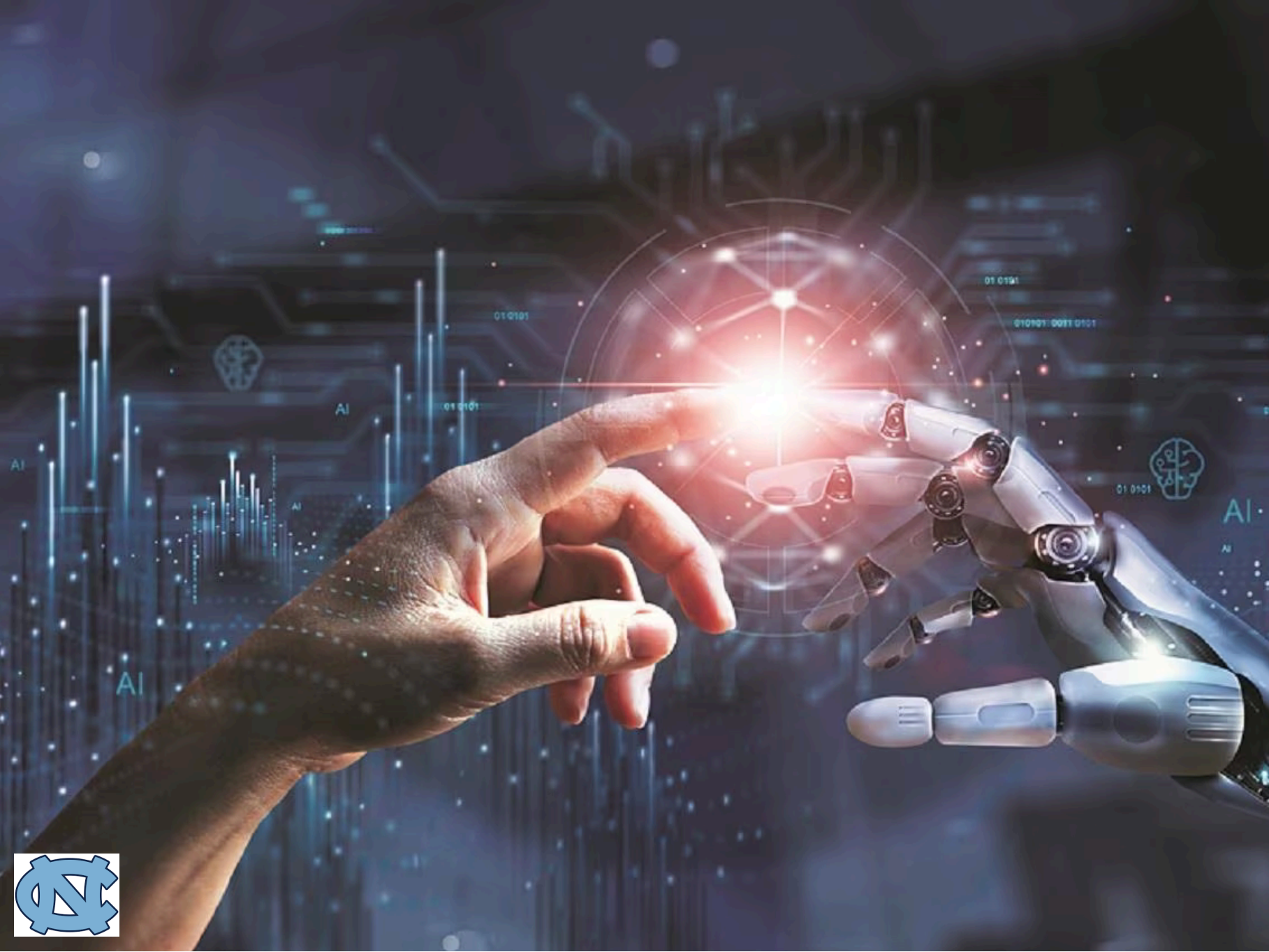
UNC-Chapel Hill

What is Statistics?

"A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

What does this mean today?

- Descriptive Statistics
- Hypothesis Testing
- Big Data
- Data Cleaning
- Data Processing
- Supervised/Unsupervised Learned


AI
AI
AI
AI
AI

Definition

Summation

$$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + \cdots + a_{n-1} + a_n$$

where i is the index of summation; a_i is an indexed variable representing each term of the sum; m is the lower bound of summation, and n is the upper bound of summation. The "$i = m$" under the summation symbol means that the index i starts out equal to m.

The index, i, is incremented by one for each successive term, stopping when $i = n$.

Example

Summation

The summation of i from $i = 1$ to 3 is:

$$\sum_{i=1}^{3} i$$

The summation of i^2 from $i = 3$ to 6 is:

$$\sum_{i=3}^{6} i^2$$

The summation of $\frac{1}{2^n}$ from $i = 0$ to ∞ is:

$$\sum_{i=0}^{\infty} \frac{1}{2^n}$$

Example

Summation

The summation of i from $i = 1$ to 3 is:

$$\sum_{i=1}^{3} i = 1 + 2 + 3 = 6$$

The summation of i^2 from $i = 3$ to 6 is:

$$\sum_{i=3}^{6} i^2 = 3^2 + 4^2 + 5^2 + 6^2 = 86$$

The summation of $\frac{1}{2^n}$ from $i = 0$ to ∞ is:

$$\sum_{i=0}^{\infty} \frac{1}{2^n} = \frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \ldots = 2 \quad \text{(Convergent series!)}$$

Definition

Arithmetic Mean (Average)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

where n is the number of observations for a given sample; x_i is an indexed variable representing each observation of a total of n observations; $\bar{x}$ is the symbol most often used for average though in practice this can be used for other purposes.

Example

Arithmetic Mean (Average)

Find the mean of the following set of numbers: $2, 4, 6, 8, 10$.

$$\frac{1}{5}\sum_{i=1}^{5} x_i = \frac{2+4+6+8+10}{5}$$

Find the weighted mean of the following set of numbers: $80, 85, 90$, and 95 with respective weights of $1, 2, 3$, and 4.

$$\frac{\sum_{i=1}^{4} w_i x_i}{\sum_{i=1}^{4} w_i} = \frac{(1\cdot 80)+(2\cdot 85)+(3\cdot 90)+(4\cdot 95)}{1+2+3+4}$$

Find the mean from a frequency distribution of the following set of numbers: $2, 3, 5, 7$. with respective frequencies of $3, 5, 2, 4$.

$$\frac{\sum_{i=1}^{4} f_i x_i}{\sum_{i=1}^{4} f_i} = \frac{(2\cdot 3)+(3\cdot 5)+(5\cdot 2)+(7\cdot 4)}{3+5+2+4}$$

Example

Arithmetic Mean (Average)

Find the mean of the following set of numbers: $2, 4, 6, 8, 10$.

$$\frac{1}{5}\sum_{i=1}^{5} x_i = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

Find the weighted mean of the following set of numbers: $80, 85, 90$, and 95 with respective weights of $1, 2, 3$, and 4.

$$\frac{\sum_{i=1}^{4} w_i x_i}{\sum_{i=1}^{4} w_i} = \frac{(1\cdot 80)+(2\cdot 85)+(3\cdot 90)+(4\cdot 95)}{1+2+3+4} = \frac{900}{10} = 90$$

Find the mean from a frequency distribution of the following set of numbers: $2, 3, 5, 7$ with respective frequencies of $3, 5, 2, 4$.

$$\frac{\sum_{i=1}^{4} f_i x_i}{\sum_{i=1}^{4} f_i} = \frac{(2\cdot 3)+(3\cdot 5)+(5\cdot 2)+(7\cdot 4)}{3+5+2+4} = \frac{59}{14} \quad \text{(Improper fraction!)}$$

Definition

Median

$$\begin{cases} x_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

where n is the number of observations in a sorted dataset; x_i is the i^{th} observation in the dataset such that for an odd number of observations, the median is the middle value, and for an even number of observations, the median is the average of the two middle values.

Example

Median

Find the median of the following set of numbers: $3, 1, 4$.

$$x_{\frac{n+1}{2}} = x_{\frac{3+1}{2}} = x_2$$

Find the median of the following set of numbers: $5, 3, 8, 7$.

$$\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_2 + x_3}{2}$$

Find the median from a frequency distribution of the following set of numbers: $2, 4, 6$ with respective frequencies of $2, 3, 4$.

$$x_{\frac{n+1}{2}} = x_{\frac{9+1}{2}} = x_5$$

Example

Median

Find the median of the following set of numbers: $3, 1, 4$.

$$x_{\frac{n+1}{2}} = x_{\frac{3+1}{2}} = x_2 = 3$$

Find the median of the following set of numbers: $5, 3, 8, 7$.

$$\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_2 + x_3}{2} = \frac{5+7}{2} = 6$$

Find the median from a frequency distribution of the following set of numbers: $2, 4, 6$ with respective frequencies of $2, 3, 4$.

$$x_{\frac{n+1}{2}} = x_{\frac{9+1}{2}} = x_5 = 4$$

Definition

Mode

$$\underset{x \in \mathbb{R}}{\operatorname{argmax}} \sum_{i=1}^{n} \left(\begin{cases} 1, & \text{if } x = x_i \\ 0, & \text{if } x \neq x_i \end{cases} \right)$$

where x is a value in the dataset, x_i represents each distinct value in the dataset, and n is the number of distinct values; the mode is the value of x for which the sum of the indicator function is maximized such that the most frequently occurring value in the dataset is identified

Example

Mode

Find the mode of the following set of numbers: 2, 2, 3, 4, 4, 4, 5.

$$\underset{x_i}{\operatorname{argmax}} \sum_{i=1}^{n} [x = x_i]$$

Find the mode of the following set of numbers: 1, 1, 2, 3, 3, 3, 4, 4, 4, 5.

$$\underset{x_i}{\operatorname{argmax}} \sum_{i=1}^{n} [x = x_i]$$

Find the mode from a frequency distribution of the following set of numbers: 1, 2, 3, 4 with respective frequencies of 4, 2, 5, 5.

$$\underset{x_i}{\operatorname{argmax}} \sum_{i=1}^{n} f_i [x = x_i]$$

Example

Mode

Find the mode of the following set of numbers: 2, 2, 3, 4, 4, 4, 5.

In this dataset, the indicator function would have the highest sum for the number 4, such that the mode of the dataset is 4.

Find the mode of the following set of numbers: 1, 1, 2, 3, 3, 3, 4, 4, 4, 5.

In this dataset, the indicator function would have the highest sum for the numbers 3 and 4, such that the dataset is bimodal with modes 3 and 4.

Find the mode from a frequency distribution of the following set of numbers: 1, 2, 3, 4 with respective frequencies of 4, 2, 5, 5.

In this distribution, the indicator function will be maximized for the values 3 and 4, such that the dataset is bimodal with modes 3 and 4.

Definition

Range

$$\max(x_i) - \min(x_i)$$

where x_i represents individual observations in a data set, and the range is a measure of the spread or dispersion of the values in the data set, calculated as the difference between the maximum and minimum observed values

Example

Range

Find the range of the following set of numbers: $5, 8, 3, 12$.

$$\max(x_i) - \min(x_i) = 12 - 3$$

Find the range of the following set of numbers: -7, -3, $2, 5, 10$.

$$\max(x_i) - \min(x_i) = 10 - (-7)$$

Find the range from a frequency distribution of the following set of numbers: $2, 4, 6, 9$ with respective frequencies of $0, 3, 2, 1$.

$$\max(x_i) - \min(x_i) = 9 - 4$$

Example

Range

Find the range of the following set of numbers: $5, 8, 3, 12$.

$$\max(x_i) - \min(x_i) = 12 - 3 = 9$$

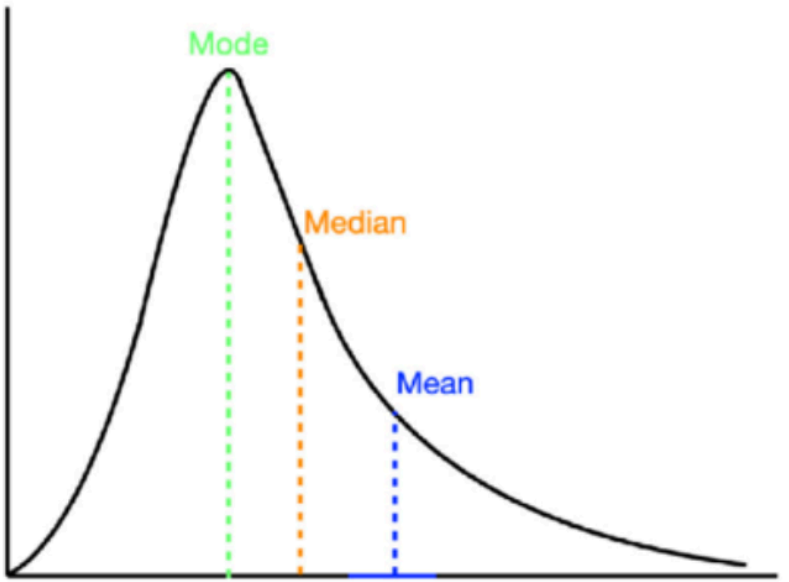
Find the range of the following set of numbers: -7, -3, $2, 5, 10$.

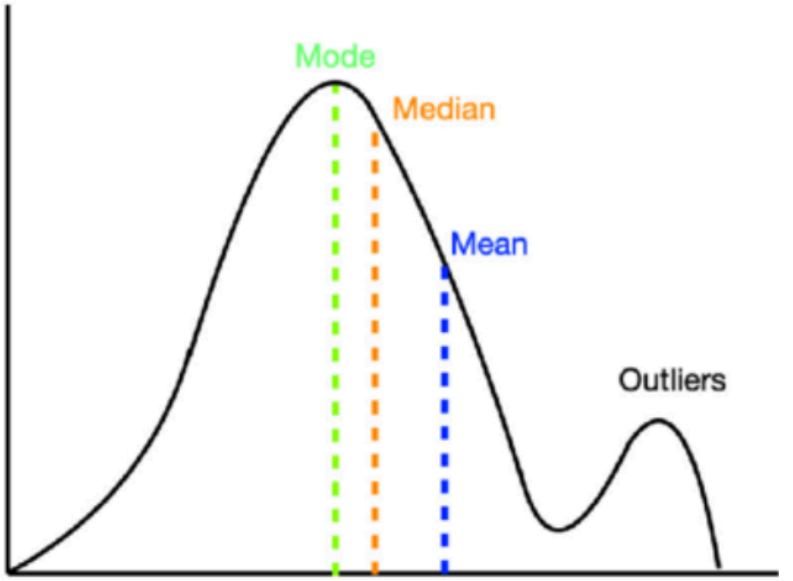
$$\max(x_i) - \min(x_i) = 10 - (-7) = 17$$

Find the range from a frequency distribution of the following set of numbers: $2, 4, 6, 9$ with respective frequencies of $0, 3, 2, 1$.

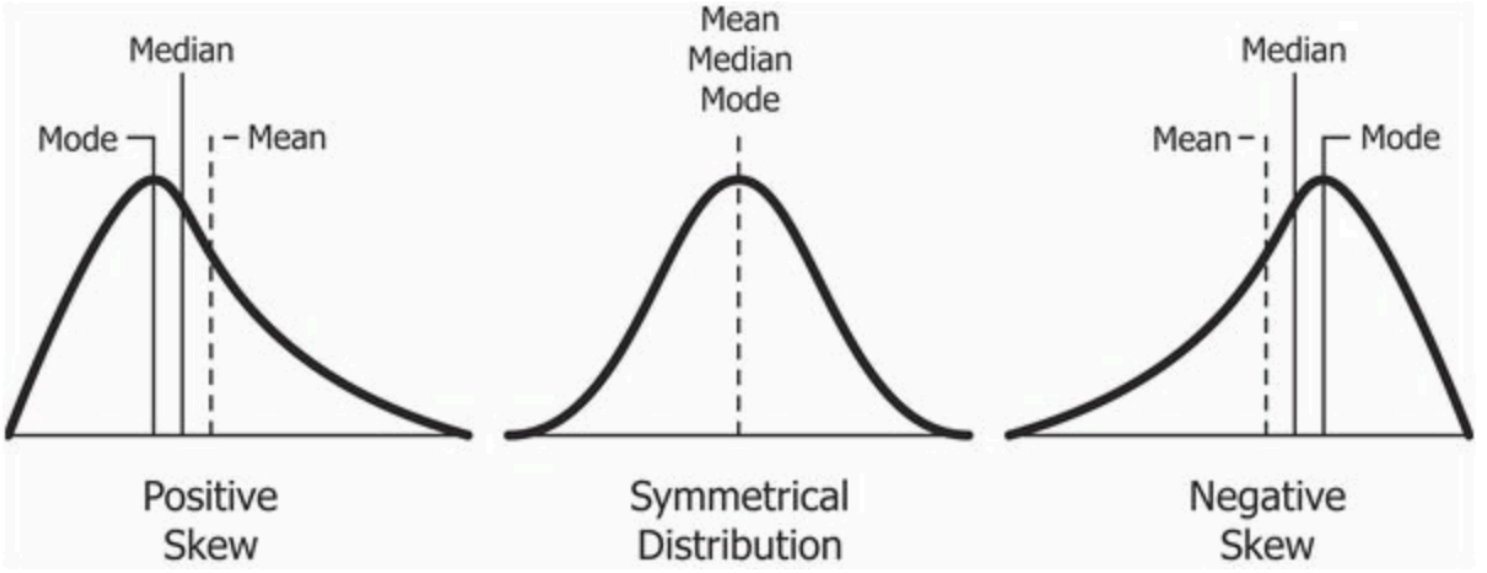
$$\max(x_i) - \min(x_i) = 9 - 4 = 5$$

When to use average vs. median?





What about now?



Definition

Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where x represents a variable, μ is the mean of the distribution, σ is the standard deviation, σ^2 is the variance, and e is Euler's number, the base of the natural logarithms. The normal distribution is symmetrical, with a bell-shaped curve where the majority of observations cluster around the mean, and probabilities for x taper off symmetrically in both directions from the mean

Are there other distributions?

Are there other distributions?

- Poisson Distribution
- Heavy-Tailed Distribution
- Tukey's Lambda Distribution
- Exponential Distribution
- Uniform Distribution
- Binomial Distribution
- Geometric Distribution
- Beta Distribution

Each distribution has its own unique properties and applications, ranging from modeling the number of events occurring in a fixed interval of time or space (Poisson Distribution), to describing outcomes with only two possible values (Binomial Distribution), to modeling waiting times (Exponential Distribution), ...

What tools exist to help with data cleaning, modeling, and analysis?

R

Python

and other point and click software like Jamovi and SPSS for those who would prefer as little programming as possible. Note that there are many other languages available for data analysis other than R and Python though today these are the two most often used languages for this purpose.

How can I learn to code with open-source?

With *open source* through GitHub, you can host and review code, manage projects, and build software alongside 50 million developers.

Pros of Open Source for Learning:

- **Collaboration:** Engage with a community of learners and developers.
- **Transparency:** Access the source code and understand the inner workings.
- **Diversity of Projects:** Contribute to projects that range from beginner-friendly to advanced.
- **Real-World Experience:** Work on projects that are used in the real world.
- **Resources and Tools:** Utilize a plethora of free tools and resources available for learning.

StatQuest

Calculating the Mean, Variance and Standard Deviation, Clearly Explained!!!

Questions?