Statistics in Data Science

overview

1. **Introduction to Statistics in Data Science**
2. **Distributions**
3. **Distribution Estimators: MoM, MLE, KDE**
4. **Point Estimates**
5. **Statistical Hypothesis Testing**
6. **Correlation**
7. **Practical Examples**

1. Introduction to Statistics in Data Science

- **Statistics** is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

- Why Statistics for Data Science?

Data gathered are all raw data, and raw data do not provide meaningful information. That's why we need statistics to collect, organize and analyze data. With statistics,

  - which observation is the most occurring?
  - Is there a difference between the two experiments?
  - Is the collected sample a representation of the population?
  - Is the result obtained significant enough to make a difference?
  - …….

These questions can be answered by statistics and transform the raw data into meaningful information.

- **Measures of Central Tendency**
- Measures of center are statistics that give us a sense of the "middle" of a numeric variable. In other words, centrality measures give you a sense of a typical value you'd expect to see. Common measures of center include the mean, median and mode.
- mean, median, and mode.

Statistics Mean

If you're trying to find the mean in statistics, what you are looking for most of the time is the average of a data set (the Arithmetic Mean).

To find the mean: add up all the numbers and then divide by the number of items in the set. For example,

The average of 1, 2, 6, 8, 10 is:

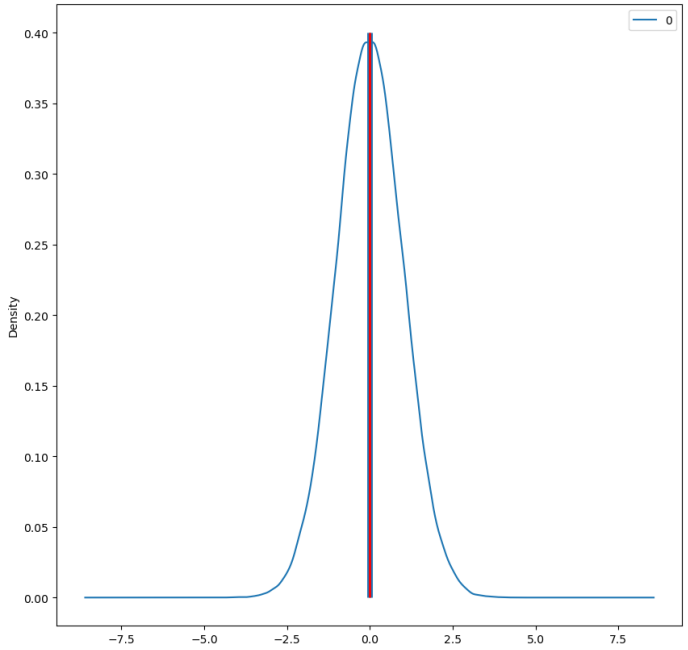
Mean vs Median

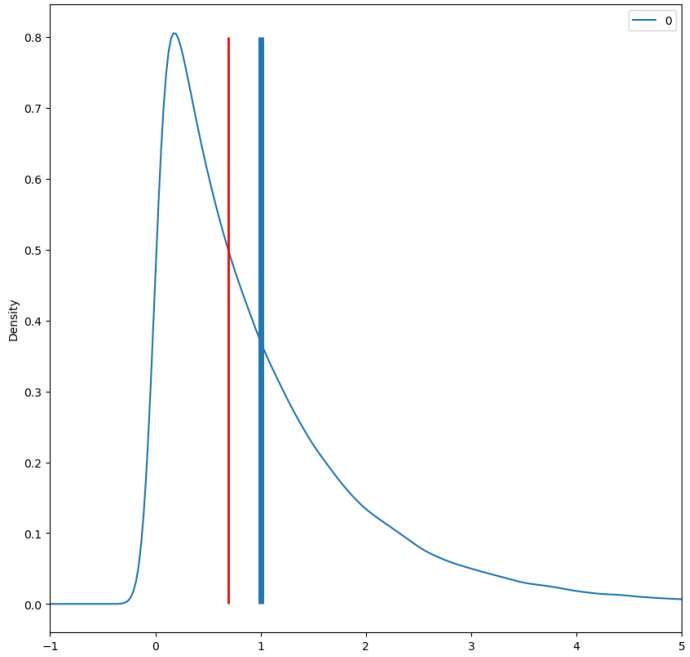
Both are measures of where the center of a data set lies, but they are usually different numbers. For example, take this list of numbers: 10,10,20,40,70.

- The ***mean*** (average) is found by adding all of the numbers together and dividing by the number of items in the set: 10 + 10 + 20 + 40 + 70 / 5 = 30.
- The ***median*** is found by ordering the set from lowest to highest and finding the exact middle. The median is just the middle number: 20.

Sometimes the two will be the same number. For example, the data set 1,2,4,6,7 has an average of 1 + 2 + 4 + 6 + 7 / 5 = 4 and a median (a middle) of 4.

- Explore the median as an alternative measure of central tendency.





Blue: mean
Red: median

Since the median tends to resist the effects of skewness and outliers, it is known a "robust" statistic.
The median generally gives a better sense of the typical value in a distribution with significant skew or outliers.

- The **mean** is commonly used with continuous numerical data. It represents a typical or "average" value in the dataset.

- The **median** is the middle value when the data is sorted in ascending or descending order. The median is robust to outliers and is often preferred when dealing with skewed data. Like the mean, it is used with numerical data.

- The **mode** of a variable is simply the value that appears most frequently. Unlike mean and median, you can take the mode of a categorical variable and numerical data and it is possible to have multiple modes.

- Use as df.mean(); df.median(); df.mode().

- **Measures of Spread**
- **Spread** in statistics refers to the degree of variability or dispersion in a dataset. It measures how data points are scattered or spread out from the central tendency.
- Measures of spread (dispersion) are statistics that describe how data varies.
- While measures of center give us an idea of the typical value, measures of spread give us a sense of how much the data tends to diverge from the typical value.
- quartiles, variance, standard deviation, skewness, and kurtosis as measures of spread.

Quartile

As noted earlier, the *median represents the 50th percentile of a data set*.

- A summary of several percentiles can be used to describe a variable's spread.
- We can extract the
  - minimum value (0th percentile),
  - first quartile (25th percentile),
  - the median - second quartile (50th percentile),
  - third quartile(75th percentile), and
  - maximum value (100th percentile)

- Quartiles are a statistical measure used to divide a dataset into four equal parts.

```
mtcars.head(10)
```

	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
5	Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
6	Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
7	Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
8	Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
9	Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4

Use df.quantile():

```
five_num = [mtcars["mpg"].quantile(0),
            mtcars["mpg"].quantile(0.25),
            mtcars["mpg"].quantile(0.50),
            mtcars["mpg"].quantile(0.75),
            mtcars["mpg"].quantile(1)]

five_num
```

```
[10.4, 15.425, 19.2, 22.8, 33.9]
```

Since these values are commonly used to describe data, they are known as the "five number summary".
They are the same percentile values returned by df.describe():

```
mtcars["mpg"].describe()
```

```
count    32.000000
mean     20.090625
std       6.026948
min      10.400000
25%      15.425000
50%      19.200000
75%      22.800000
max      33.900000
Name: mpg, dtype: float64
```

- **Variance and Standard Deviation**
- variance and standard deviation are measures of spread.

The **variance** of a distribution is the ***average of the squared deviations (differences) from the mean***.

$$variance = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Use df.var()

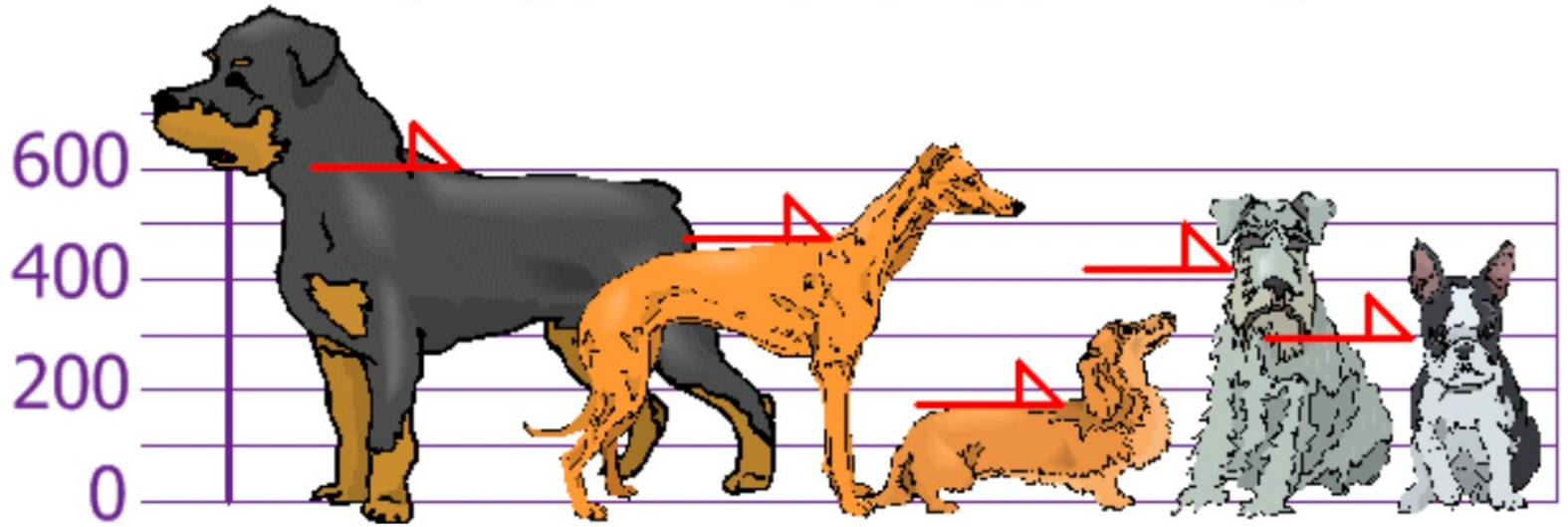
The **standard deviation** is the ***square root of the variance***.

$$std - dev = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Use df.std()

Dog Example:

You and your friends have just measured the heights of your dogs (in millimeters): dogs on graph shoulder heights



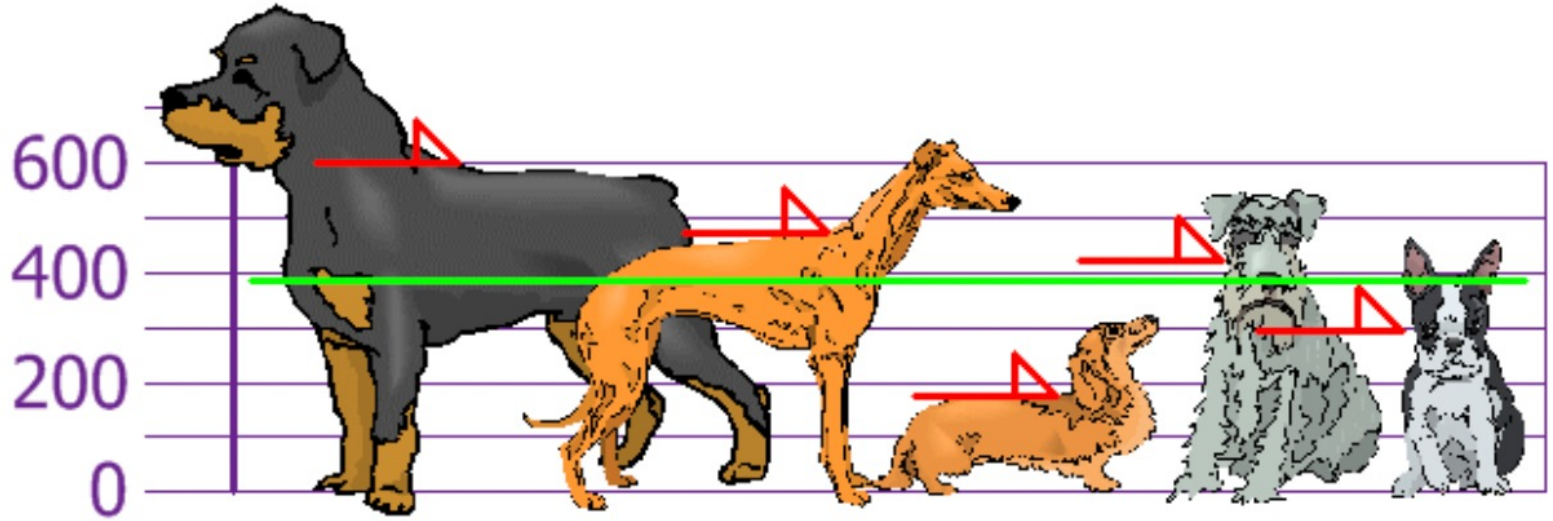
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

$$\text{Mean} = (600 + 470 + 170 + 430 + 300) / 5 = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:

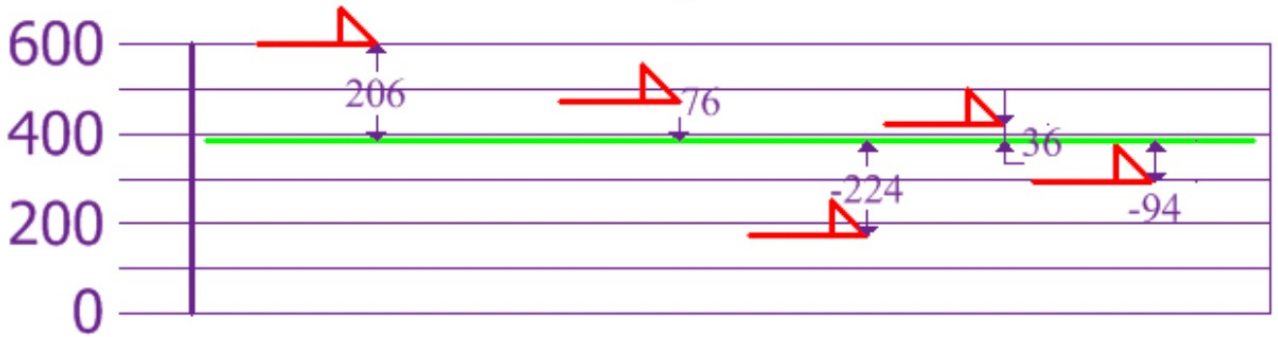


The **variance** of a distribution is the ***average of the squared deviations (differences) from the mean***.

$$variance = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Now we calculate each dog's difference from the Mean:

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



So the

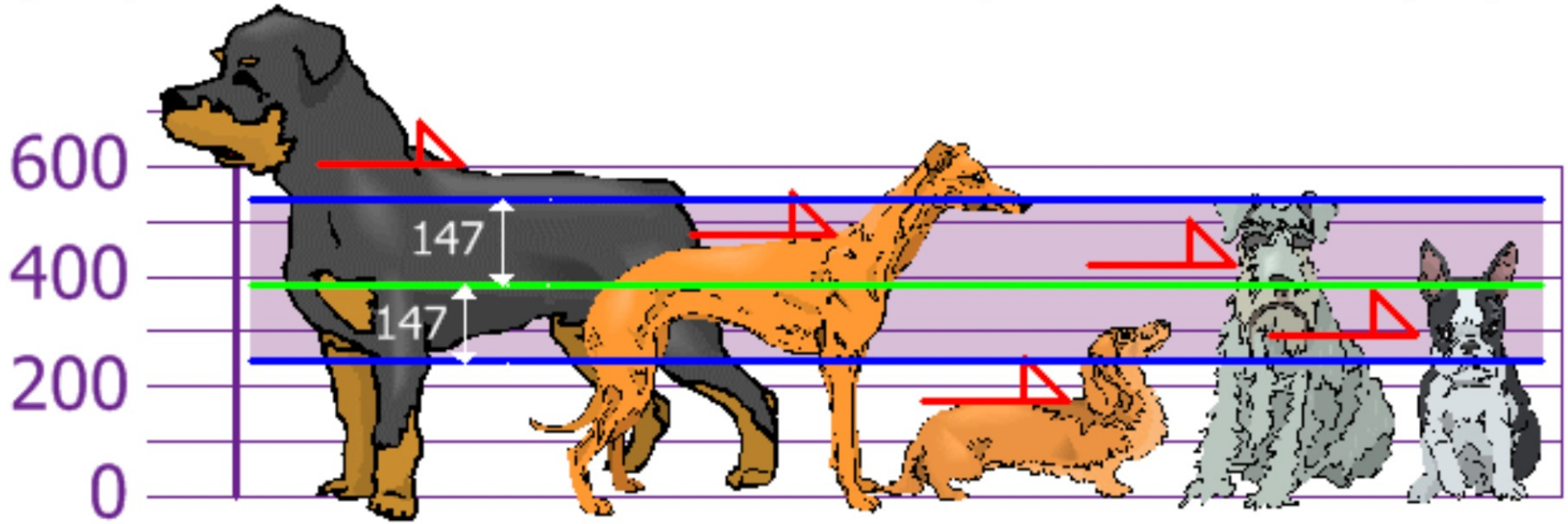
$$Variance = 21,704$$

And the Standard Deviation is just the square root of Variance, so:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

$$std - dev = \sigma = 147$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:

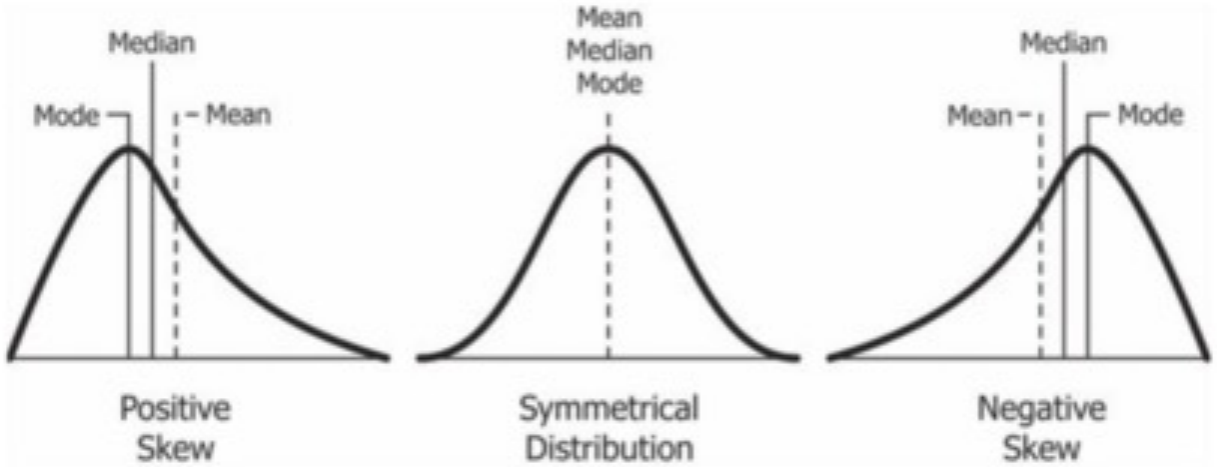


So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

- **Skewness and Kurtosis**

  Beyond measures of center and spread, descriptive statistics include measures that give you a sense of the shape of a distribution.

- **Skewness** measures the skew or asymmetry of a distribution.
- **Kurtosis** measures of the "tailedness" of a distribution.



- **Kurtosis** defines kurtosis as a measure of the "tailedness" of a distribution.

https://en.wikipedia.org/wiki/Kurtosis#:~:text=In%20probability%20theory%20and%20statistics,aspect%20of%20a%20probability%20distribution.

- Use df. skew(); df.kurt().

- https://github.com/q-tong/CS405-605-Data-Science/blob/main/Fall2023/Lecture/3.Statistics/01_Stats_Basics.ipynb

2. Distributions

- Define probability distribution.

  Probability measures how likely it is for an event to occur on a scale from 0 (the event never occurs) to 1 (the event always occurs).

- Discrete and continuous distributions.
- common probability distributions:

  Uniform distribution; Normal distribution; Exponential distribution; Poisson distribution

- Include probability density functions (PDFs) and cumulative distribution functions (CDFs).

3. Estimates

- Explain the concept of estimation in statistics.
- Discuss point estimates and interval estimates.
- Introduce the idea of sample statistics as estimators of population parameters:
  - Sample mean as an estimator of population mean.
  - Sample standard deviation as an estimator of population standard deviation.
- Discuss the properties of good estimators.

3. Distribution Estimators: MoM, MLE, KDE

- Introduce the concept of distribution estimation.
- Explain the Method of Moments (MoM) as an estimation technique based on sample moments.
- Discuss Maximum Likelihood Estimation (MLE) as an alternative technique for estimating distribution parameters.
- Mention Kernel Density Estimation (KDE) as a non-parametric method for estimating probability densities.
- Provide examples and use cases for each estimation method.

https://github.com/q-tong/CS405-605-Data-Science/blob/main/Fall2023/Lecture/3.Statistics/03_Distribution_Estimators.ipynb

4. Point Estimates

- Point estimation involves using a single value (point estimate) to approximate a population parameter. The most common point estimate is the sample mean, which is used to estimate the population mean.

- https://github.com/q-tong/CS405-605-Data-Science/blob/main/Fall2023/Lecture/3.Statistics/04_Point%20Estimates.ipynb

5. Statistical Hypothesis Testing

- Define hypothesis testing and its significance in statistics.
- The steps involved in hypothesis testing:
  - Formulate null and alternative hypotheses.
  - Choose a significance level (alpha).
  - Collect and analyze data.
  - Make a decision based on the test statistic and p-value.
  - Draw conclusions.
- Explain Type I and Type II errors.

6. Correlation

- Define correlation and its role in statistics.
- Discuss the Pearson correlation coefficient (r) as a measure of linear association between two variables.
- Interpret the sign and magnitude of the correlation coefficient.
- Illustrate correlation using scatterplots.

7. Practical Examples

- Include practical examples and case studies that demonstrate the application of statistical concepts in data science.
- Show real-world scenarios where understanding distributions, estimation, hypothesis testing, correlation, and distribution estimators are crucial.