Statistics -2

Quantiles and percentiles :

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

There are several types of quantiles used in statistical analysis, including:

a. Quartiles: Divide the data into four equal parts, Q1 (25th percentile), Q2 (50th percentile or median), and Q3 (75th percentile).

b. Deciles: Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).

c. Percentiles: Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).

d. Quintiles: Divides the data into 5 equal parts

Things to remember while calculating these measures:

1. Data should be sorted from low to high

2. You are basically finding the location of an observation

3. They are not actual values in the data

4. All other tiles can be easily derived from Percentiles

Percentile

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

Formula to calculate the percentile value:

PL = P/100 *(N+1)

where:

• PL = the desired percentile value location

• N = the total number of observations in the dataset

• p = the percentile rank (expressed as a percentage)

Example:

Find the 75th percentile score from the below data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step1 - Sort the data

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

* percentile :→

→ statical measure that Represent the percentage of observation in a dataset that fall below a particular value

(Basically in 100 student you gain 93 percentile मतलब 93 Percent log तुमसे पीछे है और 7 % तुमसे आगे ।)

formula →

$$PL = \frac{P}{100}(N+1)$$

PL → Desired percentile value location

N → total No. of observation in a dataset

P → Percentile Rank

Example

① find 75th Percentile score from the data set.

78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Step ① sort the data first

78	82	84	88	91	93	94	96	98	99
1	2	3	4	5	6	7	8	9	10

8.25

$$PL = \frac{75}{100}(10+1) = \frac{33}{4} = 8.25$$

$96 + \{0.25(98-96)$

$= 96 + 0.25 \times 2$

$= \boxed{96.5 = 75^{th} \text{ percentile}}$

96	98
↓	↓
8	9

50^{th} percentile $PL = \frac{50}{100}(10+1) = 0.5 \times 11 = 5.5$

91	93
5	6

$91 + 0.5(93-91)$

$\Rightarrow 91 + 0.5 \times 2 = \boxed{92 = 50^{th} \text{ percentile}}$

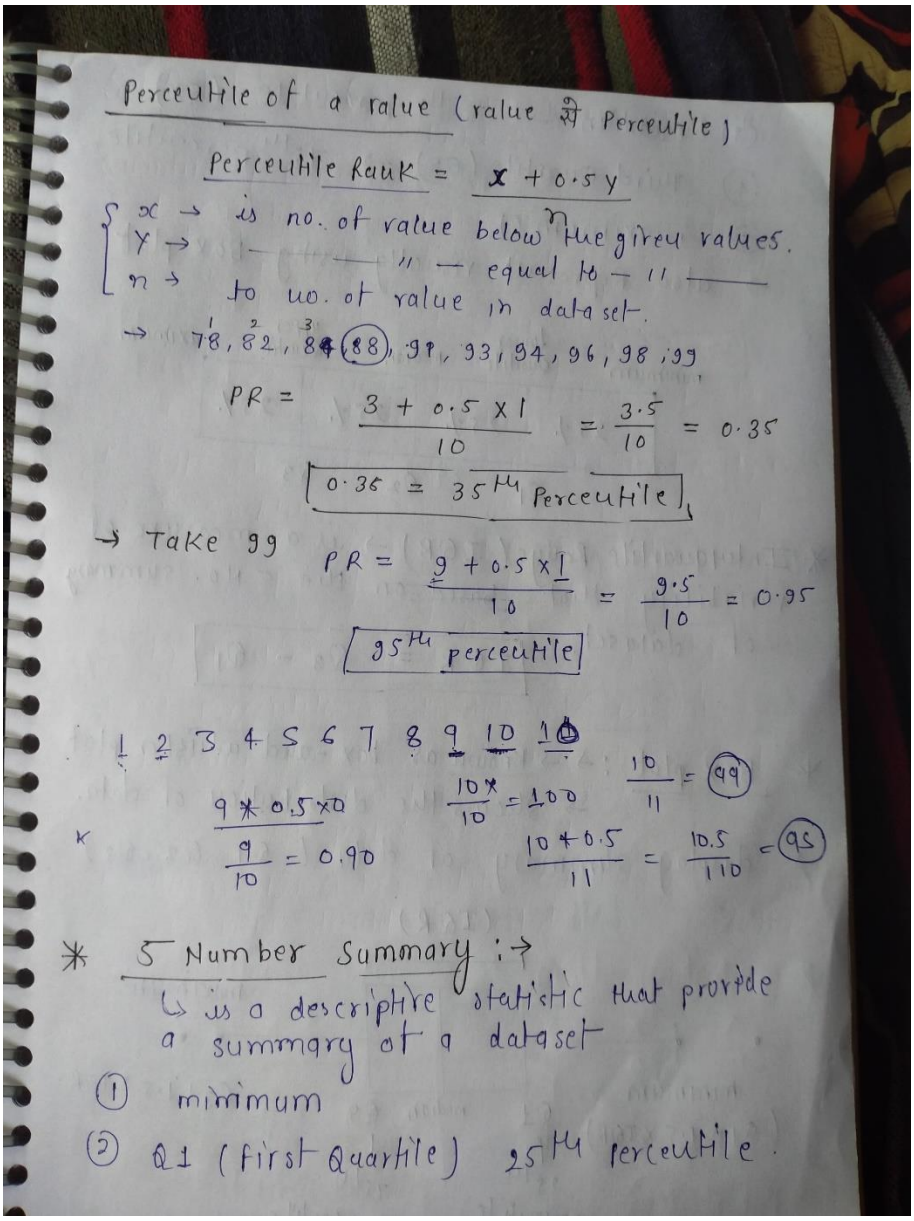
Percentile of a value

Percentile rank = (X+0.5*Y)/N

X = number of values below the given value

Y = number of values equal to the given value

N = total number of values in the dataset



5 number summary

13 March 2023 06:57

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. **Minimum value**: The smallest value in the dataset.

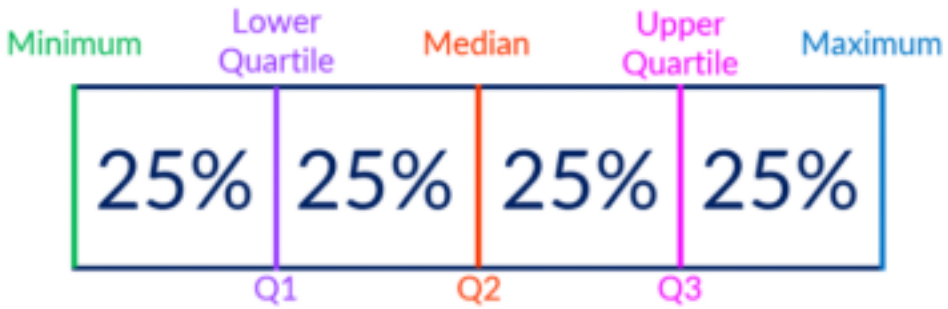
2. **First quartile (Q1)**: The value that separates the lowest 25% of the data from the rest of the dataset.

3. **Median (Q2)**: The value that separates the lowest 50% from the highest 50% of the data.

4. **Third quartile (Q3)**: The value that separates the lowest 75% of the data from the highest 25% of the data.

5. **Maximum value**: The largest value in the dataset.

The five-number summary is often represented visually using a box plot, which displays the range of the dataset, the median, and the quartiles. The five-number summary is a useful way to quickly summarize the central tendency, variability, and distribution of a dataset.



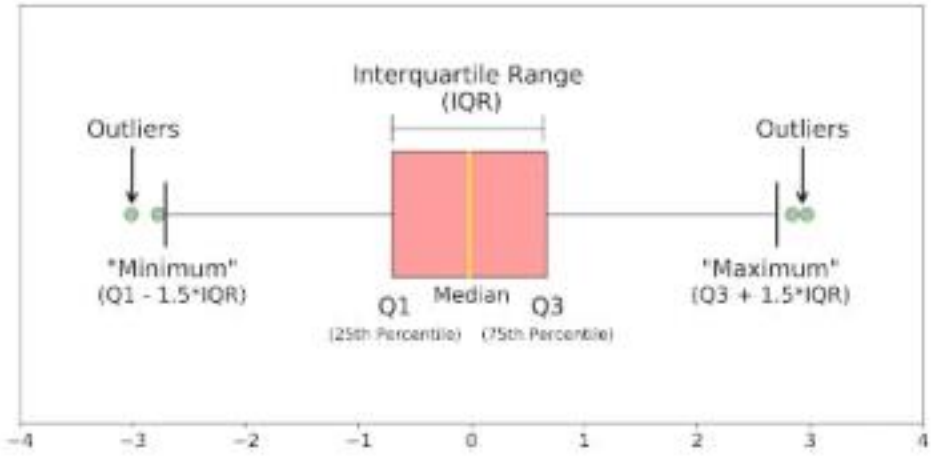
Interquartile Range

The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

$$IQR = Q3 - Q1$$

1. What is a boxplot

A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that shows the distribution of the data. The box plot displays a summary of the data, including the minimum and maximum values, the first quartile (Q1), the median (Q2), and the third quartile (Q3).



2. How to create a boxplot with example

How to create a box plot :→

6	213	241	260	281	290	314	321	350	1500
1	2	3	4	5	6	7	8	9	10

$Q1 = \frac{25}{100} \times 11 = \frac{11}{4} = 2.75$

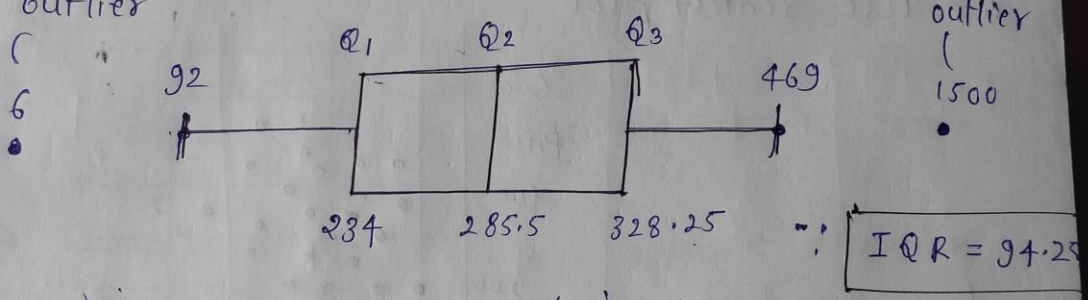
$213 + 0.75\,(241 - 213) = \boxed{234}$

$Q2 = \frac{50}{100} \times 11 = 5.5$

$281 + 0.5\,(290 - 281) = \boxed{285.5}$

$Q3 = \frac{75}{100} \times 11 = \frac{33}{4} = 8.25$

$321 + 0.25\,(350 - 321) = \boxed{328.25}$



$\therefore$ $\boxed{IQR = 94.25}$

$\text{minimum} = Q_1 - 1.5(IQR) = 234 - 1.5(94.25) = 92.6$

$\text{maximum} = Q_3 + 1.5(94.25) = 469.6$

---

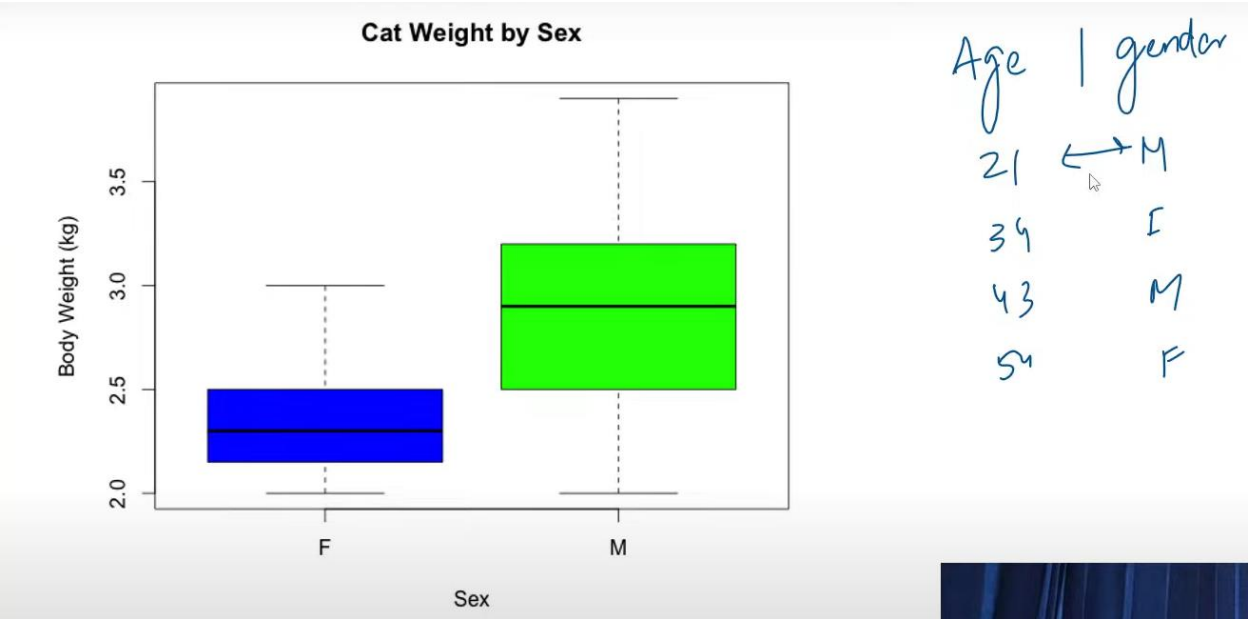
* Easy way to see distribution of data
* Tell about skewness of data
* identified outlier.
* compare 2 category data

} Benifits

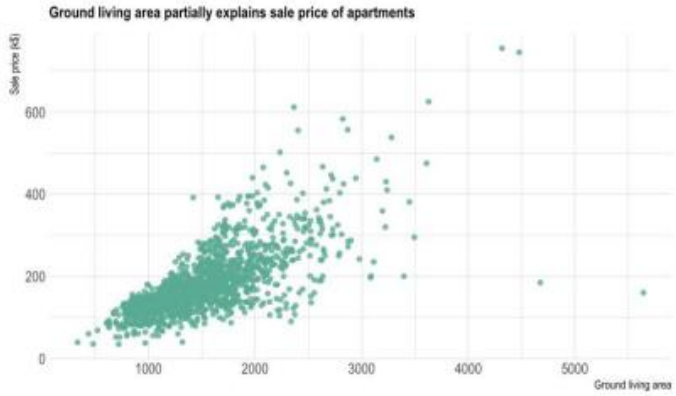
1. Benefits of a Boxplot

- Easy way to see the distribution of data
- Tells about skewness of data
- Can identify outliers
- Compare 2 categories of data

Side by side box plot :



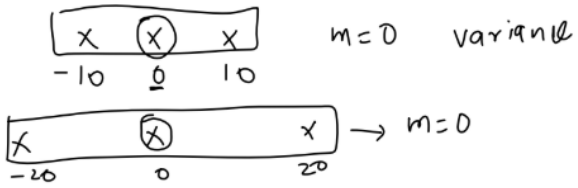
Scatter plot : when we have 2 numerical column then we create scatter plot it show the co -relation between two numerical column .

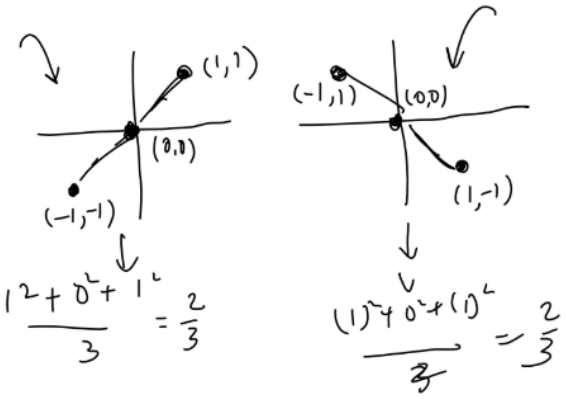


Covariance

13 March 2023 06:57

- What problem does Covariance solve?

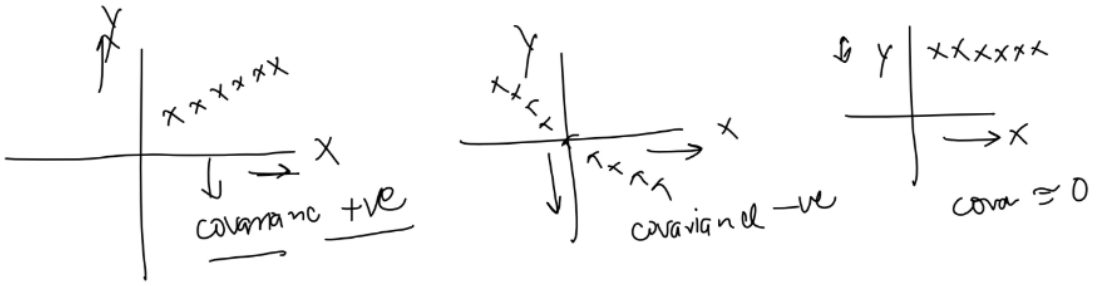




- What is covariance and how is it interpreted?

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together, such that when one variable increases, does the other variable also increase, or does it decrease?

If the covariance between two variables is positive, it means that the variables tend to move together in the same direction. If the covariance is negative, it means that the variables tend to move in opposite directions. A covariance of zero indicates that the variables are not linearly related.



- How is it calculated?

Covariance Formula	
Population	**Sample**
$\sigma_{xy} = \frac{\sum(X-\mu_x)(Y-\mu_y)}{N}$	$s_{xy} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{n-1}$
X, Y – The Value of X and Y in the Population *μ_x, μ_y – The population Mean of X and Y* *N – Total Number of Observations*	*X, Y – The Value of X and Y in the Sample Data* *$\bar{x}, \bar{y}$ – The Sample Mean of X and Y* *n - Total Number of Observations*

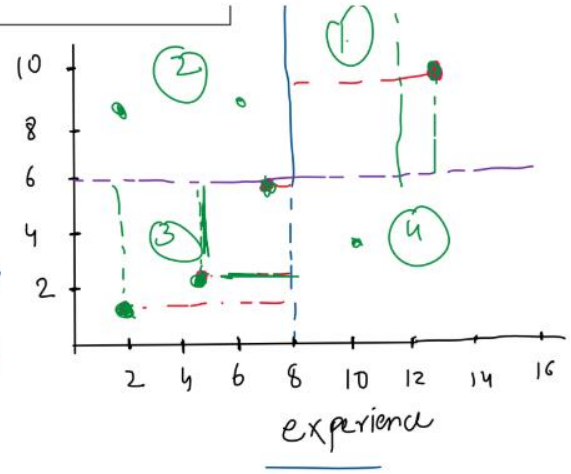
Exp(x)	Salary(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	-1
12	12	4	6	24
13	10	5	4	20

$\bar{X} = 8 \quad \bar{Y} = 6$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$n-1$

$\frac{85}{4} =$ COV = 21.5



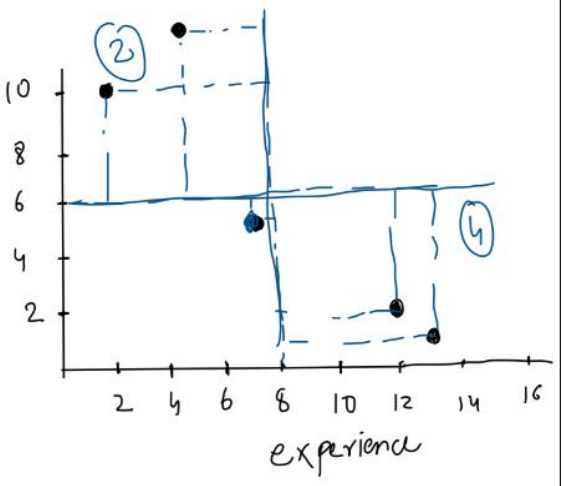
- Measures how two variables change together.
  - **Positive Covariance**: Variables increase together.
  - **Negative Covariance**: One increases, the other decreases.
  - **Zero Covariance**: No linear relationship.
- **Limitation**: Doesn't indicate the strength of the relationship.

Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10	−6	4	−24
5	12	−3	6	−18
8	5	0	−1	0
12	2	4	−4	−16
13	1	5	−5	−25

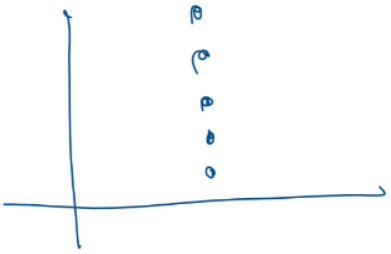
$\bar{X} = 8$ $\bar{y} = 6$

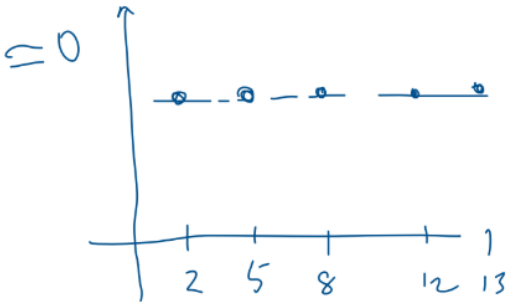
$\frac{-83}{4}$

-21.3

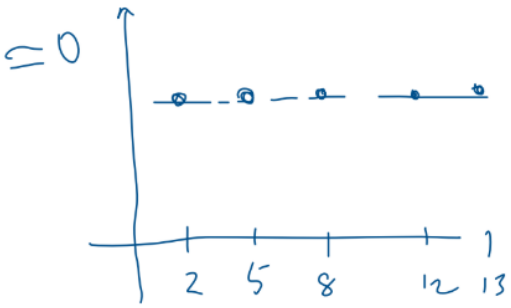


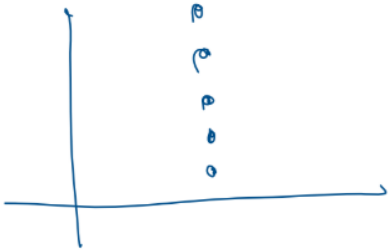
Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10			
5	10			
8	10			
12	10			
13	10			





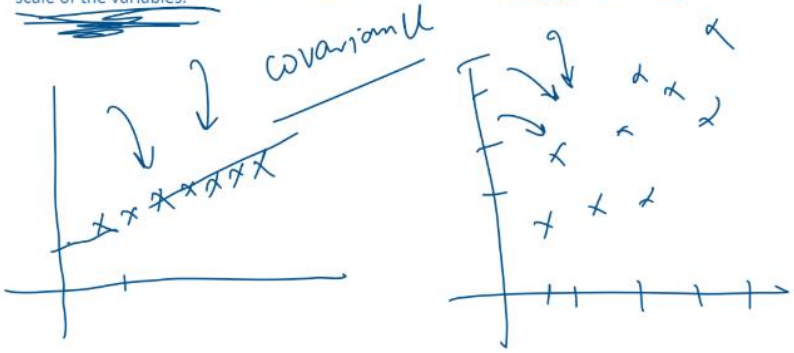
Backlogs(x)	package(y)	X-Xmean	Y-Ymean	(X-Xmean)*(Y-Ymean)
2	10			
5	10			
8	10			
12	10			
13	10			





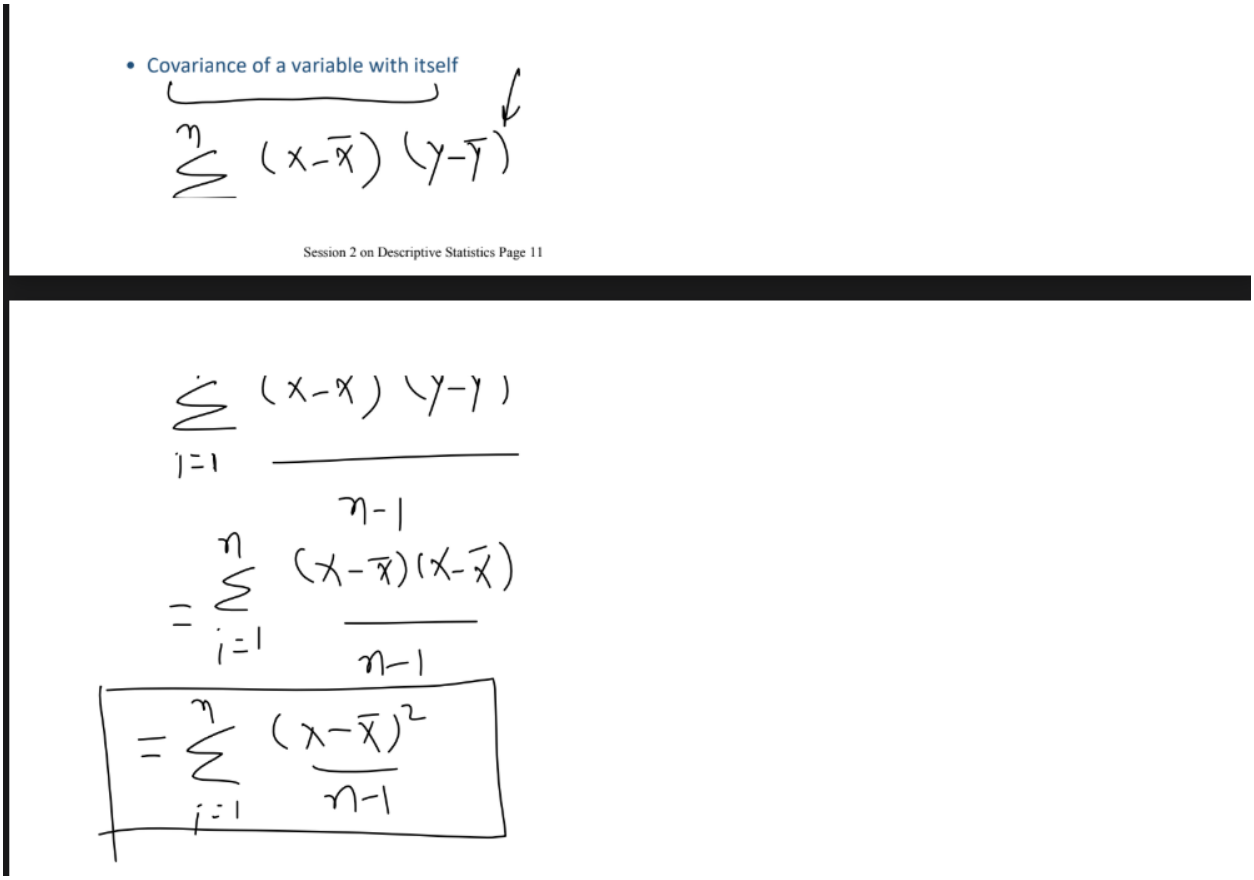
- Disadvantages of using Covariance

One limitation of covariance is that it does not tell us about the strength of the relationship between two variables, since the magnitude of covariance is affected by the scale of the variables.



- Covariance of a variable with itself

$$\sum (x-\bar{x})(y-\bar{y})$$



Real-World Example of Covariance

Scenario: Monthly Temperature vs. Ice Cream Sales

- Imagine you're analyzing the relationship between **temperature** and **ice cream sales** over several months.
    - When temperature increases, ice cream sales also tend to increase (positive covariance).
    - Conversely, if temperature decreases, ice cream sales decrease.

Interpretation:

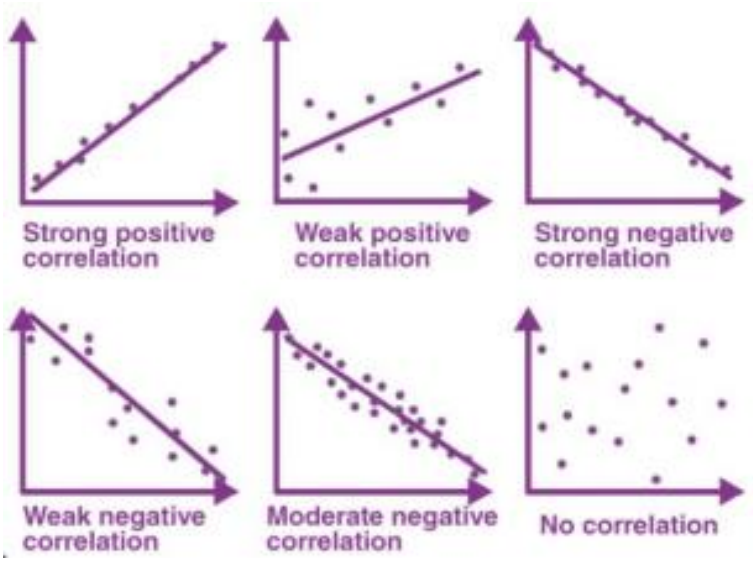
- Covariance tells you whether the two variables move together (positive) or in opposite directions (negative), but it doesn't show the strength of this relationship.

7. Correlation

- Quantifies the strength and direction of a relationship between variables.
- Measured using the **correlation coefficient**:
  - **+1**: Perfect positive correlation.
  - **0**: No correlation.
  - **-1**: Perfect negative correlation.

Correlation

. What problem does Correlation solve?



- Can we quantify this weak and strong relationship?
- 2. What is correlation?
- Correlation refers to a statistical relationship between two or more variables. Specifically, it measures the degree to which two variables are related and how they tend to change together.
- Correlation is often measured using a statistical tool called the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative correlation, a correlation coefficient of 0 indicates no correlation, and a correlation coefficient of 1 indicates a perfect positive correlation.

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Real-World Example of Correlation

Scenario: Study Hours vs. Exam Scores

- A study finds that students who study more tend to score higher in exams.
    - The **correlation coefficient** for study hours and exam scores is **+0.85** (a strong positive relationship).

Another Example: Height vs. Weight

- Taller people generally weigh more.
    - The correlation coefficient might be around **+0.7**, indicating a moderately strong positive relationship.

Interpretation:

- Correlation provides both the **direction** (positive/negative) and the **strength** of the relationship between variables, unlike covariance.

Correlation vs. Causation:

Correlation

- Describes a statistical relationship between two variables, showing how they move together.
- It **does not imply one variable causes the other** to change.
- Example: More umbrellas sold are correlated with more rain, but umbrellas don't cause rain.

Causation

- Means that one variable **directly causes a change** in another.
- To prove causation, you often need controlled experiments or strong evidence.
- Example: Increasing water intake leads to better hydration.

Key Difference:

- **Correlation**: Indicates association (variables may move together, but for unknown reasons).
- **Causation**: Proves one variable directly affects the other.

- Correlation shows a relationship but not the cause.
    - Example: Firefighters and fire damage are correlated, but firefighters don’t cause more damage. The severity of the fire is the actual cause.

Correlation vs. Causation Example

Scenario: Ice Cream Sales and Shark Attacks

- **Observation**: Data shows a positive correlation between **ice cream sales** and **shark attacks**.
    - When ice cream sales increase, shark attacks also increase.

Does this mean ice cream causes shark attacks?

- **No!** This is a case of correlation, not causation.
- **Reason**: Both are influenced by a third variable: **hot weather**.
    - Hot weather drives more people to the beach (increasing shark attacks) and boosts ice cream sales.

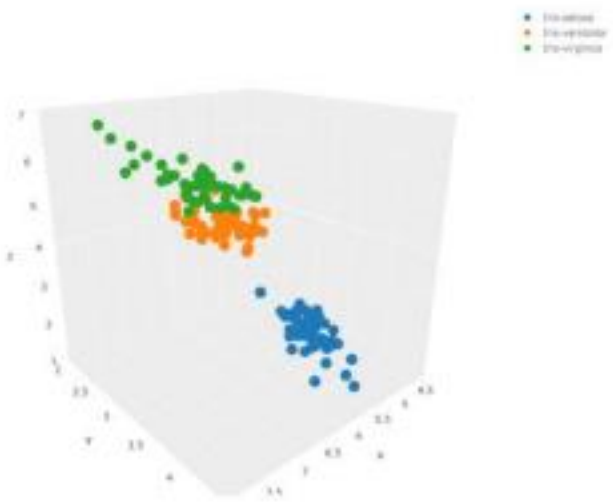
Key Takeaway:

- **Correlation**: Ice cream sales and shark attacks are related.
- **Causation**: Eating ice cream doesn't cause shark attacks; the real factor is hot weather.

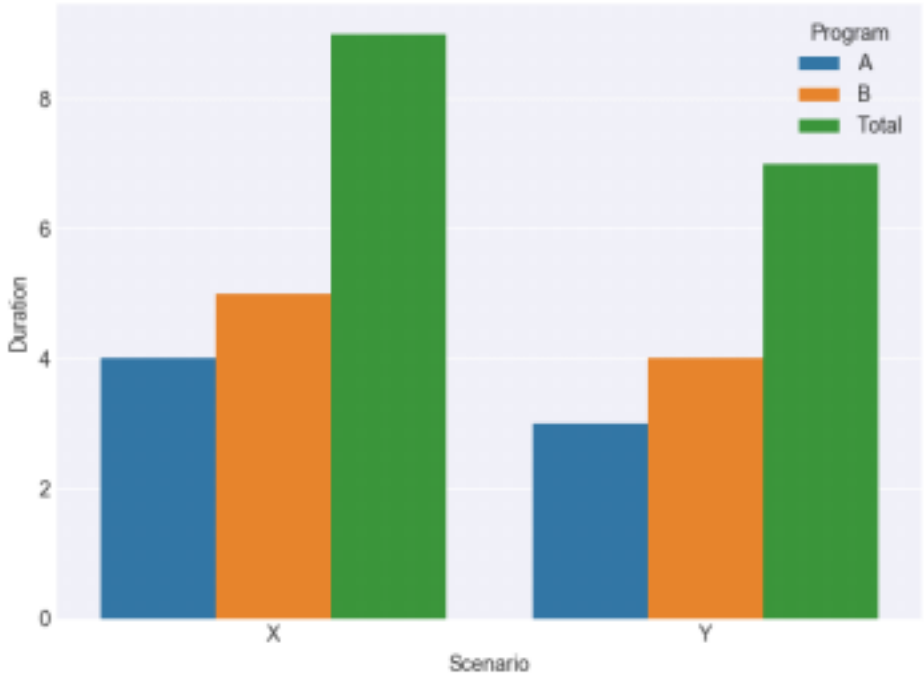
Always investigate other possible factors before assuming causation!

Visualizing Multiple Variables

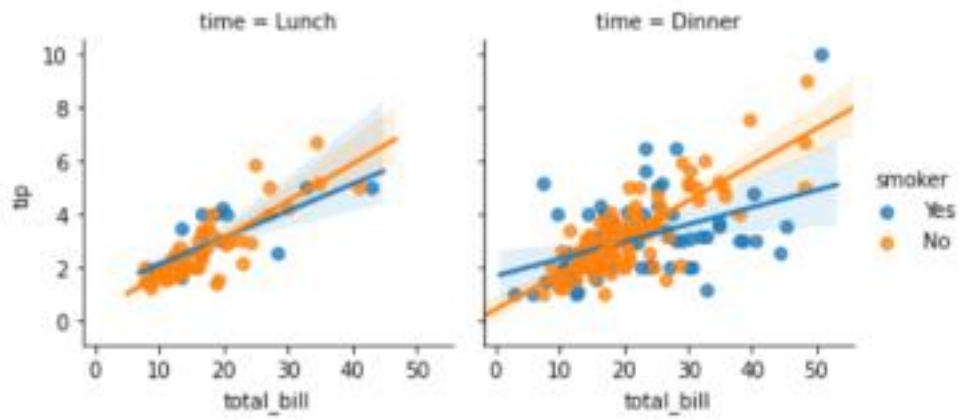
1. 3D Scatter Plots



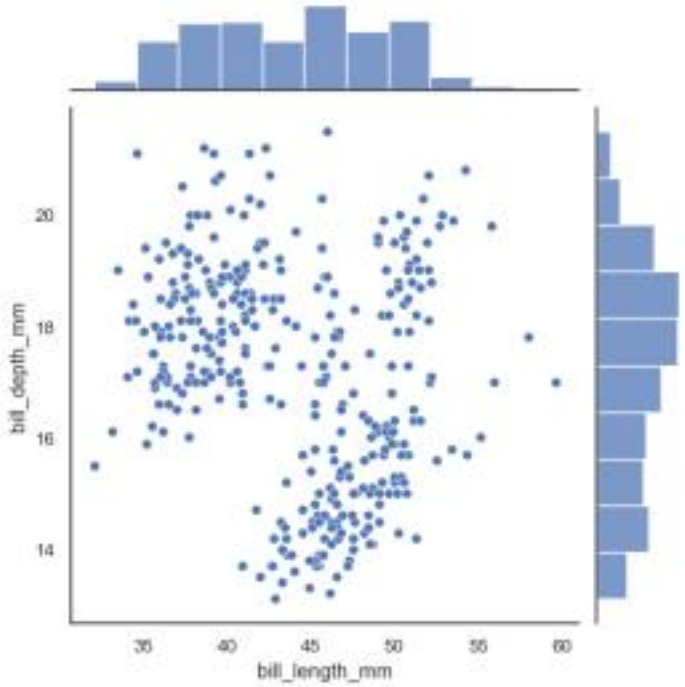
2. Hue Parameter



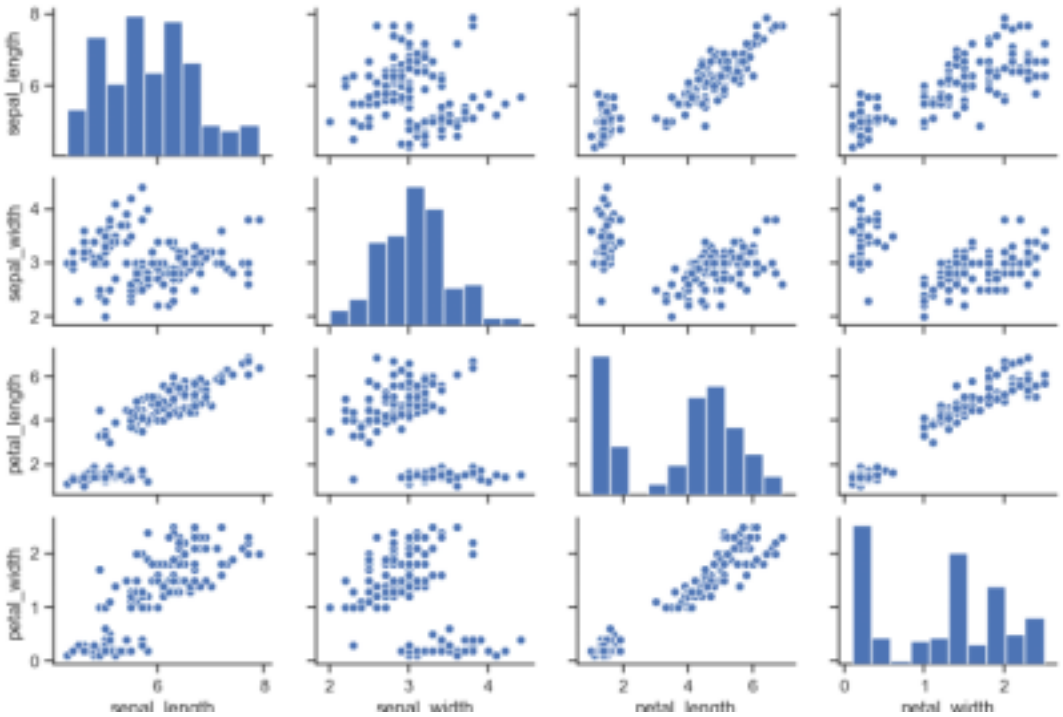
3. Facetgrids



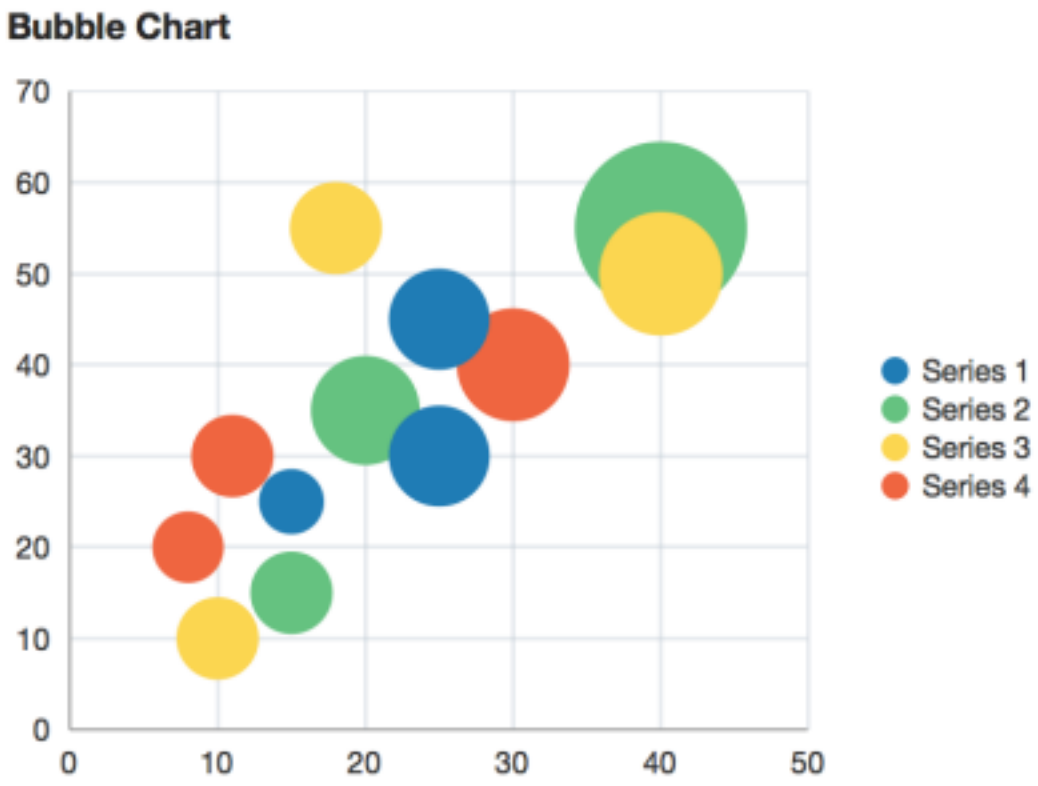
4. Jointplots



5. Pairplots



6. Bubble Plots



Summary of a notes :

1. Quantiles and Percentiles

- **Quantiles** divide data into equal parts to understand its distribution.
    - **Quartiles**: Split data into 4 parts (Q1 = 25%, Q2 = 50% or median, Q3 = 75%).
    - **Deciles**: Split data into 10 parts (e.g., D1 = 10%, D2 = 20%).
    - **Percentiles**: Split data into 100 parts (P1 = 1%, P99 = 99%).
    - **Quintiles**: Split data into 5 parts.
- **Percentiles** show the percentage of data points below a certain value. Example: The 75th percentile means 75% of the data lies below that value.

Tips:

1. Sort data first.
2. Percentiles help derive all other quantiles.
3. Percentiles are not exact data values but derived positions.

---

2. Five-Number Summary

A quick summary of data using:

1. **Minimum**: Smallest value.
2. **Q1 (First Quartile)**: 25% mark.
3. **Median (Q2)**: Middle value (50% mark).
4. **Q3 (Third Quartile)**: 75% mark.
5. **Maximum**: Largest value.

Visualization: Use a **box plot** to display this summary. It shows spread, outliers, and skewness.

---

3. Interquartile Range (IQR)

- Measures the spread of the middle 50% of data.
- **Formula**: IQR =Q3−Q1.
- Helps identify variability and potential outliers.

---

4. Boxplots

A visual tool to summarize data using the five-number summary.
Benefits:

1. Easily view data distribution.
2. Identify outliers.
3. Compare multiple datasets.

---

5. Scatterplots

A graph showing relationships between two variables as points on a 2D grid.

- **X-axis**: Independent variable.
- **Y-axis**: Dependent variable.
- Identifies patterns like positive/negative relationships.

---

6. Covariance

- Measures how two variables change together.
    - **Positive Covariance**: Variables increase together.
    - **Negative Covariance**: One increases, the other decreases.
    - **Zero Covariance**: No linear relationship.
- **Limitation**: Doesn't indicate the strength of the relationship.

---

7. Correlation

- Quantifies the strength and direction of a relationship between variables.
- Measured using the **correlation coefficient**:
    - **+1**: Perfect positive correlation.
    - **0**: No correlation.
    - **-1**: Perfect negative correlation.

Correlation vs. Causation:

- Correlation shows a relationship but not the cause.
    - Example: Firefighters and fire damage are correlated, but firefighters don't cause more damage. The severity of the fire is the actual cause.

---

8. Visualizing Multiple Variables

- **3D Scatter Plots**: Add depth to scatterplots.
- **Hue Parameter**: Differentiates data with color.
- **FacetGrids**: Splits data into grids based on categories.
- **Jointplots**: Combines scatter and distribution plots.
- **Pairplots**: Shows scatterplots for all variable pairs.
- **Bubble Plots**: Uses the size of points to show additional data.

Real-World Example of Covariance

Scenario: Monthly Temperature vs. Ice Cream Sales

- Imagine you're analyzing the relationship between **temperature** and **ice cream sales** over several months.
    - When temperature increases, ice cream sales also tend to increase (positive covariance).
    - Conversely, if temperature decreases, ice cream sales decrease.

Interpretation:

- Covariance tells you whether the two variables move together (positive) or in opposite directions (negative), but it doesn't show the strength of this relationship.

Real-World Example of Correlation

Scenario: Study Hours vs. Exam Scores

- A study finds that students who study more tend to score higher in exams.
    - The **correlation coefficient** for study hours and exam scores is **+0.85** (a strong positive relationship).

Another Example: Height vs. Weight

- Taller people generally weigh more.
    - The correlation coefficient might be around **+0.7**, indicating a moderately strong positive relationship.

Interpretation:

- Correlation provides both the **direction** (positive/negative) and the **strength** of the relationship between variables, unlike covariance.

Key Difference:

- Covariance gives a general sense of the relationship's direction.
- Correlation standardizes the relationship, making it easier to interpret strength and compare across datasets.

Let me know if you need further clarification!

Correlation vs. Causation Example

Scenario: Ice Cream Sales and Shark Attacks

- **Observation**: Data shows a positive correlation between **ice cream sales** and **shark attacks**.
    - When ice cream sales increase, shark attacks also increase.

Does this mean ice cream causes shark attacks?

- **No!** This is a case of correlation, not causation.
- **Reason**: Both are influenced by a third variable: **hot weather**.
    - Hot weather drives more people to the beach (increasing shark attacks) and boosts ice cream sales.

Key Takeaway:

- **Correlation**: Ice cream sales and shark attacks are related.
- **Causation**: Eating ice cream doesn't cause shark attacks; the real factor is hot weather.

Always investigate other possible factors before assuming causation!

Correlation vs. Causation

Correlation

- Describes a statistical relationship between two variables, showing how they move together.
- It **does not imply one variable causes the other** to change.
- Example: More umbrellas sold are correlated with more rain, but umbrellas don’t cause rain.

Causation

- Means that one variable **directly causes a change** in another.

- To prove causation, you often need controlled experiments or strong evidence.
- Example: Increasing water intake leads to better hydration.

Key Difference:

- **Correlation**: Indicates association (variables may move together, but for unknown reasons).
- **Causation**: Proves one variable directly affects the other.

Remember: "Correlation does not imply causation."