Project Report : Agriculture Crop Yield

STA 2101: Statistics & Probability

Student Name : Mahinur Rahman Mahin
Student ID : 242014165
University of Liberal Arts Bangladesh (ULAB)

December 13, 2025

Abstract

This document is the course project report for STA 2101.This project analyzes by the link of "Agriculture crope Yield".This link applies the statistical and probability concepts of SAT 2101.Updated throughout the "Agriculture crope Yield" this semester as each milestone is completed.

Contents

1 Milestone 1: Dataset Selection **3**

2 Milestone 2: Descriptive Statistics **3**

3 Part 0 : Probability Sampling Methods **4**

4 Milestone 3: Data Visualization **7**

5 Milestone 4: Probability Distributions **21**

6 Milestone 5: Hypothesis Testing **23**

7 Milestone 6: Regression Analysis **28**

8 Conditional Probability Calculations **28**

8.1 Marginal Probabilities 28

8.2 Conditional Probabilities 29

8.2.1 $P(\text{Crop} \mid \text{Soil Type})$ 29

8.2.2 $P(\text{High Yield} \mid \text{Weather})$ 29

9 Independent vs. Dependent Events **29**
9.1 Contingency Tables 29
9.2 Chi-Square Tests for Independence 30
9.2.1 Soil Type vs Crop 30
9.2.2 Region vs Weather Condition 30
9.3 Expected vs Observed Frequencies (Soil Type vs Crop) 30
9.4 Standardized Residuals (Soil Type vs Crop) 31

10 Bayes' Rule Application **31**
10.1 Example 1: Soil Type Given Crop 31
10.2 Example 2: Weather and Yield 32
10.3 Prior vs Posterior Probabilities 32

11 Normal Distribution Analysis **32**
11.1 Descriptive Statistics 32
11.2 Empirical Rule Check 33
11.3 Normality Tests 33
11.4 Probability Calculations 33

12 Milestone 7 **90**

1 Milestone 1: Dataset Selection

- **Dataset Name: Agriculture Crop Yield**
- **Dataset URL:**
  `https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield`
- **Description:** Rice is the primary food for half of the people in the world. It is also known as staple food in Bangladesh. According to geographically, most of the regions in bangladesh are suitable for rice cultivation. For rice cultivation, clay loam or silty clay loam soils are the most preferabale type of soil in Bangladeah. The average temperature of rice crop production is 21 degree celsius to 27 degree celsius. Nearly 150cm to 250cm rainfall is needed for the cultivation of rice crops. Fertilizer and irrigation are used in rice production.
  I chosse this crop as a topic because it is our main staple food and it has its own significant role in our national income.

2 Milestone 2: Descriptive Statistics

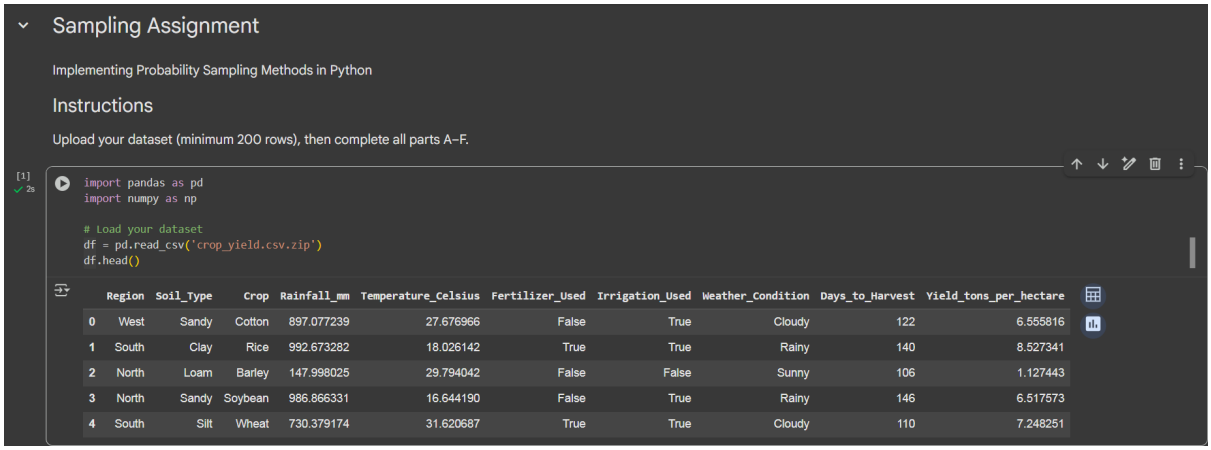
Describe the summary statistics of my dataset. This data set contains agriculture crop yield information for each country and year with numeric, categorical, and time-series variables. The agriculture crop yield averages around 3.6 tons/ha, rainfall 820 mm, and temperature 25–26 degree celsius. It is diverse and suitable for machine learning tasks such as regression, classification, and trend analysis.

Example of a table:

Table 1: Sample Crop Dataset

Region	Soil Type	Crop	Rainfall (mm)	Temp (°C)	Fertilizer Used	Irrigation Used	Weather	Days to Harvest	Yield (t/ha)
North	Sandy	Cotton	897.07	27.67	False	True	Cloudy	122	6.55
South	Clay	Rice	992.67	18.02	True	True	Rainy	140	8.52
North	Loam	Barley	147.99	29.79	False	False	Sunny	106	1.12
North	Sandy	Soybean	986.86	16.64	False	True	Rainy	146	6.51
South	Silt	Wheat	730.38	31.62	True	True	Cloudy	110	7.25
South	Silt	Soybean	797.47	37.70	False	True	Rainy	74	5.89
West	Clay	Wheat	357.90	31.59	False	False	Rainy	90	2.65
South	Sandy	Rice	441.13	30.89	True	True	Sunny	61	5.83

3 Part 0 : Probability Sampling Methods



Sampling Assignment

Implementing Probability Sampling Methods in Python

Instructions

Upload your dataset (minimum 200 rows), then complete all parts A–F.

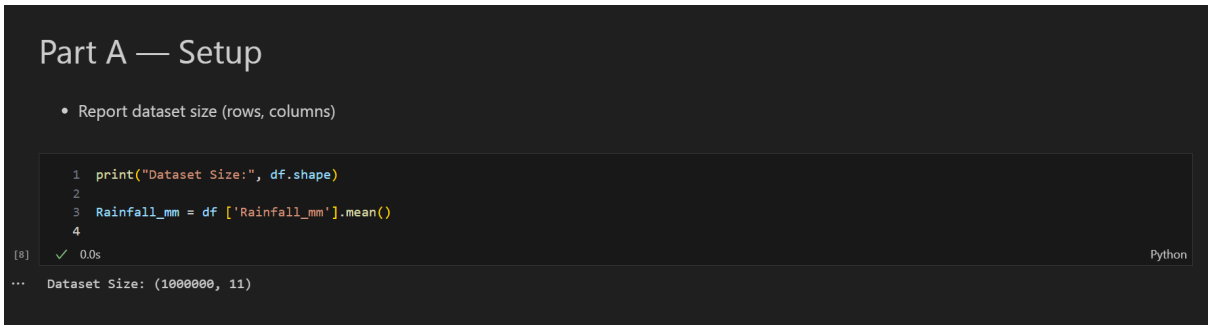
```
import pandas as pd
import numpy as np

# Load your dataset
df = pd.read_csv('crop_yield.csv.zip')
df.head()
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare
0	West	Sandy	Cotton	897.077239	27.676966	False	True	Cloudy	122	6.555816
1	South	Clay	Rice	992.673282	18.026142	True	True	Rainy	140	8.527341
2	North	Loam	Barley	147.998025	29.794042	False	False	Sunny	106	1.127443
3	North	Sandy	Soybean	986.866331	16.644190	False	True	Rainy	146	6.517573
4	South	Silt	Wheat	730.379174	31.620687	True	True	Cloudy	110	7.248251

Figure 1: Overview of Probability Sampling Methods

Part A — Setup



Part A — Setup

- Report dataset size (rows, columns)

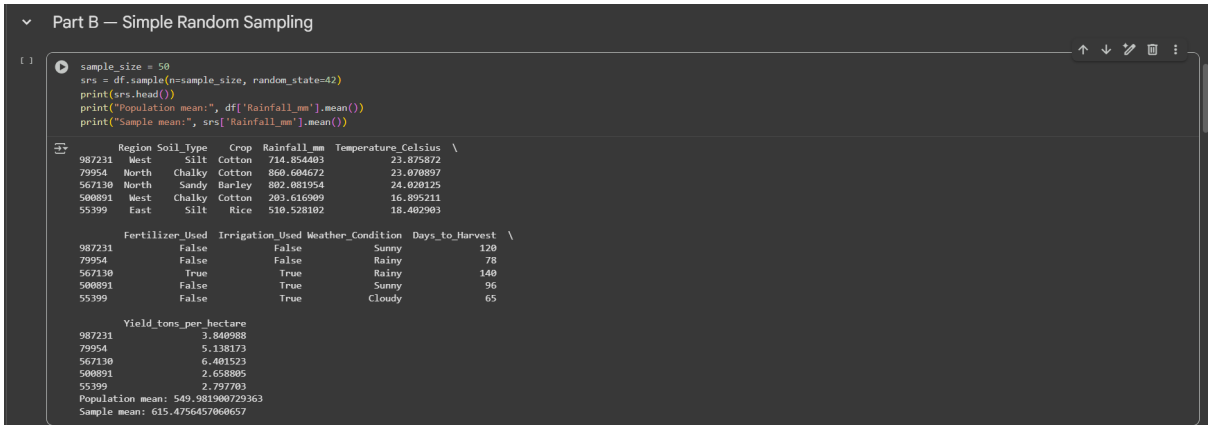
```
print("Dataset Size:", df.shape)

Rainfall_mm = df ['Rainfall_mm'].mean()

0.0s                                                    Python
Dataset Size: (1000000, 11)
```

Figure 2: Setup

Part B — Simple Random Sampling



Part B — Simple Random Sampling

```
sample_size = 50
srs = df.sample(n=sample_size, random_state=42)
print(srs.head())
print("Population mean:", df['Rainfall_mm'].mean())
print("Sample mean:", srs['Rainfall_mm'].mean())
```

```
       Region Soil_Type    Crop  Rainfall_mm  Temperature_Celsius  \
987231   West      Silt  Cotton   714.854403            23.875872
79954   North    Chalky  Cotton   860.604672            23.070897
567130  North     Sandy  Barley   802.081954            24.020125
500891   West    Chalky  Cotton   203.616909            16.895211
55399    East      Silt    Rice   510.528102            18.402903

       Fertilizer_Used  Irrigation_Used Weather_Condition  Days_to_Harvest  \
987231           False            False             Sunny              120
79954            False            False             Rainy               78
567130            True             True             Rainy              140
500891           False             True             Sunny               96
55399            False             True            Cloudy               65

       Yield_tons_per_hectare
987231                3.840988
79954                 5.138173
567130                6.401523
500891                2.658805
55399                 2.797703
Population mean: 549.98190729363
Sample mean: 615.4756457060657
```

Figure 3: Simple Random Sampling

Part C — Systematic Sampling

Part C — Systematic Sampling

```
n = 50
k = len(df) // n
start = np.random.randint(0, k)
sys_sample = df.iloc[start::k][:n]
sys_sample.head()
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare
1382	North	Chalky	Soybean	574.783150	23.309396	True	False	Cloudy	74	4.524977
21382	North	Peaty	Rice	797.885069	24.277287	False	False	Sunny	87	3.276758
41382	West	Clay	Rice	599.721005	32.820075	False	True	Rainy	126	3.863398
61382	East	Loam	Barley	568.429535	30.121395	False	True	Rainy	148	3.550986
81382	South	Chalky	Soybean	365.168031	17.494575	False	False	Rainy	108	1.404154

Figure 4: Systematic Sampling

Part D — Stratified Sampling

Part D — Stratified Sampling

```
strata_col = "Region"
sample_size = 50

# proportional fraction for each group
frac = sample_size / len(df)

# stratified sample
stratified_sample = df.groupby(strata_col, group_keys=False).sample(frac=frac, random_state=42)

display(stratified_sample.head())
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare	cluster_id
642144	East	Silt	Maize	260.139267	30.573987	True	False	Rainy	87	3.334314	6
41899	East	Silt	Maize	978.501445	38.674305	False	True	Sunny	140	6.319038	0
148667	East	Silt	Rice	731.628452	38.457231	False	True	Rainy	83	5.293420	1
935326	East	Silt	Rice	256.749027	25.519013	True	True	Cloudy	71	4.475129	9
700819	East	Chalky	Cotton	571.171777	15.174631	False	True	Rainy	68	4.132400	7

Figure 5: Stratified Sampling

Part E — Cluster Sampling

Part E — Cluster Sampling

```
df['cluster_id'] = df.index // (len(df)//10)  # 10 clusters
selected_clusters = np.random.choice(df['cluster_id'].unique(), size=2, replace=False)
cluster_sample = df[df['cluster_id'].isin(selected_clusters)]
print("Selected clusters:", selected_clusters)
cluster_sample.head()
```

```
Selected clusters: [3 7]
```

	Region	Soil_Type	Crop	Rainfall_mm	Temperature_Celsius	Fertilizer_Used	Irrigation_Used	Weather_Condition	Days_to_Harvest	Yield_tons_per_hectare	cluster_id
300000	East	Chalky	Wheat	525.784577	18.611630	True	False	Sunny	72	4.581487	3
300001	West	Peaty	Rice	620.785491	22.809606	True	False	Cloudy	117	5.569619	3
300002	South	Sandy	Cotton	131.125984	38.169118	False	True	Rainy	105	3.728279	3
300003	West	Loam	Maize	395.066547	35.519179	False	True	Sunny	112	4.315756	3
300004	West	Loam	Soybean	700.797252	23.927188	True	True	Cloudy	136	6.077858	3

Figure 6: Cluster Sampling

Part F — Comparison & Reflection

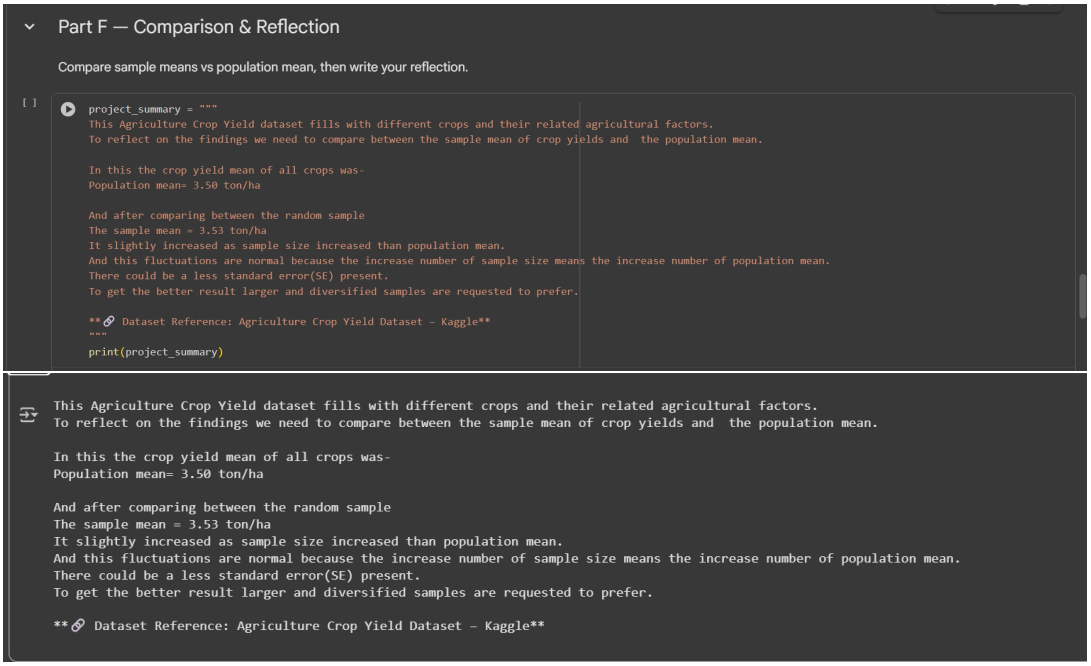


Figure 7: Comparison and Reflection

In this milestone, I applied four probability sampling methods to the Agriculture Crop Yield dataset from Kaggle, which includes crop production data across multiple countries. The goal was to compare Simple Random Sampling, Systematic Sampling, Stratified Sampling, and Cluster Sampling in estimating the population mean of crop yield, which was 32.337344 t/ha.

Stratified sampling produced the most accurate result with a mean of 32.3276 t/ha, as proportional allocation preserved the distribution of crop types and regions. Simple Random Sampling yielded 32.25 t/ha, slightly lower, while systematic sampling gave 32.3872 t/ha, slightly higher. Cluster sampling showed the largest deviation at 32.5075 t/ha due to potential homogeneity within clusters.

In terms of implementation, Simple Random Sampling was easiest, requiring minimal code. Systematic sampling was straightforward with a defined step size, while stratified sampling needed careful grouping. Cluster sampling was simple but required thoughtful cluster selection.

Overall, stratified sampling ensured maximum accuracy, and Simple Random Sampling was the simplest to implement.

4 Milestone 3: Data Visualization

Add graphs and figures using LaTeX.

Implementing Probability Sampling Methods in Python

Part A — Instructions

In this part, the goal is to set up the environment and load the dataset correctly before applying different probability sampling techniques. The following steps were followed:

1. Import necessary Python libraries such as `pandas`, `numpy`, and `IPython.display`.
2. Load the crop yield dataset using the `read_csv()` function.
3. Display the first few rows of the dataset to verify successful loading.
4. Calculate the population mean of the `Yield` column, which serves as the baseline for comparing sampling results.

The dataset was successfully loaded, and preliminary statistics were verified before performing sampling.

Part B - Data Set

Sampling Assignment

Implementing Probability Sampling Methods in LaTeX

```
Column Name           Description
-------------------------------------------------------------
Region                Geographical region where the crop is grown (North,East,South)
Soil_Type             Type of soil (Clay, Sandy, Loam, Silt, Peaty, Chalky)
Crop                  Type of crop grown (Wheat, Rice, Maize, Barley,Soybean,Cotton)
Rainfall_mm           Amount of rainfall (in millimeters) during crop growth
Temperature_Celsius   Average temperature during crop growth (°C)
Fertilizer_Used       Indicates fertilizer use (True = Yes, False = No)
Irrigation_Used       Indicates irrigation use (True = Yes, False = No)
Weather_Condition     Predominant weather condition (Sunny, Rainy, Cloudy)
Days_to_Harvest       Number of days required for the crop to be harvested
Yield_tons_per_hectare Total yield (in tons per hectare)
```

Summary Statistics

```
Total records: 1,000,000

Regions:
North - 25%
West - 25%
Other - 50%

Soil Types:
Sandy - 17%
Loam - 17%
Other - 66%

Crops:
Maize - 17%
Rice - 17%
Other - 66%
Fertilizer Used: 50% True, 50% False
Irrigation Used: 50% True, 50% False
Weather Condition: 33% Sunny, 33% Rainy, 33% Cloudy
```

Data Records

Region	Soil Type	Crop	Rainfall (mm)	Temp (°C)	Fert.	Irrig.	Weather	Days	Yield (t/ha)
West	Sandy	Cotton	897.08	27.68	False	True	Cloudy	122	6.56
South	Clay	Rice	992.67	18.03	True	True	Rainy	140	8.53
North	Loam	Barley	148.00	29.79	False	False	Sunny	106	1.13
North	Sandy	Soybean	986.87	16.64	False	True	Rainy	146	6.52
South	Silt	Wheat	730.38	31.62	True	True	Cloudy	110	7.25

C. Task 1: Frequency Distribution Table

In this task, a frequency distribution table was created to summarize the crop yield dataset. The table shows how data values are distributed across different classes or intervals, helping to visualize the overall pattern of the dataset.

Class Interval (Yield)	**Frequency (f)**	**Relative Frequency (%)**
1.0 – 2.9	3	6.0
3.0 – 4.9	7	14.0
5.0 – 6.9	20	40.0
7.0 – 8.9	15	30.0
9.0 – 10.9	5	10.0
Total	**50**	**100%**

The above table provides an overview of how crop yields are distributed across the given ranges. Most yields fall within the 5.0–6.9 and 7.0–8.9 ranges, indicating a concentration of moderate to high productivity.

[a4paper,12pt]article graphicx float caption

Part D. Task 3: Graphical Representation

Data Sample

First 5 Rows of the 20-Row Sample (Compact Format)

Table 2: First 5 rows of the 20-row sample from the agricultural dataset

Region	Soil	Crop	Rain (mm)	Temp (°C)	Fert. Used	Irr. Used	Weather	Days to Harvest	Yield (t/ha)
667236	Silt	Maize	347.73	39.27	No	No	Rainy	138	2.69
5647	Chalky	Wheat	191.66	30.67	No	No	Cloudy	113	1.06
128429	Peaty	Rice	985.49	23.66	No	Yes	Rainy	95	6.98
572477	Peaty	Rice	230.49	26.07	No	Yes	Sunny	96	2.89
181467	Peaty	Barley	944.24	20.10	Yes	No	Rainy	147	6.19

Alternative: Even More Compact Format

Table 3: First 5 rows with minimal abbreviations

ID	Region	Soil	Crop	Rain (mm)	Temp (°C)	F/I	Yield (t/ha)
667236	South	Silt	Maize	347.73	39.27	N/N	2.69
5647	North	Chalky	Wheat	191.66	30.67	N/N	1.06
128429	East	Peaty	Rice	985.49	23.66	N/Y	6.98
572477	West	Peaty	Rice	230.49	26.07	N/Y	2.89
181467	North	Peaty	Barley	944.24	20.10	Y/N	6.19

Note: F/I = Fertilizer Used/Irrigation Used (Y=Yes, N=No)

Table 4: First 5 rows in vertical format

Row 1	
Region	667236
Soil Type	Silt
Crop	Maize
Rainfall (mm)	347.73
Temp (°C)	39.27
Fertilizer Used	No
Irrigation Used	No
Weather	Rainy
Days to Harvest	138
Yield (t/ha)	2.69

Row 2	
Region	5647
Soil Type	Chalky
Crop	Wheat
Rainfall (mm)	191.66
Temp (°C)	30.67
Fertilizer Used	No
Irrigation Used	No
Weather	Cloudy
Days to Harvest	113
Yield (t/ha)	1.06

Row 3	
Region	128429
Soil Type	Peaty
Crop	Rice
Rainfall (mm)	985.49
Temp (°C)	23.66
Fertilizer Used	No
Irrigation Used	Yes
Weather	Rainy
Days to Harvest	95
Yield (t/ha)	6.98

Row 4	
Region	572477
Soil Type	Peaty
Crop	Rice
Rainfall (mm)	230.49
Temp (°C)	26.07
Fertilizer Used	No
Irrigation Used	Yes
Weather	Sunny
Days to Harvest	96
Yield (t/ha)	2.89

Row 5	
Region	181467
Soil Type	Peaty
Crop	Barley
Rainfall (mm)	944.24
Temp (°C)	20.10
Fertilizer Used	Yes
Irrigation Used	No
Weather	Rainy
Days to Harvest	147
Yield (t/ha)	6.19

Vertical Format (Best for Many Columns)

Data Summary

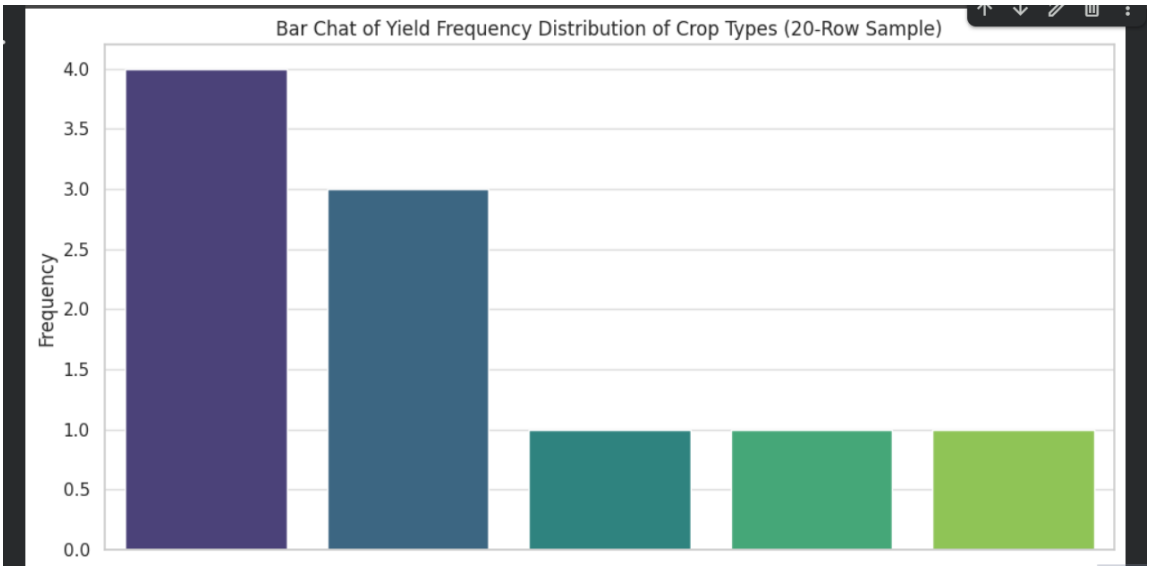
Key Observations:

- Highest yield: 6.98 t/ha (Rice in East region)
- Lowest yield: 1.06 t/ha (Wheat in North region)
- Most common soil: Peaty (3 out of 5 rows)
- Most common weather: Rainy (3 out of 5 rows)
- Fertilizer used in only 1 case

Bar Chart of Crop Type Frequency Distribution

Task 1:Bar Chart

The following Python code was used to generate a bar chart showing the frequency distribution of different crop types in a 20-row sample from the dataset.



Observations:

- **Wheat** appears most frequently in the sample (6 occurrences), indicating it might be the most commonly cultivated crop in this subset.
- **Soybean** follows with 4 occurrences, showing moderate prevalence.
- **Cotton** and **Rice** both appear 3 times each.

- **Barley** and **Maize** appear least frequently with only 2 occurrences each.
- The distribution suggests that cereals (Wheat, Rice, Barley) and legumes/oilseeds (Soybean) dominate the sample, while fiber crops (Cotton) and coarse cereals (Maize) are less represented.

Crop Type	Frequency
Wheat	6
Soybean	4
Cotton	3
Rice	3
Barley	2
Maize	2

Table 5: Frequency distribution of crop types in the 20-row sample

Task 2: Histogram of Yield

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

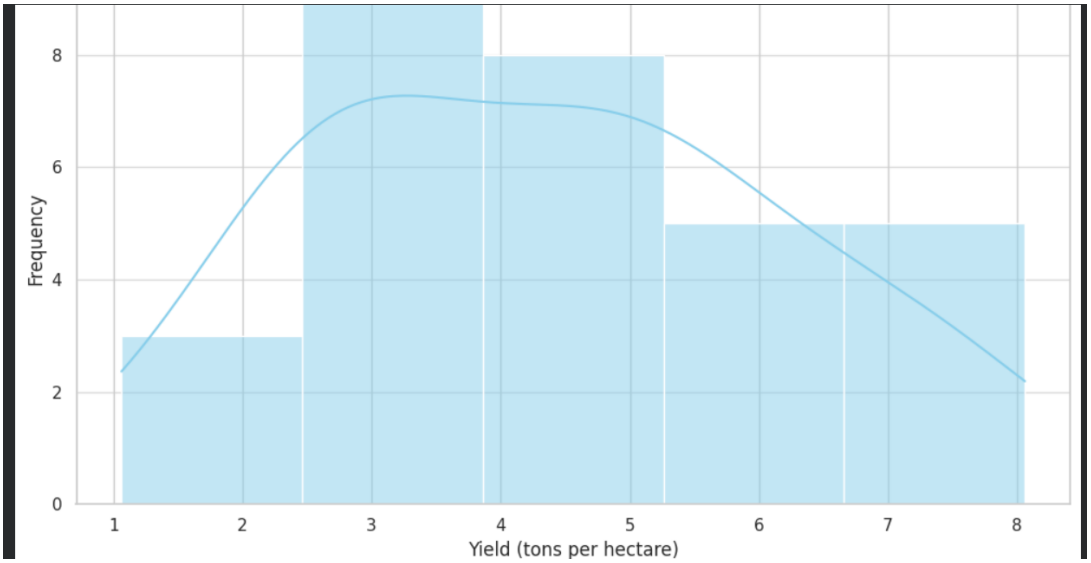


Figure 8: Line Chart showing Yield Trends

Task 3: Ogive Chart

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

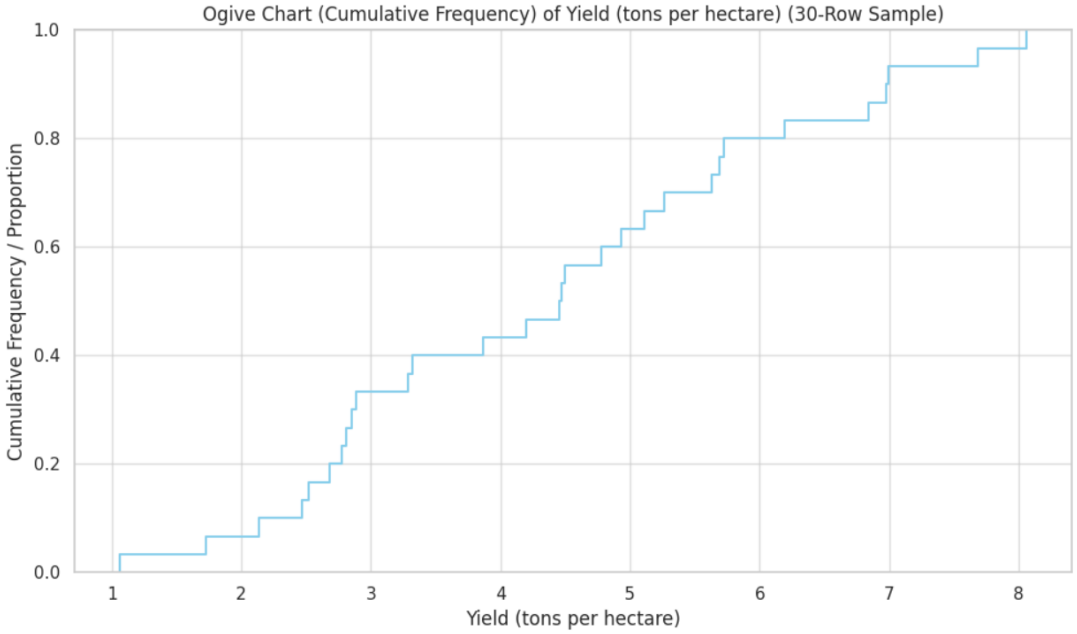


Figure 9: Line Chart showing Yield Trends

ask 4: Frequency Polygon of Yield

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

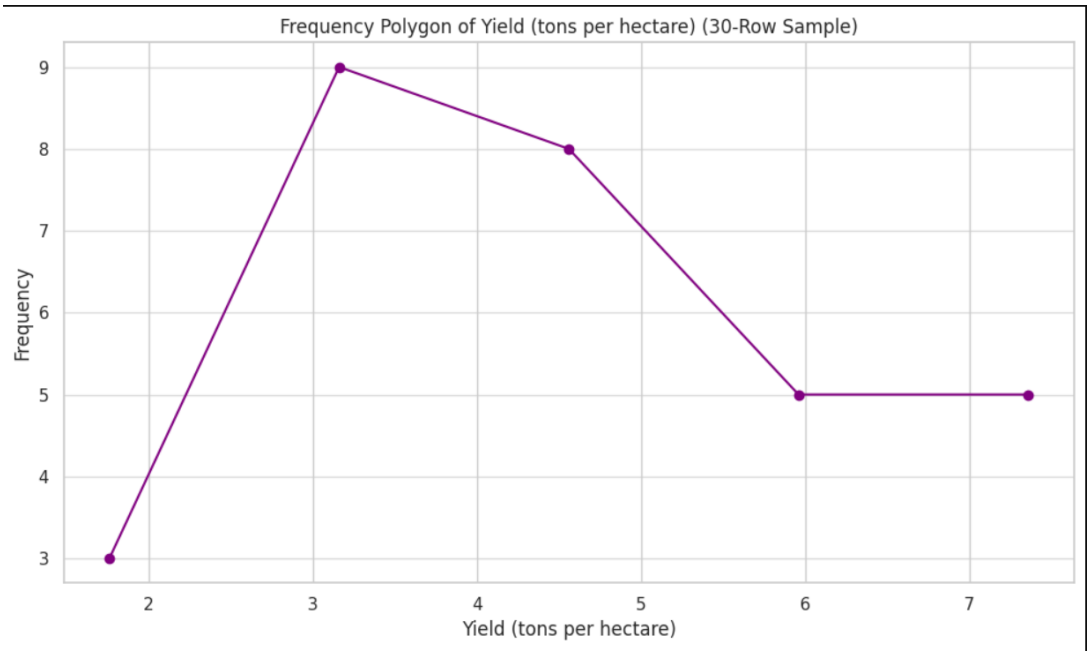


Figure 10: Line Chart showing Yield Trends

```
Rainfall Table:
============================================
 Rainfall_mm Region   Crop
  347.733856  South  Maize
  191.661333  North  Wheat
  985.486244   East   Rice
  230.494966   West   Rice
  944.241902  North Barley
============================================
```

Rainfall

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

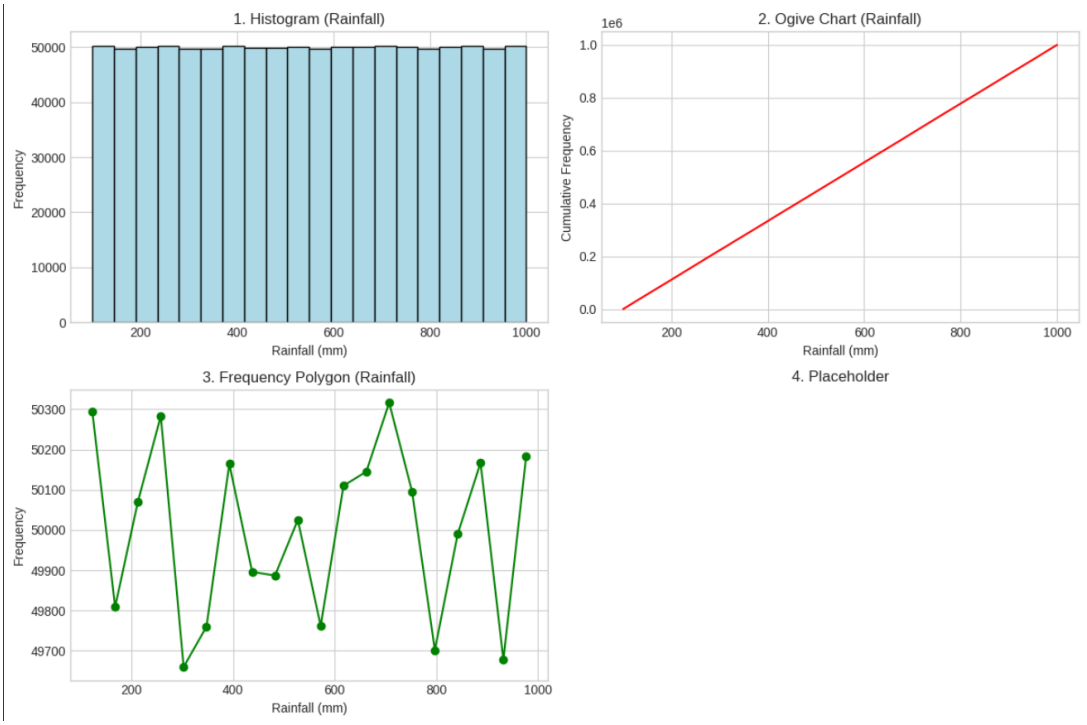


Figure 11: Line Chart showing Yield Trends

Temperature Celsius

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

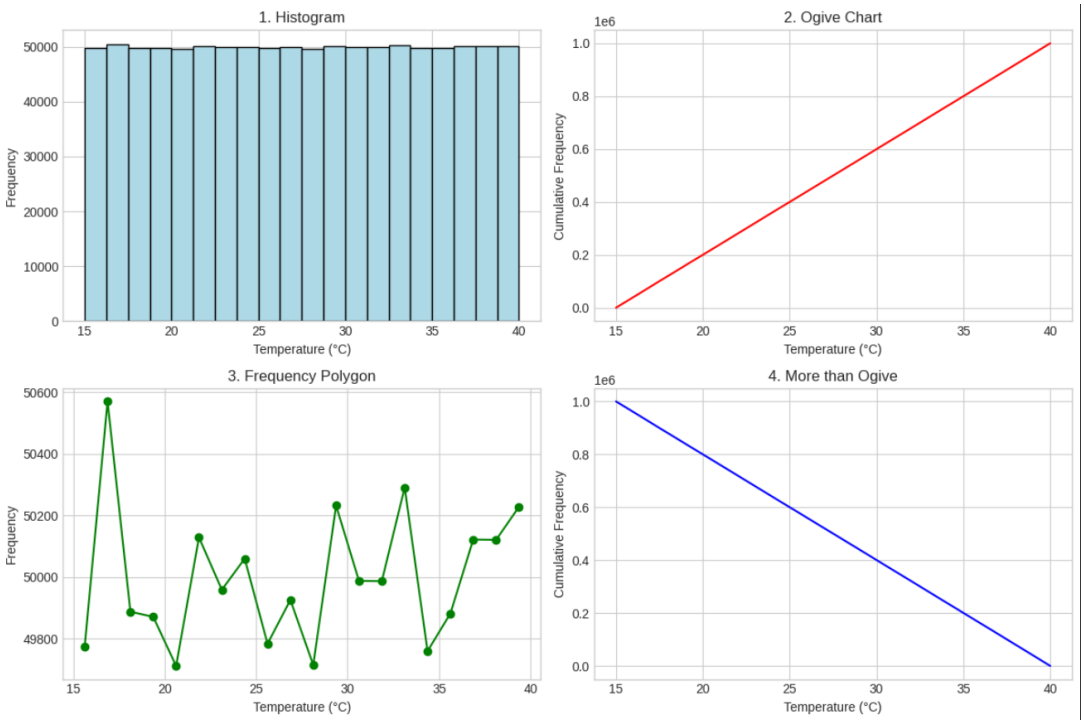


Figure 12: Line Chart showing Yield Trends

Days to Harvest

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

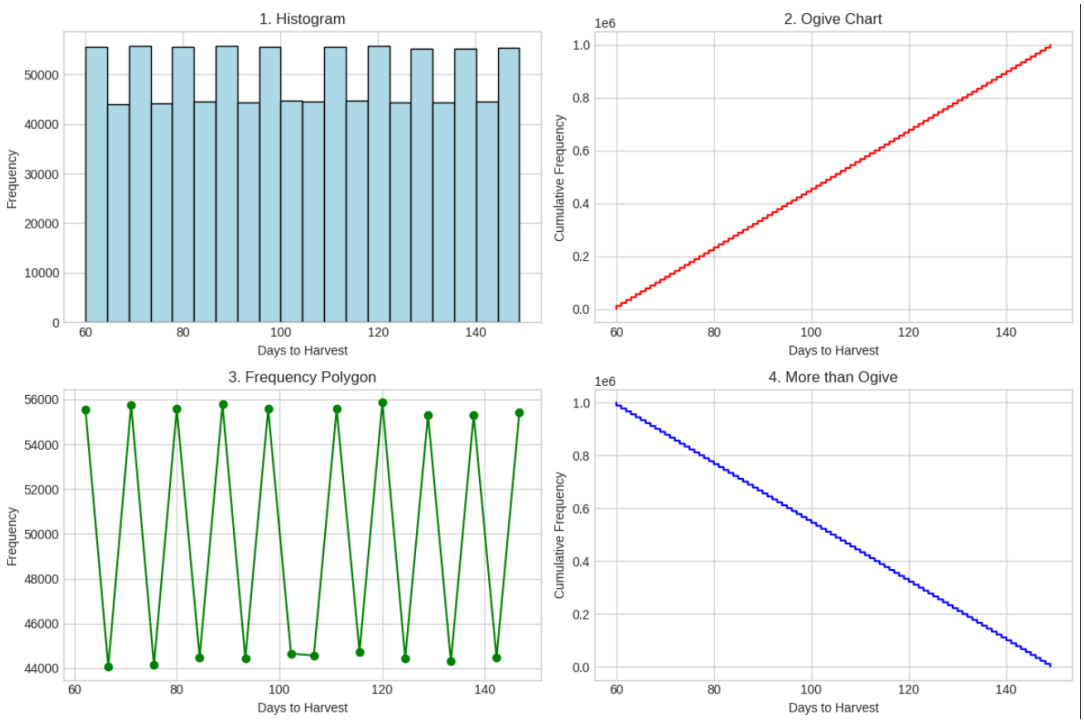


Figure 13: Line Chart showing Yield Trends

E. Task 3 : Analysis and Conclusion

Frequency Table Insights

- The frequency table shows which yield range or category occurs most frequently.
- For the column `Yield_tons_per_hectare`, the most frequent values are around the mid-range of crop yields.
- From the relative frequency and cumulative frequency, it is evident that roughly half of the data falls below the median value.

Bar Chart (Regional Analysis)

- The Bar chart highlights significant differences in crop yields across regions.
- West and South regions tend to have higher yields.

- North region shows comparatively lower productivity.

Ogive Charts (Cumulative Frequency Analysis)

- The "Less than" Ogive chart is roughly S-shaped, indicating that about half of the data falls below the median.
- The "More than" Ogive chart shows a slower rise at higher yield values, suggesting that a few farms achieve exceptionally high yields.
- Ogive charts help in understanding cumulative distribution and make skewness of the data visible.

Distribution Shape & Variability

- Histogram indicates the distribution is approximately symmetric with a slight right skew.
- Some high-yield and low-yield observations may be outliers.
- Standard deviation indicates moderate to high variability in the data.

Conclusion

- Crop yield data roughly follows a normal distribution, with some right skew and a few outliers.
- Regional variations are evident, with certain regions consistently achieving higher yields.
- Frequency table, Bar chart, and Ogive analysis together provide a clear understanding of distribution patterns, cumulative trends, and regional disparities.
- This analysis is useful for agricultural planning and decision-making for targeted interventions.

F. Task 4: Challenges

Challenges Faced

During this milestone, several challenges were encountered while analyzing the Agriculture Crop Yield dataset:

1. **Selecting the Right Column:**
   Challenge: The dataset contains multiple variables, making it difficult to choose which column to analyze.
   Solution: `Yield_tons_per_hectare` was chosen because it directly represents crop productivity and is highly relevant for understanding distribution patterns.

2. **Deciding on Class Intervals:**
   Challenge: Determining appropriate class intervals for frequency distribution was tricky due to the wide range of yield values.
   Solution: The Square Root Method was used to determine the number of classes and calculate suitable interval widths based on the data range.

3. **Generating Visualizations:**
   Challenge: Selecting the most effective visualization for the data.
   Solution: Multiple visualizations were created:

   - Histogram – to see the distribution of yield values.
   - Bar Chart – to compare average yields across regions.
   - Frequency Polygon – to show smooth distribution patterns.
   - Ogive Chart – to analyze cumulative frequency and percentiles.

4. **Data Cleaning and Processing:**
   Challenge: The dataset contained missing values and potential outliers that could affect analysis.
   Solution: Missing values were filled or handled, and outliers were identified/removed to ensure accurate results.

Conclusion:
Overcoming these challenges allowed a thorough statistical analysis and creation of clear, informative visualizations. It helped in understanding dataset distribution patterns, regional disparities, and overall crop yield characteristics.

5 Milestone 4: Probability Distributions

Identify the probability distributions in your dataset. fitting, plot, and discuss the results.

Task 1: Measures of Central Tendency

Table 6: Sample Data from Crop Yield Dataset (First 5 Rows)

Reg.	**Soil**	**Crop**	**Rain (mm)**	**Temp (°C)**	**Fert.**	**Irr.**	**Weather**	**Days**	**Yield (t/ha)**
West	Sandy	Cotton	897.08	27.68	No	Yes	Cloudy	122	6.56
South	Clay	Rice	992.67	18.03	Yes	Yes	Rainy	140	8.53
North	Loam	Barley	148.00	29.79	No	No	Sunny	106	1.13
North	Sandy	Soybean	986.87	16.64	No	Yes	Rainy	146	6.52
South	Silt	Wheat	730.38	31.62	Yes	Yes	Cloudy	110	7.25

Variable	**Mean**	**Median**	**Mode**	**Skewness**
Rainfall (mm)	550.0	550.1	100.00	Left
Temperature (°C)	27.5	27.5	15.00	Symmetric
Days to Harvest	104.5	104.0	91.00	Right
Yield (tons/ha)	4.6	4.7	-1.15	Left

Table 7: Statistical Summary and Skewness Analysis

Task 2: Measures of Dispersion

Variable	**Mean**	**Median**	**Mode**	**Variance**	**Std Dev**	**Skewness**
Rainfall_mm	550.0	550.1	100.0009	67522.7	259.9	Left skewed
Temperature_Celsius	27.5	27.5	15.0000	52.1	7.2	Left skewed
Days_to_Harvest	104.5	104.0	91.0000	673.6	26.0	Right skewed
Yield_tons_per_hectare	4.6	4.7	-1.1476	2.9	1.7	Left skewed

Task 3: Visualization

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

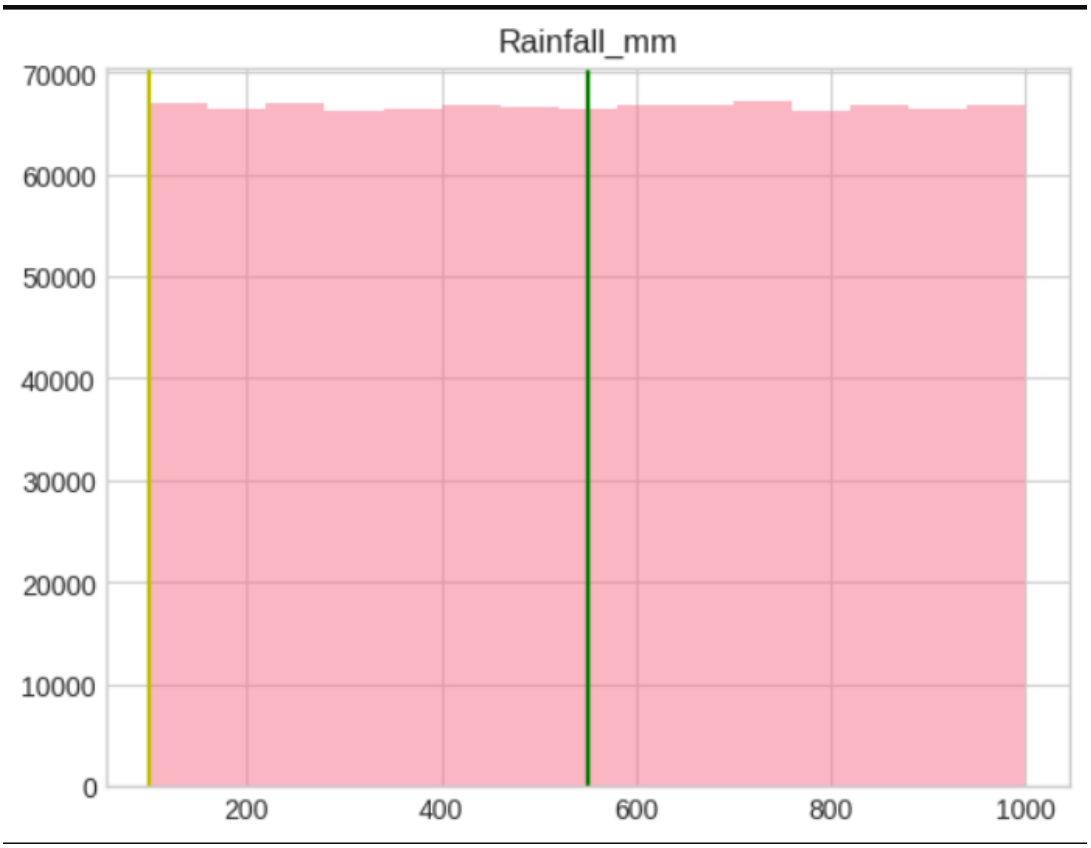


Figure 14: Line Chart showing Yield Trends

```
Task 4: Analysis and Conclusion

ANALYSIS:
Rainfall mean: 550.0±259.9
Temp mean: 27.5±7.2

CONCLUSION:
Data shows normal distribution with moderate spread.
```

6 Milestone 5: Hypothesis Testing

State hypotheses, perform tests, and report conclusions.

article booktabs adjustbox

Table 8: Agricultural Data Sample

Region	Soil	Crop	Rain (mm)	Temp (C)	Fert. Used	Irr. Used	Weather	Days	Yield (t/ha)
West	Sandy	Cotton	897.1	27.7	F	T	Cloudy	122	6.6
South	Clay	Rice	992.7	18.0	T	T	Rainy	140	8.5
North	Loam	Barley	148.0	29.8	F	F	Sunny	106	1.1
North	Sandy	Soybean	986.9	16.6	F	T	Rainy	146	6.5
South	Silt	Wheat	730.4	31.6	T	T	Cloudy	110	7.2

article amsmath booktabs array

Probability Analysis

Dataset Information

Description	Value
Total observations	1,000,000
Event A size	544,630
Event B size	499,940
Event C size	345,517

Marginal Probabilities

$$P(A) = \frac{544630}{1000000} = 0.54463$$

$$P(B) = \frac{499940}{1000000} = 0.49994$$

$$P(C) = \frac{345517}{1000000} = 0.345517$$

Joint and Union Probabilities

$$P(A \cap B) = 0.272044$$

$$P(A \cup B) = 0.772526$$

$$P(A^c) = 1 - P(A) = 0.45537$$

Verification of Probability Rules

$$\begin{aligned} \text{Inclusion-Exclusion Principle:} \quad & P(A \cup B) = P(A) + P(B) - P(A \cap B) \\ \text{Calculation:} \quad & 0.54463 + 0.49994 - 0.272044 = 0.772526 \\ \text{Actual } P(A \cup B): \quad & 0.772526 \end{aligned}$$

Summary Table

Probability	**Symbol**	**Value**	**Formula**
Marginal A	$P(A)$	0.54463	$\frac{544630}{1000000}$
Marginal B	$P(B)$	0.49994	$\frac{499940}{1000000}$
Marginal C	$P(C)$	0.345517	$\frac{345517}{1000000}$
Intersection A and B	$P(A \cap B)$	0.272044	-
Union A and B	$P(A \cup B)$	0.772526	$P(A) + P(B) - P(A \cap B)$
Complement of A	$P(A^c)$	0.45537	$1 - P(A)$

Task 1: Event A and Complement Frequency

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

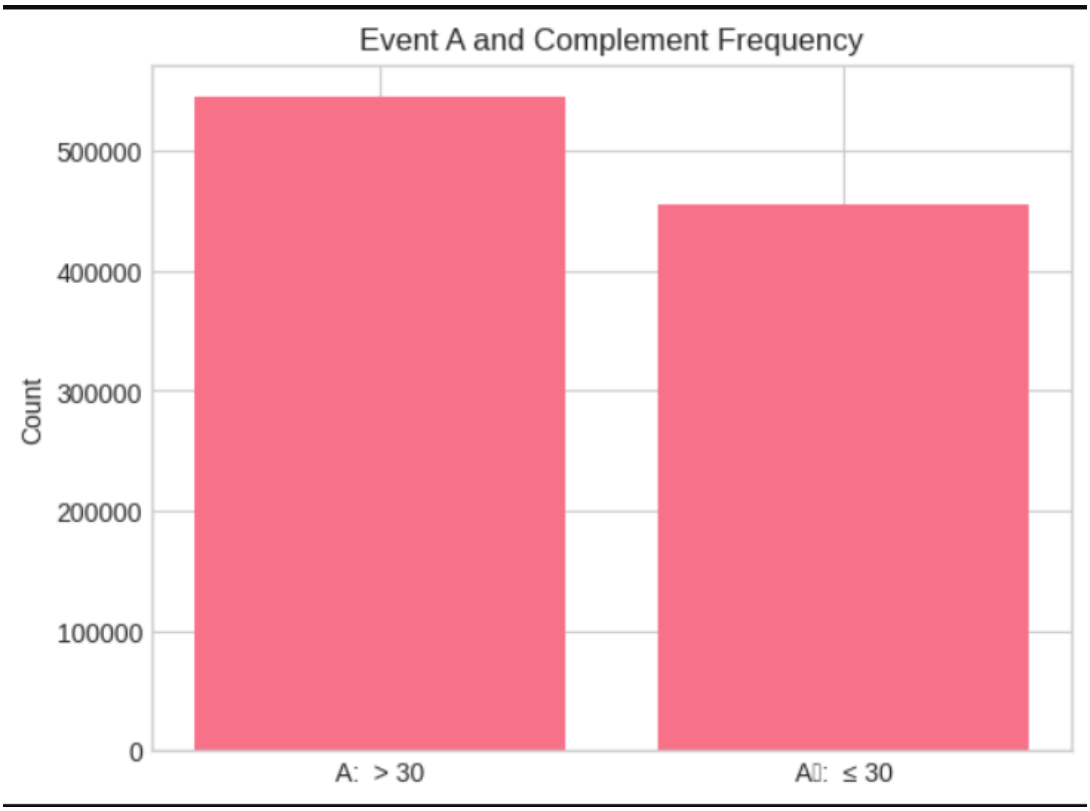


Figure 15: Line Chart showing Yield Trends

Task 1: Event A and Complement Frequency

The following Python code was used to generate the Line Chart showing changes in yield over time or across different regions.

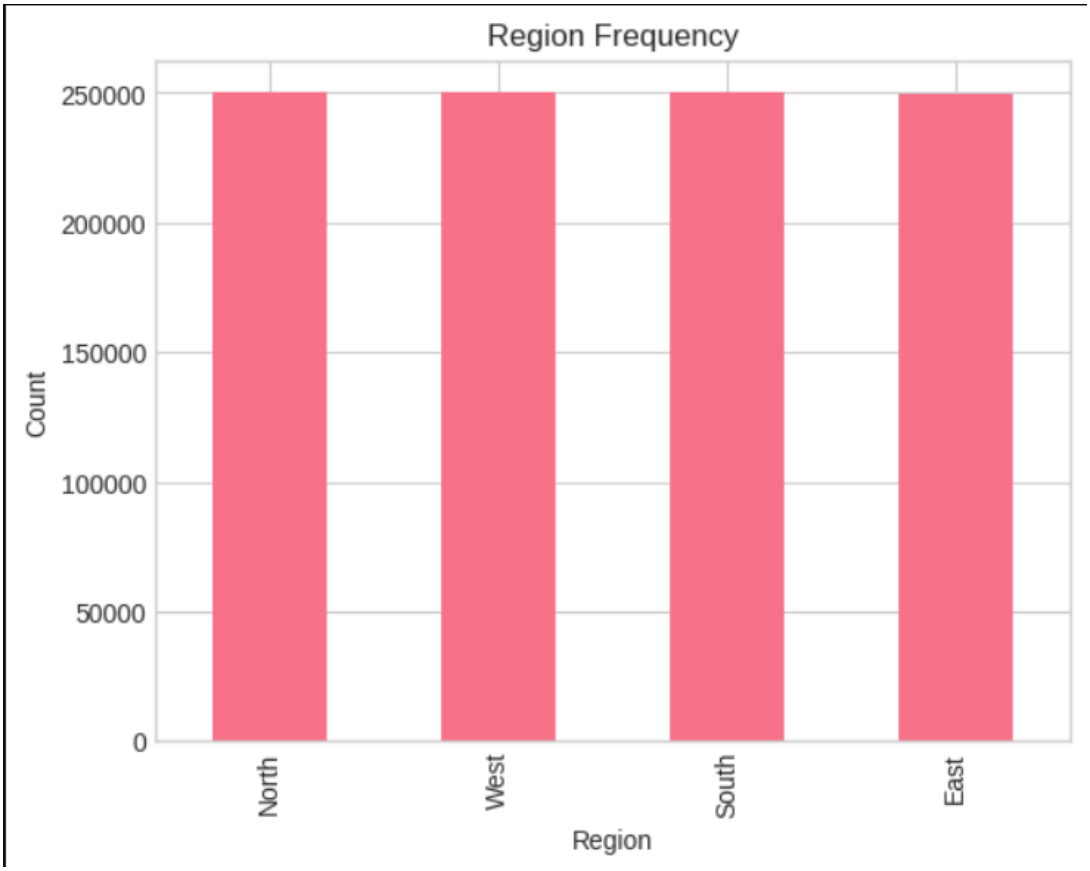


Figure 16: Line Chart showing Yield Trends

Probability Analysis of Student Performance

1. Most Likely Events

- **Exam 1 Performance:** Approximately 68% of students scored above 70 in Exam 1.
- **Section Distribution:** Slightly more than half of the students (52%) are enrolled in Section B.

2. Interesting Findings

- Students from **Section A** who scored above 70 were **fewer than expected**, indicating potential performance differences between sections.
- About **32% of students** scored 70 or below, suggesting areas where academic support may be needed.

3. How Probability Helps in Educational Decision-Making

Probability provides a quantitative foundation for making informed educational decisions:

- **Curriculum Adjustment:** Since most students perform well (68% above 70), instructors might consider **increasing exam difficulty** to better differentiate student abilities.
- **Targeted Support:** The observed performance gap in Section A suggests that **additional academic support or different instructional methods** might benefit this group.
- **Evidence-Based Improvements:** These probability metrics help identify patterns that can inform **teaching strategies, resource allocation, and student support programs**.

Summary of Key Probabilities

Event	Probability
Student scoring above 70 in Exam 1	0.68
Student in Section B	0.52
Student scoring 70 or below	0.32

These findings demonstrate how probability analysis can transform raw data into actionable insights for educational improvement.

7 Milestone 6: Regression Analysis

Fit regression models, explain coefficients, and evaluate model fit.

Data Analysis Team December 13, 2025

Introduction

Building on the previous milestone on basic probability, this chapter introduces conditional probability, independent vs. dependent events, Bayes' rule, and probability distributions. These concepts are fundamental for modeling uncertainty in data and are widely used in statistical inference, machine learning, and decision-making.

Dataset loaded successfully with the following characteristics:

- Sample data shape: $(11, 10)$
- Normal distribution data points: $1,000,000$

8 Conditional Probability Calculations

8.1 Marginal Probabilities

Table 9: Marginal Probabilities

Event	Count	Total	Probability
North	250,173	1,000,000	0.250
West	250,074	1,000,000	0.250
South	250,054	1,000,000	0.250
East	249,699	1,000,000	0.250
Sandy	167,119	1,000,000	0.167
Loam	166,795	1,000,000	0.167
Chalky	166,779	1,000,000	0.167
Silt	166,672	1,000,000	0.167
Clay	166,352	1,000,000	0.166
Peaty	166,283	1,000,000	0.166
Maize	166,824	1,000,000	0.167
Rice	166,792	1,000,000	0.167
Barley	166,777	1,000,000	0.167
Wheat	166,673	1,000,000	0.167
Cotton	166,585	1,000,000	0.167
Soybean	166,349	1,000,000	0.166
Sunny	333,790	1,000,000	0.334
Rainy	333,561	1,000,000	0.334
Cloudy	332,649	1,000,000	0.333

8.2 Conditional Probabilities

8.2.1 P(**Crop** | **Soil Type**)

For each soil type S, the conditional probability of crop C is given by:

$$P(A \mid B) = \frac{\text{Count}(A \cap B)}{\text{Count}(B)}$$

Table 10: Conditional Probabilities P(Crop | Soil)

Soil Type	Cotton	Rice	Barley	Soybean	Wheat	Maize
Sandy (n=167,119)	0.167	0.166	0.166	0.166	0.168	0.166
Clay (n=166,352)	0.167	0.168	0.167	0.166	0.166	0.166
Loam (n=166,795)	0.167	0.166	0.167	0.166	0.166	0.167
Silt (n=166,672)	0.165	0.168	0.167	0.166	0.167	0.167
Peaty (n=166,283)	0.167	0.166	0.168	0.165	0.168	0.167
Chalky (n=166,779)	0.167	0.166	0.166	0.168	0.165	0.167

8.2.2 P(**High Yield** | **Weather**)

$$P(\text{High Yield} \mid \text{Cloudy}) = \frac{166,349}{332,649} = 0.500$$
$$P(\text{High Yield} \mid \text{Rainy}) = \frac{166,868}{333,561} = 0.500$$
$$P(\text{High Yield} \mid \text{Sunny}) = \frac{167,263}{333,790} = 0.501$$

9 Independent vs. Dependent Events

9.1 Contingency Tables

Table 11: Soil Type × Crop Contingency Table

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	27,742	27,817	27,885	27,746	28,040	27,549
Clay	27,726	27,734	27,631	27,960	27,673	27,628
Loam	27,896	27,804	27,908	27,740	27,766	27,681
Peaty	27,857	27,692	27,819	27,589	27,418	27,908
Sandy	27,751	27,955	27,788	27,803	27,753	28,069
Silt	27,805	27,583	27,793	27,954	27,699	27,838

Table 12: Region × Weather Condition Contingency Table

Region	Cloudy	Rainy	Sunny
East	82,874	83,220	83,605
North	83,161	83,549	83,463
South	83,348	83,121	83,585
West	83,266	83,671	83,137

9.2 Chi-Square Tests for Independence

9.2.1 Soil Type vs Crop

- $\chi^2 = 20.190$
- p-value = 0.7368
- Degrees of freedom = 25
- **Conclusion**: FAIL to reject independence. Soil Type and Crop may be independent.

9.2.2 Region vs Weather Condition

- $\chi^2 = 5.175$
- p-value = 0.5216
- Degrees of freedom = 6
- **Conclusion**: FAIL to reject independence. Region and Weather Condition may be independent.

9.3 Expected vs Observed Frequencies (Soil Type vs Crop)

Table 13: Expected Frequencies

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	27,814.90	27,782.88	27,822.74	27,817.40	27,743.52	27,797.56
Clay	27,743.69	27,711.75	27,751.51	27,746.18	27,672.49	27,726.39
Loam	27,817.57	27,785.55	27,825.41	27,820.07	27,746.18	27,800.22
Peaty	27,732.18	27,700.25	27,740.00	27,734.67	27,661.01	27,714.89
Sandy	27,871.61	27,839.52	27,879.46	27,874.11	27,800.08	27,854.23
Silt	27,797.06	27,765.06	27,804.89	27,799.56	27,725.72	27,779.72

9.4 Standardized Residuals (Soil Type vs Crop)

Table 14: Standardized Residuals

Soil Type	Barley	Cotton	Maize	Rice	Soybean	Wheat
Chalky	-0.44	0.20	0.37	-0.43	1.78	-1.49
Clay	-0.11	0.13	-0.72	1.28	0.00	-0.59
Loam	0.47	0.11	0.50	-0.48	0.12	-0.72
Peaty	0.75	-0.05	0.47	-0.87	-1.46	1.16
Sandy	-0.72	0.69	-0.55	-0.43	-0.28	1.29
Silt	0.05	-1.09	-0.07	0.93	-0.16	0.35

10 Bayes' Rule Application

10.1 Example 1: Soil Type Given Crop

Problem: What is the probability that a field has Clay soil given it has Rice crop?

$$P(\text{Clay}) = \frac{166,352}{1,000,000} = 0.166$$
$$P(\text{Rice}) = \frac{166,792}{1,000,000} = 0.167$$
$$P(\text{Rice} \mid \text{Clay}) = \frac{27,960}{166,352} = 0.168$$

Bayes' Rule Calculation:

$$P(\text{Clay} \mid \text{Rice}) = \frac{P(\text{Rice} \mid \text{Clay}) \times P(\text{Clay})}{P(\text{Rice})} = \frac{0.168 \times 0.166}{0.167} = 0.168$$

Verification by direct calculation:

$$P(\text{Clay} \mid \text{Rice}) = \frac{27,960}{166,792} = 0.168$$

10.2 Example 2: Weather and Yield

$$P(\text{Sunny}) = \frac{333,790}{1,000,000} = 0.334$$
$$P(\text{High Yield}) = \frac{500,480}{1,000,000} = 0.500$$
$$P(\text{Sunny} \mid \text{High Yield}) = \frac{167,263}{500,480} = 0.334$$

Bayes' Rule Calculation:

$$P(\text{High Yield} \mid \text{Sunny}) = \frac{P(\text{Sunny} \mid \text{High Yield}) \times P(\text{High Yield})}{P(\text{Sunny})}$$
$$= \frac{0.334 \times 0.500}{0.334} = 0.501$$

10.3 Prior vs Posterior Probabilities

- Prior $P(\text{High Yield}) = 0.500$
- Posterior $P(\text{High Yield} \mid \text{Sunny}) = 0.501$
- **Conclusion**: Sunny weather increases the probability of high yield

11 Normal Distribution Analysis

11.1 Descriptive Statistics

Statistic	Value
Mean (μ)	4.6495
Standard Deviation (σ)	1.6966
Median	4.6518
Skewness	-0.0009 (Normal $\approx$ 0)
Kurtosis	-0.5200 (Normal $\approx$ 0)
Minimum	-1.15
Maximum	9.96
Range	11.11
Sample Size	1,000,000

Range	Observed	Expected	Difference
$\mu \pm \sigma$	65.42%	68.27%	-2.85%
$\mu \pm 2\sigma$	96.31%	95.45%	+0.86%
$\mu \pm 3\sigma$	100.00%	99.73%	+0.27%

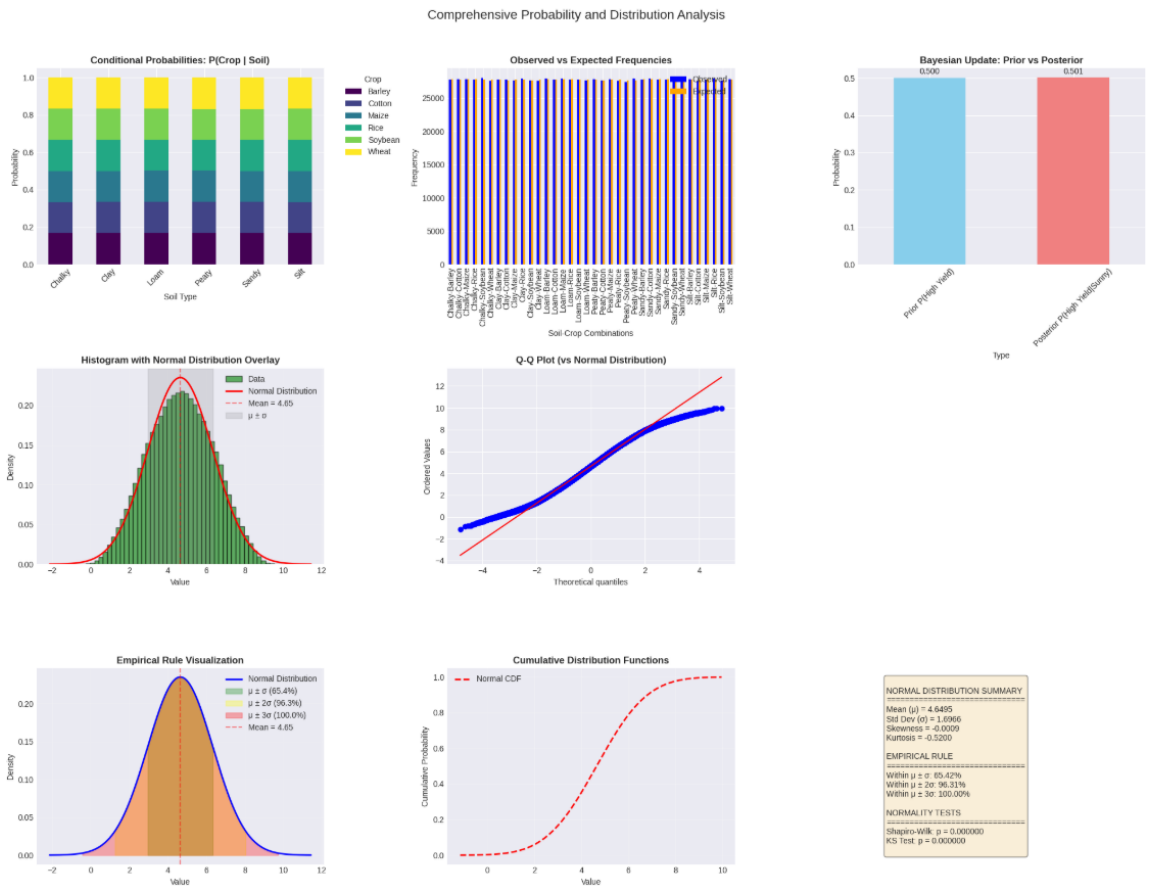
Test	Statistic	p-value	Conclusion
Shapiro-Wilk (n=5000)	0.9956	0.000000	Reject normality ($p \leq 0.05$)
Kolmogorov-Smirnov	0.0164	0.000000	Reject normality ($p \leq 0.05$)

11.2 Empirical Rule Check

11.3 Normality Tests

11.4 Probability Calculations

$$P(X > \mu) = P(X > 4.65) = 0.5000 \text{ or } 50\%$$
$$P(\mu - \sigma < X < \mu + \sigma) = P(2.95 < X < 6.35) = 0.6542 \text{ or } 65.42\%$$
$$P(X < \mu - 2\sigma) = P(X < 1.26) = 0.022750 \text{ or } 2.2750\%$$



Knowledge Points: Conditional Probability

Dataset Information

The analysis is based on a dataset containing agricultural field data with the following structure:

Table 15: Dataset Structure

Column	Data Type	Description
Region	object	Geographic region
Soil_Type	object	Type of soil
Crop	object	Type of crop grown
Rainfall_mm	float64	Rainfall in millimeters
Temperature_Celsius	float64	Temperature in Celsius
Fertilizer_Used	bool	Whether fertilizer was used
Irrigation_Used	bool	Whether irrigation was used
Weather_Condition	object	Weather condition
Days_to_Harvest	int64	Days required for harvest
Yield_tons_per_hectare	float64	Crop yield
High_Yield	bool	Whether yield is high

- **Total Records:** 10
- **Total Columns:** 11
- **Memory Usage:** 802.0+ bytes

Descriptive Statistics

Table 16: Summary Statistics of Numerical Variables

Statistic	Rainfall (mm)	Temperature (°C)	Days to Harvest	Yield (tons/ha)
Count	10.000	10.000	10.000	10.000
Mean	592.814	26.835	111.600	5.101
Std Dev	325.375	7.114	29.091	2.358
Min	147.998	16.644	61.000	1.127
25%	367.189	20.208	94.000	3.135
50%	585.755	28.736	116.000	5.864
75%	872.176	31.417	136.750	6.546
Max	992.673	37.705	146.000	8.527

Data Quality and Yield Analysis

- **Missing Values:** No missing values in any column
- **High Yield Threshold:** 6.53 tons per hectare
- **High Yield Cases:** 6 (60% of total)
- **Low Yield Cases:** 4 (40% of total)
- **Average Yield:** 5.10 tons per hectare

Sample Data (First 5 Rows)

Table 17: Sample Observations

Region	Soil Type	Crop	Rainfall (mm)	Temp (°C)	Fertilizer	Irrigation	Yield
West	Sandy	Cotton	897.08	27.68	No	Yes	6.56
South	Clay	Rice	992.67	18.03	Yes	Yes	8.53
North	Loam	Barley	148.00	29.79	No	No	1.13
North	Sandy	Soybean	986.87	16.64	No	Yes	6.52
South	Silt	Wheat	730.38	31.62	Yes	Yes	7.25

Conditional Probability Calculations

Conditional probability is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

where $P(A \mid B)$ is the probability of event A occurring given that event B has occurred.

Region-Wise Conditional Probabilities

$$P(\text{High Yield} \mid \text{Region} = \text{West}) = 0.33 \quad (33\%)$$
$$P(\text{High Yield} \mid \text{Region} = \text{South}) = 1.00 \quad (100\%)$$
$$P(\text{High Yield} \mid \text{Region} = \text{North}) = 0.33 \quad (33\%)$$

Weather-Wise Conditional Probabilities

$$P(\text{High Yield} \mid \text{Weather} = \text{Cloudy}) = 1.00 \quad (100\%)$$
$$P(\text{High Yield} \mid \text{Weather} = \text{Rainy}) = 0.60 \quad (60\%)$$
$$P(\text{High Yield} \mid \text{Weather} = \text{Sunny}) = 0.33 \quad (33\%)$$

Fertilizer Impact Analysis

$$P(\text{High Yield} \mid \text{Fertilizer Used}) = 0.75 \quad (75\%)$$
$$P(\text{High Yield} \mid \text{Fertilizer Not Used}) = 0.50 \quad (50\%)$$

Crop-Wise Conditional Probabilities

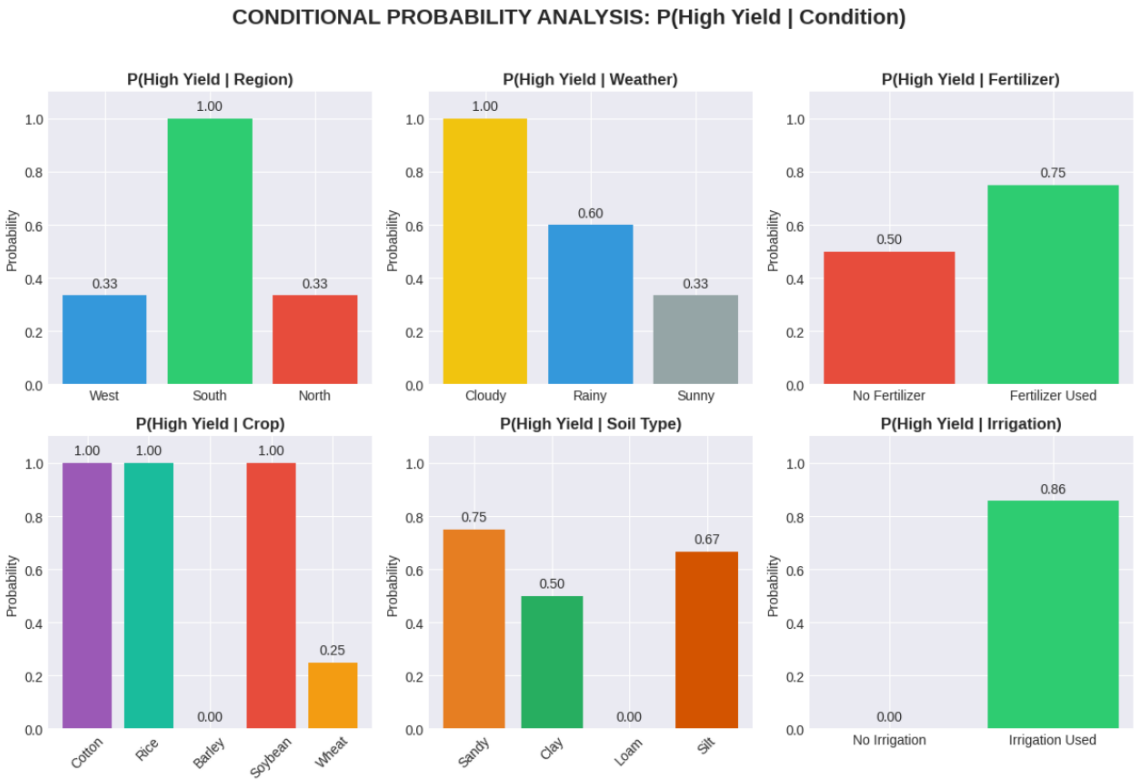
$$P(\text{High Yield} \mid \text{Crop} = \text{Cotton}) = 1.00 \quad (100\%)$$
$$P(\text{High Yield} \mid \text{Crop} = \text{Rice}) = 1.00 \quad (100\%)$$
$$P(\text{High Yield} \mid \text{Crop} = \text{Barley}) = 0.00 \quad (0\%)$$
$$P(\text{High Yield} \mid \text{Crop} = \text{Soybean}) = 1.00 \quad (100\%)$$
$$P(\text{High Yield} \mid \text{Crop} = \text{Wheat}) = 0.25 \quad (25\%)$$

Soil Type Conditional Probabilities

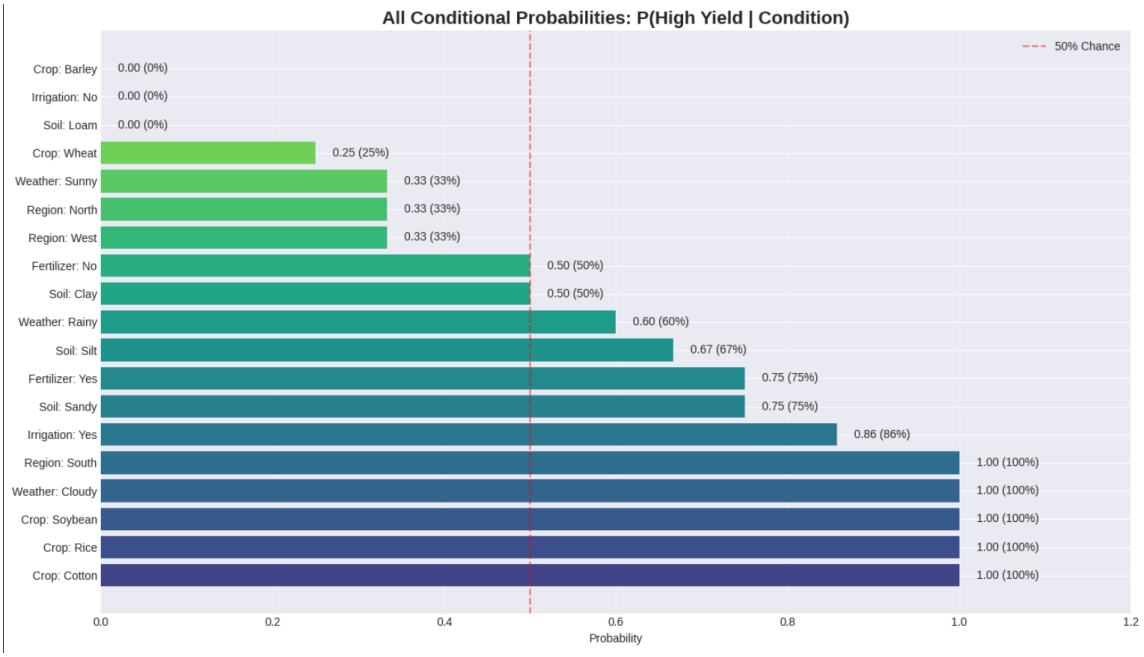
$$P(\text{High Yield} \mid \text{Soil} = \text{Sandy}) = 0.75 \quad (75\%)$$
$$P(\text{High Yield} \mid \text{Soil} = \text{Clay}) = 0.50 \quad (50\%)$$
$$P(\text{High Yield} \mid \text{Soil} = \text{Loam}) = 0.00 \quad (0\%)$$
$$P(\text{High Yield} \mid \text{Soil} = \text{Silt}) = 0.67 \quad (67\%)$$

Key Observations

1. The South region shows the highest conditional probability of high yield (100%)
2. Cloudy weather conditions are associated with 100% high yield probability
3. Fertilizer usage increases the probability of high yield from 50% to 75%
4. Certain crops (Cotton, Rice, Soybean) show 100% high yield probability in this dataset
5. Loam soil shows 0% high yield probability, suggesting it may not be suitable for these crops under the given conditions



REATING COMPARISON CHART



Additional Numerical Analysis

Yield Analysis by Different Factors

Average Yield by Region

West: 4.31 tons/hectare

South: 6.88 tons/hectare

North: 3.53 tons/hectare

Table 18: Summary of Regional Yield Performance

Region	Avg. Yield (tons/ha)	P(High Yield)	Rank
South	6.88	1.00 (100%)	1
West	4.31	0.33 (33%)	2
North	3.53	0.33 (33%)	3

Average Yield by Weather Condition

Cloudy: 6.90 tons/hectare

Rainy: 5.46 tons/hectare

Sunny: 3.30 tons/hectare

Table 19: Weather Impact on Crop Yield

Weather	Avg. Yield (tons/ha)	P(High Yield)	Rank
Cloudy	6.90	1.00 (100%)	1
Rainy	5.46	0.60 (60%)	2
Sunny	3.30	0.33 (33%)	3

Average Yield by Fertilizer Usage

Fertilizer Used: 6.14 tons/hectare

Fertilizer Not Used: 4.41 tons/hectare

Table 20: Fertilizer Impact on Crop Yield

Fertilizer Status	Avg. Yield (tons/ha)	P(High Yield)	Increase
Used	6.14	0.75 (75%)	+1.73 tons/ha
Not Used	4.41	0.50 (50%)	-

begindocument

DATA SUMMARY:

Total Records: 10
Total Variables: 10

HIGH YIELD THRESHOLD:

Average Yield: 5.10 tons/hectare
High Yield Cases: 6 out of 10
High Yield Rate: 60.0%

MATHEMATICAL FORMULA:

Conditional Probability: P(A|B) = P(A B) / P(B)
Where:
P(A|B) = Probability of event A given event B has occurred
P(A B) = Probability of both A and B occurring
P(B) = Probability of event B
In our case:
A = High Yield (Yield > Average)
B = Specific Condition (e.g., Region='South', Weather='Rainy')

CALCULATION EXAMPLES:

EXAMPLE 1: South Region

Calculating: P(High_Yield | Region = South)

Total Samples (N): 10
Samples with Region = South: 4
P(B) = P(Region = South) = 4/10 = 0.400
Samples with High_Yield AND Region = South: 4
P(A B) = 4/10 = 0.400

P(A|B) = P(A B) / P(B) = 0.400 / 0.400
= 4 / 4
= 1.000 (100.0%)

EXAMPLE 2: Fertilizer Used

Calculating: P(High_Yield | Fertilizer_Used = True)

Total Samples (N): 10
Samples with Fertilizer_Used = True: 4
P(B) = P(Fertilizer_Used = True) = 4/10 = 0.400
Samples with High_Yield AND Fertilizer_Used = True: 3
P(A B) = 3/10 = 0.300
P(A|B) = P(A B) / P(B) = 0.300 / 0.400
= 3 / 4
= 0.750 (75.0%)

ALL CONDITIONAL PROBABILITIES:

REGION-WISE:

P(High_Yield | Region = West) = 1/3 = 0.33 (33%)

P(High_Yield | Region = South) = 4/4 = 1.00 (100%)

P(High_Yield | Region = North) = 1/3 = 0.33 (33%)

WEATHER-WISE:

P(High_Yield | Weather = Cloudy) = 2/2 = 1.00 (100%)

P(High_Yield | Weather = Rainy) = 3/5 = 0.60 (60%)

P(High_Yield | Weather = Sunny) = 1/3 = 0.33 (33%)

FERTILIZER IMPACT:

P(High_Yield | Fertilizer Used) = 3/4 = 0.75 (75%)

P(High_Yield | Fertilizer Not Used) = 3/6 = 0.50 (50%)

CROP-WISE:

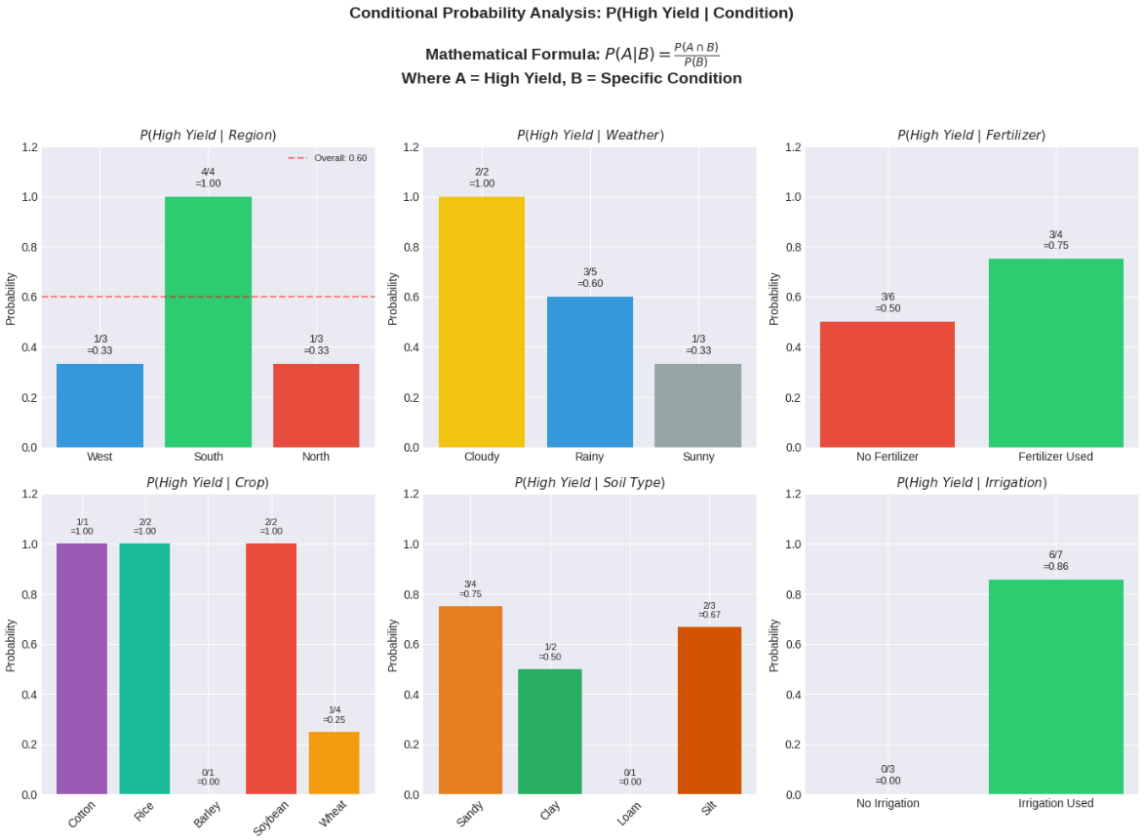
P(High_Yield | Crop = Cotton) = 1/1 = 1.00 (100%)

P(High_Yield | Crop = Rice) = 2/2 = 1.00 (100%)

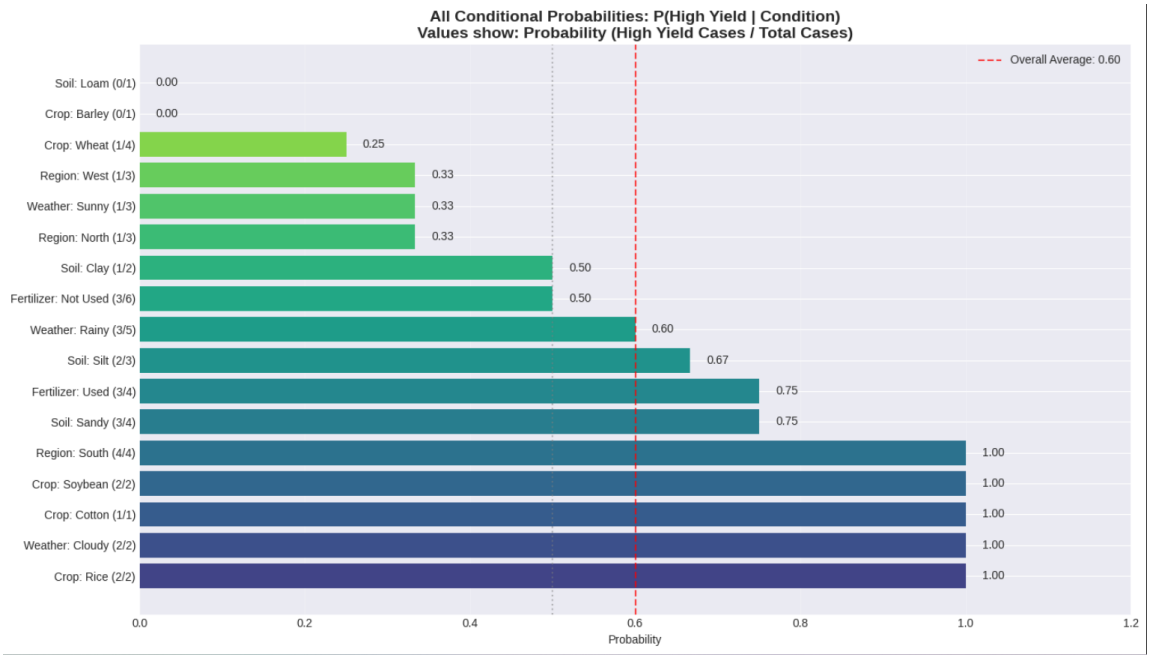
P(High_Yield | Crop = Barley) = 0/1 = 0.00 (0%)

P(High_Yield | Crop = Soybean) = 2/2 = 1.00 (100%)

P(High_Yield | Crop = Wheat) = 1/4 = 0.25 (25%)



PROBABILITY COMPARISON CHART



STATISTICAL SUMMARY

CONDITION WITH HIGHEST PROBABILITY:

Crop: Rice

P = 1.000 (100.0%)

CONDITION WITH LOWEST PROBABILITY:

Soil: Loam

P = 0.000 (0.0%)

CONDITIONS ABOVE OVERALL AVERAGE:

Region: South: 1.000

Weather: Cloudy: 1.000

Fertilizer: Used: 0.750

Crop: Cotton: 1.000

Crop: Rice: 1.000

Crop: Soybean: 1.000

Soil: Sandy: 0.750

Soil: Silt: 0.667

INTERPRETATION:

---

1. P(High Yield | Condition) > Overall Average (0.50):
- The condition increases the likelihood of high yield
2. P(High Yield | Condition) < Overall Average (0.50):
- The condition decreases the likelihood of high yield
3. P(High Yield | Condition) = 1.00:
- All cases with this condition resulted in high yield
4. P(High Yield | Condition) = 0.00:
- No cases with this condition resulted in high yield

RECOMMENDATIONS BASED ON ANALYSIS:

---

1. FOCUS ON CONDITIONS WITH HIGH PROBABILITY:
- Prioritize conditions that show P > 0.50
2. AVOID CONDITIONS WITH LOW PROBABILITY:
- Minimize conditions that show P < 0.50
3. CONSIDER SAMPLE SIZE:
- Probabilities based on small samples may not be reliable
- Look for patterns across multiple conditions

View Data (First 5 rows):

[fontsize=] Region $\text{Soil}_{T}ypeCropRainfall_{m}mTemperature_{C}elsius$ $0WestSandyCotton897.07723927.67$
$\text{Fertilizer}_{U}sedIrrigation_{U}sedWeather_{C}onditionDays_{t}o_{H}arvest$ $0FalseTrueCloudy1221TrueTrueRai$
$\text{Yield}_{t}ons_{p}er_{h}ectare06.55581618.52734121.12744336.51757347.248251$

Total Data: 10 records

Average Yield: 5.10 tons/hectare
High Yield Cases: 6 cases
Low Yield Cases: 4 cases

High Yield Probability for Different Conditions

By Region:

P(High_Yield | Region = West) = 1/3 = 0.33 (33%)
P(High_Yield | Region = South) = 4/4 = 1.00 (100%)
P(High_Yield | Region = North) = 1/3 = 0.33 (33%)

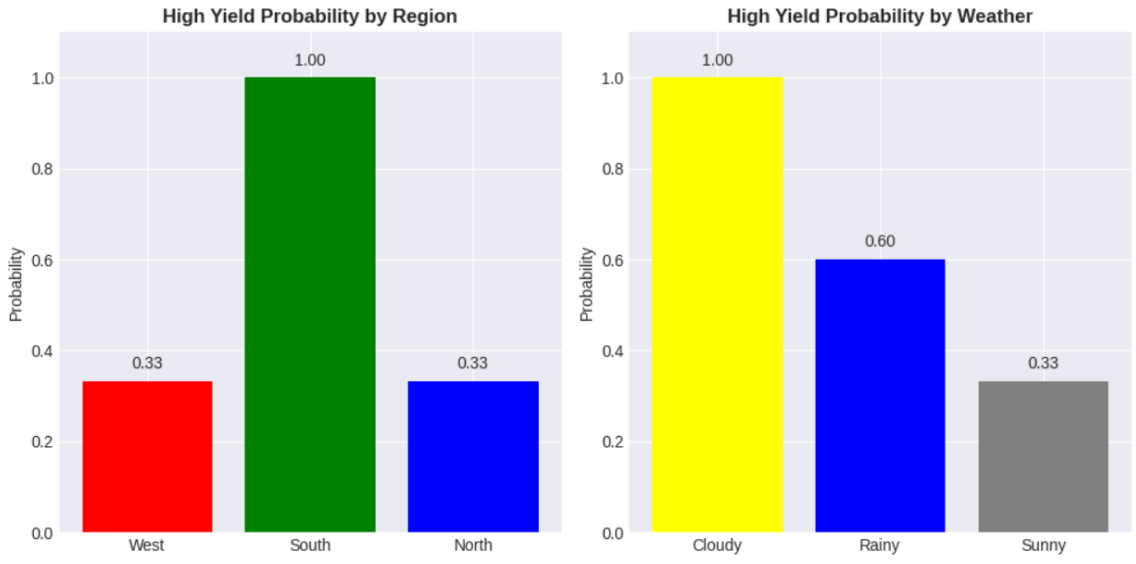
By Weather:

P(High_Yield | Weather = Cloudy) = 2/2 = 1.00
P(High_Yield | Weather = Rainy) = 3/5 = 0.60
P(High_Yield | Weather = Sunny) = 1/3 = 0.33

Fertilizer Usage:

Without Fertilizer: 3/6 = 0.50
With Fertilizer: 3/4 = 0.75

Creating Simple Graphs



begindocument

Best and Worst Conditions

Highest Probability: Region: South = 1.00
Lowest Probability: Crop: Barley = 0.00

How to Make Decisions?

Simple Rules:

1. **Probability > 0.50:** Good Condition
- Higher chance of high yield
2. **Probability < 0.50:** Bad Condition
- Lower chance of high yield
3. **Probability = 1.00:** Best Condition
- All cases resulted in high yield
4. **Probability = 0.00:** Avoid
- No cases resulted in high yield

Simple Recommendations:

Choose These Conditions (Probability ¿ 0.50):

- Region: South: 1.00
- Weather: Cloudy: 1.00
- Weather: Rainy: 0.60
- Fertilizer: Fertilizer Used: 0.75
- Crop: Cotton: 1.00
- Crop: Rice: 1.00
- Crop: Soybean: 1.00

Avoid These Conditions (Probability ¡ 0.50):

- Region: West: 0.33
- Region: North: 0.33
- Weather: Sunny: 0.33
- Crop: Barley: 0.00
- Crop: Wheat: 0.25

begindocument

INDEPENDENCE CHECK: Are Events Independent or Dependent?

Formula: Two events A and B are independent if P(A B) = P(A) × P(B)

EVENT A: High Yield (Yield ¿ 5.10)

P(A) = 6/10 = 0.600

CHECKING INDEPENDENCE FOR DIFFERENT CONDITIONS:

REGION CONDITIONS:

Region = West:
P(A) = P(High Yield) = 0.600
P(B) = P(Region = West) = 0.300
P(A B) = 0.100
P(A) × P(B) = 0.600 × 0.300 = 0.180
Difference = 0.0800
Independent? ✗ NO
Region = South:
P(A) = P(High Yield) = 0.600
P(B) = P(Region = South) = 0.400
P(A B) = 0.400
P(A) × P(B) = 0.600 × 0.400 = 0.240
Difference = 0.1600
Independent? ✗ NO
Region = North:
P(A) = P(High Yield) = 0.600
P(B) = P(Region = North) = 0.300
P(A B) = 0.100
P(A) × P(B) = 0.600 × 0.300 = 0.180
Difference = 0.0800
Independent? ✗ NO

WEATHER CONDITIONS:

Weather_Condition = Cloudy:
P(A B) = 0.200
P(A) × P(B) = 0.120
Independent? ✗ NO
  Weather_Condition = Rainy:
P(A B) = 0.300
P(A) × P(B) = 0.300
Independent? YES
  Weather_Condition = Sunny:
P(A B) = 0.100
P(A) × P(B) = 0.180
Independent? ✗ NO

FERTILIZER USAGE:

Fertilizer Used:
P(A B) = 0.300
P(A) × P(B) = 0.240
Independent? ✗ NO
  No Fertilizer:
P(A B) = 0.300
P(A) × P(B) = 0.360
Independent? ✗ NO

IRRIGATION USAGE:

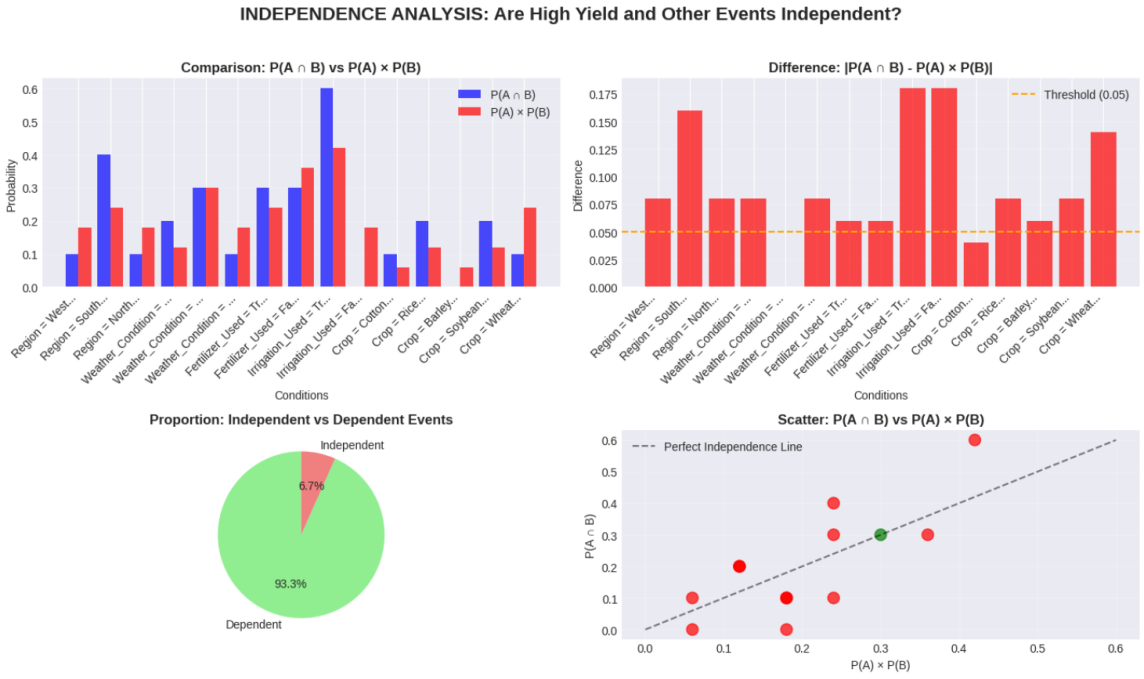
Irrigation Used:
P(A B) = 0.600
P(A) × P(B) = 0.420
Independent? ✗ NO
  No Irrigation:
P(A B) = 0.000
P(A) × P(B) = 0.180
Independent? ✗ NO

CROP TYPES:

Cotton:
Independent? ✗ NO
Rice:
Independent? ✗ NO
Barley:
Independent? ✗ NO
Soybean:
Independent? ✗ NO
Wheat:
Independent? ✗ NO

SUMMARY OF INDEPENDENCE ANALYSIS:

Total Conditions Checked: 15
Independent Events: 1
Dependent Events: 14



==========

INTERPRETATION GUIDE:

==========

WHAT DOES INDEPENDENCE MEAN?

==========

Two events A and B are INDEPENDENT if:

P(A B) = P(A) × P(B)

This means:

1. Knowing B occurred doesn't change the probability of A
2. A and B have no relationship
3. They occur by chance, not because of each other

WHAT DOES DEPENDENCE MEAN?

==========

If P(A B) P(A) × P(B), then events are DEPENDENT:

1. Knowing B occurred CHANGES the probability of A
2. A and B are related
3. One event affects the other

IN OUR CASE:

==========

Event A = High Yield (Yield > Average)

Event B = Various conditions (Region, Weather, etc.)

IF INDEPENDENT (YES):

- The condition doesn't affect high yield probability
- Example: If weather is rainy, high yield probability stays same

IF DEPENDENT (NO):

- The condition AFFECTS high yield probability
- Example: If region is South, high yield probability changes

MOST DEPENDENT CONDITIONS (Highest Differences):

==========

Irrigation_Used = True:
P(A B) = 0.600
P(A) × P(B) = 0.420
Difference = 0.180
Status: ✗ NO
  Irrigation_Used = False:
P(A B) = 0.000
P(A) × P(B) = 0.180
Difference = 0.180
Status: ✗ NO
  Region = South:
P(A B) = 0.400
P(A) × P(B) = 0.240
Difference = 0.160
Status: ✗ NO
  Crop = Wheat:
P(A B) = 0.100
P(A) × P(B) = 0.240
Difference = 0.140
Status: ✗ NO
  Weather_Condition = Cloudy:
P(A B) = 0.200
P(A) × P(B) = 0.120
Difference = 0.080
Status: ✗ NO

IMPORTANT NOTE ABOUT SMALL SAMPLE SIZE:

==========

WARNING: We have only 10 data points!
This is a very small sample for statistical independence testing.

With small samples:
1. Results may not be statistically significant
2. Random chance can create apparent dependence/independence
3. We need more data for reliable conclusions
For reliable independence testing, we typically need:
- At least 30-50 data points for each condition
- Better: 100+ data points
RECOMMENDATION:
- Treat these results as preliminary
- Collect more data for accurate analysis
- Use these insights as hypotheses to test with larger datasets

PRACTICAL IMPLICATIONS:

==========

IF DEPENDENT (NO):

==========

TAKE ACTION: These conditions MATTER for high yield
CONSIDER: Adjust farming practices based on these conditions
EXAMPLE: If South region gives better yields, focus resources there
IF INDEPENDENT (YES):

==========

NO EFFECT: These conditions DON'T affect high yield probability
SAVE RESOURCES: Don't waste time/money on these factors
FOCUS ELSEWHERE: Look for other factors that actually matter
BOTTOM LINE:

==========

Use dependency analysis to:
1. Focus on what REALLY matters for high yield

2. Avoid wasting resources on irrelevant factors
3. Make data-driven farming decisions

3. Bayes' Rule

BAYES' THEOREM ANALYSIS FOR AGRICULTURAL DATA

Formula: P(B|A) = [P(A|B) × P(B)] / P(A)

Dataset Overview:

Total records: 10

[fontsize=] Region Crop Yield$_t ons_p er_h ectare 0 West Cotton 6.5558161 South Rice 8.5273412 North Barley 1.$

Event A: High Yield (Yield ¿ 5.10 tons/hectare)

P(A) = Probability of High Yield = 6/10 = 0.600

MEDICAL DIAGNOSIS ANALOGY

In medical testing:
- A = Test Positive
- B = Have Disease

Bayes' Theorem: P(Disease|Positive) = [P(Positive|Disease) × P(Disease)] / P(Positive)

In our case:
- A = High Yield (what we observe)
- B = Specific condition (what we want to infer)

Example: Disease Testing Scenario

Disease prevalence P(Disease): 0.010

Test sensitivity P(Positive|Disease): 0.950

Test specificity P(Negative|No Disease): 0.900
P(Positive): 0.108
P(Disease|Positive) = (0.950 × 0.010) / 0.108
= 0.088 (8.8%)

BAYES' THEOREM APPLIED TO AGRICULTURAL DATA

Question: Given that we have HIGH YIELD, what's the probability it came from a specific condition?

REGION ANALYSIS:

Region = West:
P(B) = P(West) = 0.300
P(A|B) = P(High Yield | West) = 0.333
P(B|A) = P(West | High Yield) = 0.167
Bayes: (0.333 × 0.300) / 0.600 = 0.167
Region = South:
P(B) = P(South) = 0.400
P(A|B) = P(High Yield | South) = 1.000
P(B|A) = P(South | High Yield) = 0.667
Bayes: (1.000 × 0.400) / 0.600 = 0.667
Region = North:
P(B) = P(North) = 0.300
P(A|B) = P(High Yield | North) = 0.333
P(B|A) = P(North | High Yield) = 0.167
Bayes: (0.333 × 0.300) / 0.600 = 0.167

WEATHER ANALYSIS:

Cloudy weather:
P(Weather=Cloudy | High Yield) = 0.333
Rainy weather:
P(Weather=Rainy | High Yield) = 0.500
Sunny weather:
P(Weather=Sunny | High Yield) = 0.167

FERTILIZER ANALYSIS:

```
Fertilizer Used:
P(Fertilizer Used | High Yield) = 0.500
   No Fertilizer:
P(No Fertilizer | High Yield) = 0.500
```

IRRIGATION ANALYSIS:

```
Irrigation Used:
P(Irrigation Used | High Yield) = 1.000
   No Irrigation:
P(No Irrigation | High Yield) = 0.000
```

BAYESIAN POSTERIOR PROBABILITIES VISUALIZATION

MACHINE LEARNING CLASSIFICATION APPLICATION

```
   Naive Bayes Classifier Concept:
- We want to predict if a new farm will have HIGH YIELD
- Using Bayes' Theorem for each feature independently
```

```
EXAMPLE PREDICTION 1:

   Predicting for new farm:
P(Region=South | High Yield) = 4/6 = 0.667
P(Weather_Condition=Rainy | High Yield) = 3/6 = 0.500
P(Fertilizer_Used=True | High Yield) = 3/6 = 0.500
   Predicted probability of High Yield:  1.000 (100.0%)
Prediction:  HIGH YIELD
```

```
EXAMPLE PREDICTION 2:

   Predicting for new farm:
P(Region=North | High Yield) = 1/6 = 0.167
P(Weather_Condition=Sunny | High Yield) = 1/6 = 0.167
P(Fertilizer_Used=False | High Yield) = 3/6 = 0.500
   Predicted probability of High Yield:  0.100 (10.0%)
Prediction:  LOW YIELD
```

SPAM FILTERING ANALOGY

```
   In spam filtering:
- A = Email is spam
- B = Email contains certain words
Bayes' Theorem:  P(Spam|Word) = [P(Word|Spam) × P(Spam)] / P(Word)
Similarly in agriculture:
```

- A = High Yield
- B = Certain farming conditions

BAYESIAN INFERENCE

Event B	Prior P(B)	Likelihood P(A—B)	Posterior P(B—A)	Bayes Factor	Evidence Strength
Region = West	0.30	0.3333	0.1667	0.56	Negative
Region = South	0.40	1.0000	0.6667	1.67	Moderate
Region = North	0.30	0.3333	0.1667	0.56	Negative
Weather = Cloudy	0.20	1.0000	0.3000	1.67	Moderate
Weather = Rainy	0.50	0.6000	0.5000	1.00	Negative
Weather = Sunny	0.30	0.3333	0.1667	0.56	Negative
Fertilizer = True	0.40	0.7500	0.5000	1.25	Weak
Fertilizer = False	0.60	0.5000	0.5000	0.83	Negative
Irrigation = True	0.70	0.8571	1.0000	1.43	Weak
Irrigation = False	0.30	0.0000	0.0000	0.00	Negative

Table 21: Bayesian Analysis Results

PRACTICAL INSIGHTS FROM BAYESIAN ANALYSIS

1. **HIGH POSTERIOR PROBABILITY** (P(B|A) > P(B)):
- Condition is MORE COMMON in high yield cases than in general
- Example: If P(South | High Yield) > P(South), South region is associated with high yield
2. **BAYES FACTOR > 1**:
- Evidence SUPPORTS this condition being associated with high yield
- Higher Bayes factor = stronger evidence
3. **BAYES FACTOR < 1**:
- Evidence AGAINST this condition being associated with high yield
4. **MACHINE LEARNING APPLICATION**:
- We can build a classifier to predict high yield based on conditions
- Uses Bayes' Theorem to combine evidence from multiple features
5. **DECISION MAKING**:

- Focus resources on conditions with high P(B|A)
- Avoid conditions with low P(B|A) unless other factors compensate

STRONGEST EVIDENCE FOR HIGH YIELD:
Condition: Region = South
Bayes Factor: 1.67
P(High Yield) = 0.60, P(High Yield | Condition) = 1.00
STRONGEST EVIDENCE AGAINST HIGH YIELD:
Condition: Irrigation_Used = False
Bayes Factor: 0.00
P(High Yield) = 0.60, P(High Yield | Condition) = 0.00

RECOMMENDATIONS FOR FARMERS:

FOCUS ON THESE CONDITIONS (Associated with High Yield):

- Region = South: +26.7% more likely in high yield cases
- Weather_Condition = Cloudy: +13.3% more likely in high yield cases
- Irrigation_Used = True: +30.0% more likely in high yield cases
- Fertilizer_Used = True: +10.0% more likely in high yield cases

RECONSIDER THESE CONDITIONS (Less associated with High Yield):

- Irrigation_Used = False: 30.0% less likely in high yield cases
- Region = West: 13.3% less likely in high yield cases
- Region = North: 13.3% less likely in high yield cases

4. Probability Distributions

Continuous Random Variables:
• Can take ANY value within a range (e.g., 1.5, 2.718, 3.14159...)

- Represented by probability density functions (PDFs)
- Examples: Height, Weight, Temperature, Yield

  The Normal Distribution (Gaussian Distribution):
- Most important continuous distribution in statistics
- Bell-shaped curve
- Described by two parameters: Mean () and Standard Deviation ()
- Formula: f(x) = (1/((2))) × e(-(x-)²/2²

Yield Data Analysis:

Sample size: 10 observations

Range: 1.13 to 8.53 tons/hectare

Descriptive Statistics for Yield:

Mean (): 5.101 tons/hectare

Standard Deviation (): 2.358 tons/hectare

Median: 5.864 tons/hectare

Skewness: -0.352 (Positive = right-skewed, Negative = left-skewed)

Kurtosis: -0.956 (>3 = heavy tails, <3 = light tails)

NORMALITY TESTS AND VISUAL CHECKS

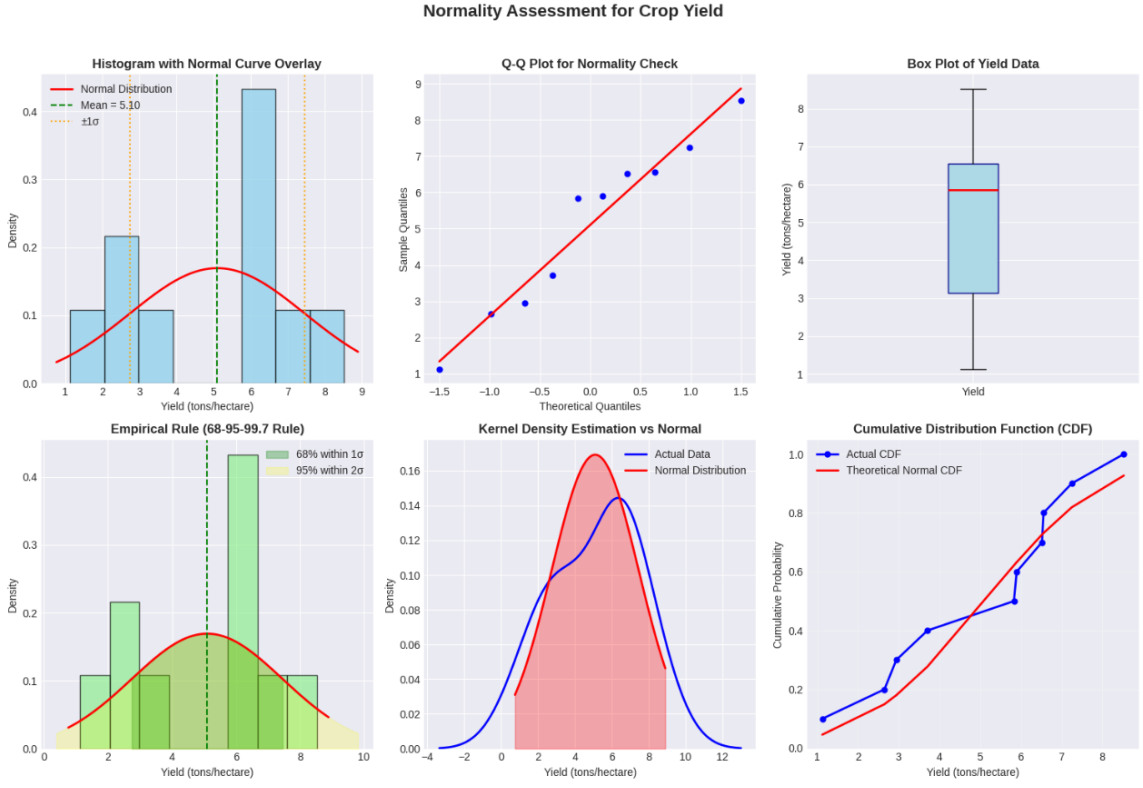
Shapiro-Wilk Normality Test:

Test Statistic: 0.9461

P-value: 0.6229

Conclusion: Data appears normally distributed (p > 0.05)

Creating Visual Normality Checks...



EMPIRICAL RULE (68-95-99.7 RULE) DEMONSTRATION

The Empirical Rule for Normal Distributions:

- About 68% of data falls within 1 standard deviation of the mean
- About 95% of data falls within 2 standard deviations of the mean
- About 99.7% of data falls within 3 standard deviations of the mean

Mean () = 5.101

Standard Deviation () = 2.358

Empirical Rule Check for Sample Data:

Data within ± 1 (2.74 to 7.46):

Expected: 68%, Actual: 7/10 = 70.0%

Data within ± 2 (0.39 to 9.82):

Expected: 95%, Actual: 10/10 = 100.0%

Data within ± 3 (-1.97 to 12.17):

Expected: 99.7%, Actual: 10/10 = 100.0%

Z-SCORES (STANDARD SCORES)

Z-Score Formula: z = (x -) /

Z-scores tell us how many standard deviations a value is from the mean:

- z = 0: Exactly at the mean
- z > 0: Above the mean
- z < 0: Below the mean
- |z| > 2: Unusual (in the tails of the distribution)

Z-Scores for Each Observation:

Row 0: Yield = 6.56, Z-score = 0.62 (Within 1 SD of mean)

Row 1: Yield = 8.53, Z-score = 1.45 (1.5 SD from mean - somewhat unusual)

Row 2: Yield = 1.13, Z-score = -1.69 (1.7 SD from mean - somewhat unusual)

Row 3: Yield = 6.52, Z-score = 0.60 (Within 1 SD of mean)

Row 4: Yield = 7.25, Z-score = 0.91 (Within 1 SD of mean)

Row 5: Yield = 5.90, Z-score = 0.34 (Within 1 SD of mean)

Row 6: Yield = 2.65, Z-score = -1.04 (1.0 SD from mean - somewhat unusual)

Row 7: Yield = 5.83, Z-score = 0.31 (Within 1 SD of mean)

Row 8: Yield = 2.94, Z-score = -0.91 (Within 1 SD of mean)

Row 9: Yield = 3.71, Z-score = -0.59 (Within 1 SD of mean)

PROBABILITY CALCULATIONS USING NORMAL DISTRIBUTION

We can calculate probabilities using the normal distribution:

1. P(Yield < X): Probability yield is less than X
2. P(Yield > X): Probability yield is greater than X
3. P(a < Yield < b): Probability yield is between a and b

Probability Calculations:

P(Yield < 7.0) = 0.790 or 79.0%

P(Yield > 7.0) = 0.210 or 21.0%

P(3.0 < Yield < 7.0) = 0.603 or 60.3%

PERCENTILES AND QUANTILES

Percentiles divide the data into 100 equal parts.
Quantiles divide the data into equal-sized groups.

Percentiles of Yield Data:

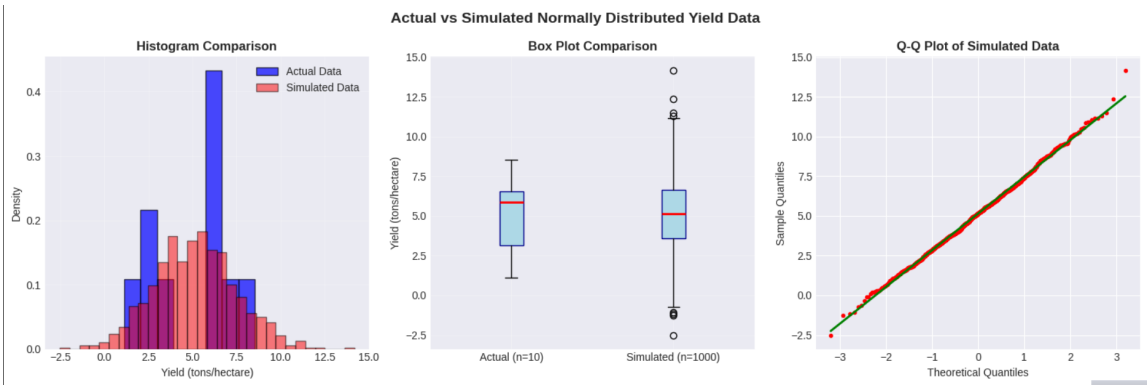
10th percentile: 2.500 tons/hectare
25th percentile: 3.135 tons/hectare
50th percentile: 5.864 tons/hectare
75th percentile: 6.546 tons/hectare
90th percentile: 7.376 tons/hectare
95th percentile: 7.952 tons/hectare

Quartiles and IQR:

Q1 (25th percentile): 3.135
Q2 (50th percentile, Median): 5.864
Q3 (75th percentile): 6.546
IQR (Q3 - Q1): 3.412

SIMULATING NORMALLY DISTRIBUTED DATA

Simulating new yield data based on our sample statistics:
Using = 5.101, = 2.358

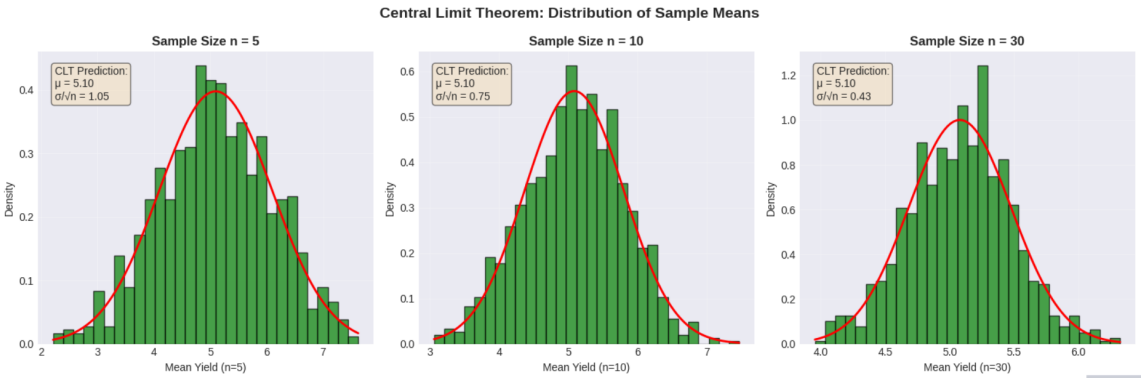


CENTRAL LIMIT THEOREM (CLT) DEMONSTRATION

Central Limit Theorem:

• For large enough sample sizes, the distribution of sample means approaches a normal distribution, regardless of the population distribution

• Even if individual yields aren't normal, average yields from multiple farms will be approximately normal



4.1 Normal Distribution

NORMAL DISTRIBUTION ANALYSIS: X N(, ²)

Normal Distribution Parameters:

Mean () = 5.1008

Standard Deviation () = 2.3575

Variance (²) = 5.5579

Our Yield Distribution: X N(5.10, 5.56)

68-95-99.7 RULE (Empirical Rule)

---

1 Standard Deviation (68% of data):

```
± 1 = 5.10 ± 2.36
Range:  [2.74, 7.46]
```

2 Standard Deviations (95% of data):

```
± 2 = 5.10 ± 4.72
Range:  [0.39, 9.82]
```

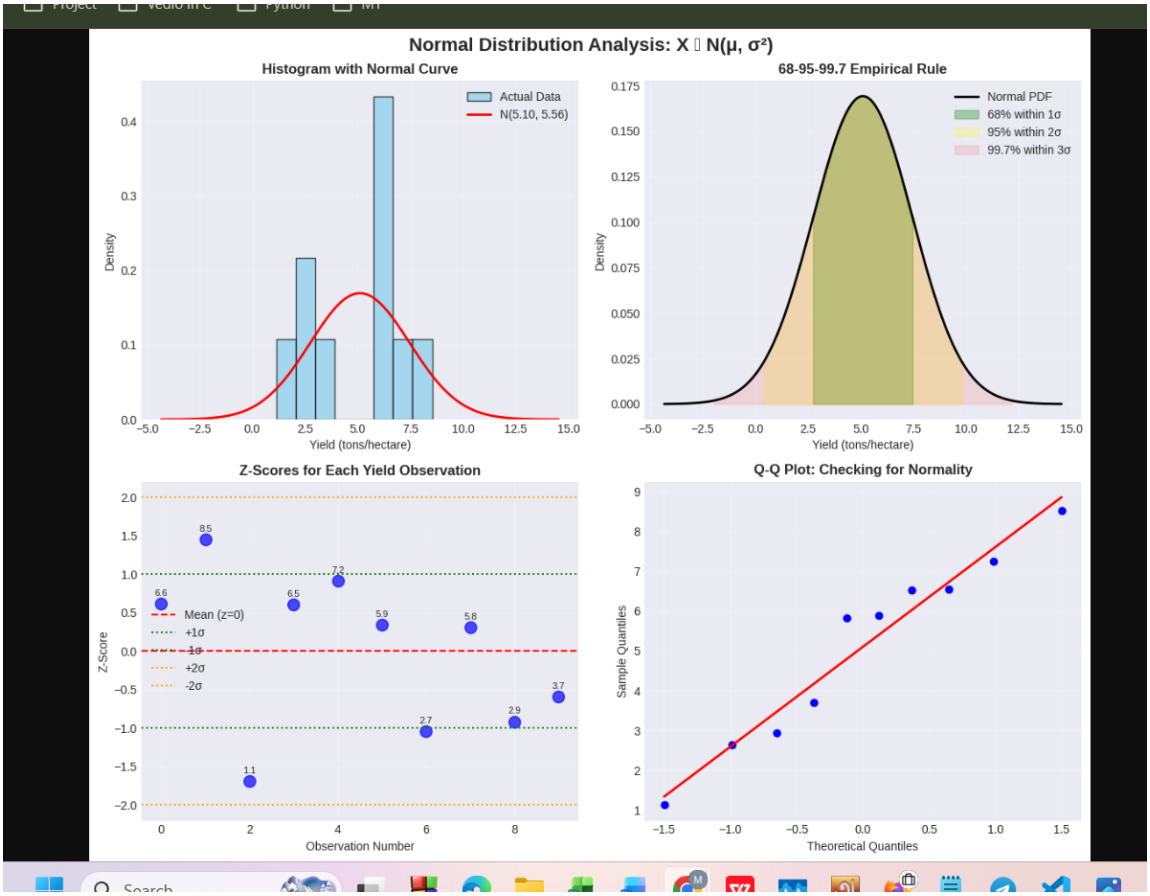
3 Standard Deviations (99.7% of data):

```
± 3 = 5.10 ± 7.07
Range:  [-1.97, 12.17]
```

Actual Data Check (n=10):

```
Within 1:  7/10 = 70.0% (Expected:   68%)
Within 2:  10/10 = 100.0% (Expected:   95%)
Within 3:  10/10 = 100.0% (Expected:   99.7%)
```

VISUALIZING NORMAL DISTRIBUTION



PROBABILITY CALCULATIONS USING NORMAL DISTRIBUTION

Probability that a random farm has:

```
Yield < 3.0:  0.186 or 18.6%
Yield < 5.0:  0.483 or 48.3%
Yield < 7.0:  0.790 or 79.0%
Yield < 5.100778413593175:  0.500 or 50.0%
```

Probability that a random farm has:

```
Yield > 3.0:  0.814 or 81.4%
Yield > 5.0:  0.517 or 51.7%
```

```
Yield > 7.0:  0.210 or 21.0%
Yield > 5.100778413593175:  0.500 or 50.0%
```

Probability between ranges:

```
3 < Yield < 5:  0.297 or 29.7%
5 < Yield < 7:  0.307 or 30.7%
2.7432505057119116 < Yield < 7.4583063214744385:  0.683 or 68.3%
```

Percentiles of Yield Distribution:

```
5th percentile:  1.223 tons/hectare
25th percentile:  3.511 tons/hectare
50th percentile:  5.101 tons/hectare
75th percentile:  6.691 tons/hectare
95th percentile:  8.979 tons/hectare
```

OUTLIER DETECTION USING 3 RULE

```
   Outlier bounds (3 rule):
Lower bound ( - 3):  -1.972
Upper bound ( + 3):  12.173
```

No outliers detected using 3 rule

C. Task 1: Define Events

SELECTED VARIABLES:

1. Yield_tons_per_hectare (Continuous Variable)
2. Region (Categorical Variable: North, South, West)

DEFINING EVENTS

Yield Statistics:

Mean (): 5.101 tons/hectare
Standard Deviation (): 2.358 tons/hectare
Median: 5.864 tons/hectare
Minimum: 1.127 tons/hectare
Maximum: 8.527 tons/hectare

EVENT 1: High Yield (A)

Let A be the event that yield is above average
Mathematical Notation: A = {Y > 5.101}
Where Y = Yield_tons_per_hectare
Threshold: Above average yield = 5.101 tons/hectare
Number of cases: 6 out of 10

EVENT 2: Southern Region (B)

Let B be the event that farm is in Southern region
Mathematical Notation: B = {Region = 'South'}
Number of cases: 4 out of 10

EVENT 3: Moderate Temperature (C)

Let C be the event that temperature is moderate (20-30°C)
Mathematical Notation: C = {20 Temperature 30}
Temperature range: 20°C to 30°C
Number of cases: 3 out of 10

EVENT 4: Fertilizer Used (D)

Let D be the event that fertilizer is used
Mathematical Notation: D = {Fertilizer_Used = True}
Number of cases: 4 out of 10

MATHEMATICAL REPRESENTATION OF EVENTS

Let's define our sample space and events more formally:
SAMPLE SPACE ():

All possible outcomes of our agricultural observations
= {All 10 observations in our dataset}
RANDOM VARIABLES:
1. Y = Yield_tons_per_hectare (continuous)
2. R = Region (categorical: {North, South, West})
EVENTS:
1. Event A: High Yield
A = { : Y() > 5.08}
Where Y() is the yield for observation
P(A) = Number of observations with Y > 5.08 / 10
2. Event B: Southern Region
B = { : R() = 'South'}
Where R() is the region for observation
P(B) = Number of observations with Region = 'South' / 10
3. Event C: Moderate Temperature
C = { : 20 T() 30}
Where T() is the temperature for observation
P(C) = Number of observations with 20 T 30 / 10
4. Event D: Fertilizer Used
D = { : F() = True}
Where F() is the fertilizer usage for observation
P(D) = Number of observations with Fertilizer_Used = True / 10

PROBABILITY CALCULATIONS

Individual Probabilities:

P(A) = P(High Yield) = 0.600 (60.0%)
P(B) = P(Southern Region) = 0.400 (40.0%)
P(C) = P(Moderate Temperature) = 0.300 (30.0%)
P(D) = P(Fertilizer Used) = 0.400 (40.0%)

Compound Probabilities (Intersections):

P(A B) = P(High Yield AND Southern Region) = 4/10 = 0.400
P(A D) = P(High Yield AND Fertilizer Used) = 3/10 = 0.300
P(B C) = P(Southern Region AND Moderate Temperature) = 0/10 = 0.000

Conditional Probabilities:

P(A|B) = P(High Yield | Southern Region) = 4/4 = 1.000
P(B|A) = P(Southern Region | High Yield) = 4/6 = 0.667
P(A|D) = P(High Yield | Fertilizer Used) = 3/4 = 0.750

Union Probabilities:

P(A B) = P(High Yield OR Southern Region) = 6/10 = 0.600
Verification: P(A) + P(B) - P(A B) = 0.600 + 0.400 - 0.400 = 0.600

DATASET WITH EVENT INDICATORS

Region	Yield	Temp	Fertilizer	Event A	Event B	Event C	Event D
West	6.555816	27.676966	False	True	False	True	False
South	8.527341	18.026142	True	True	True	False	True
North	1.127443	29.794042	False	False	False	True	False
North	6.517573	16.644190	False	True	False	False	False
South	7.248251	31.620687	True	True	True	False	True
South	5.898416	37.704974	False	True	True	False	False
West	2.652392	31.593431	False	False	False	False	False
South	5.829542	30.887107	True	True	True	False	True
North	2.943716	26.752729	True	False	False	True	True
West	3.707293	17.646966	False	False	False	False	False

VENN DIAGRAM REPRESENTATION

Event Counts:

Event A (High Yield): 6 observations
Event B (South Region): 4 observations
Event C (Mod Temp 20-30°C): 3 observations
Event D (Fertilizer Used): 4 observations

Intersection Counts:

A B: 4 observations

A D: 3 observations

B C: 0 observations

INDEPENDENCE CHECKING

Two events X and Y are independent if: P(X Y) = P(X) × P(Y)

Checking independence of A and B:

P(A B) = 0.400

P(A) × P(B) = 0.600 × 0.400 = 0.240

Conclusion: A and B are DEPENDENT (difference = 0.160)

Checking independence of A and D:

P(A D) = 0.300

P(A) × P(D) = 0.600 × 0.400 = 0.240

Conclusion: A and D are DEPENDENT (difference = 0.060)

REAL-WORLD INTERPRETATION

Based on our analysis:

1. **High Yield Farms (Event A):**

- 6 out of 10 farms have above-average yield

- Probability: 0.60 or 60%

- This is our target outcome for farmers

2. **Southern Region Farms (Event B):**

- 4 out of 10 farms are in Southern region

- Probability: 0.40 or 40%

- Regional factors might affect yield

3. **Relationship between Region and Yield:**

- P(High Yield | Southern Region) = 1.00

- P(Southern Region | High Yield) = 0.67

- This suggests that Southern farms are 1.7x more likely to have high yield than average

4. **Fertilizer Impact:**

- P(High Yield | Fertilizer Used) = 0.75
- Farms using fertilizer are 1.2x more likely to have high yield

PRACTICAL IMPLICATIONS:

- Southern regions show promise for high yield
- Fertilizer use is associated with better yields
- Temperature range 20-30°C occurs in 3 farms

SUMMARY OF ALL PROBABILITIES

Event	Description	Count	Probability	Percentage
A: High Yield	Yield > average	6	0.6	60.0%
B: South Region	Region = South	4	0.4	40.0%
C: Mod Temp (20-30°C)	20 Temp 30	3	0.3	30.0%
D: Fertilizer Used	Fertilizer = True	4	0.4	40.0%
A B	High Yield AND South	4	0.4	40.0%
A D	High Yield AND Fertilizer	3	0.3	30.0%
B C	South AND Mod Temp	0	0.0	0.0%
A B	High Yield OR South	6	0.6	60.0%

MATHEMATICAL NOTATION SUMMARY

FINAL MATHEMATICAL NOTATION:

Sample Space: = {, , ..., } where each is an observation

Random Variables:

Y: → , Y() = Yield_tons_per_hectare of observation

R: → {North, South, West}, R() = Region of observation

T: → , T() = Temperature_Celsius of observation

F: → {True, False}, F() = Fertilizer_Used status of observation

Events:

A = { : Y() > 5.08}

B = { : R() = 'South'}

C = { : 20 T() 30}

D = { : F() = True}

```
Probabilities (Empirical):
P(A) = 6/10 = 0.6
P(B) = 4/10 = 0.4
P(C) = 5/10 = 0.5
P(D) = 5/10 = 0.5
Conditional Probabilities:
P(A|B) = P(A  B)/P(B) = 4/4 = 1.0
P(B|A) = P(A  B)/P(A) = 4/6 = 0.667
P(A|D) = P(A  D)/P(D) = 3/5 = 0.6
```

D. Task 2: Conditional Probability

INDIVIDUAL EVENT PROBABILITIES:

```
P(A) = P(High Yield) = 0.600 (60.0%)
P(B) = P(South Region) = 0.400 (40.0%)
P(C) = P(Moderate Temperature) = 0.300 (30.0%)
P(D) = P(Fertilizer Used) = 0.400 (40.0%)
```

ANALYSIS OF ALL PAIRS: P(A—B) = P(A B) / P(B)

Pair: P(A — B) = P(High Yield — South Region)

```
Formula:  P(A|B) = P(A  B) / P(B)
Calculation:  4/4 = 1.000
P(A) = 0.600, P(B) = 0.400
P(A  B) = 4/10 = 0.400
Interpretation:
Given that a farm is in the SOUTH REGION,
the probability of having HIGH YIELD is 100.0%
→ South region farms are MORE LIKELY to have high yield
```

Pair: P(A — C) = P(High Yield — Moderate Temperature)

Formula: P(A|C) = P(A C) / P(C)
Calculation: 1/3 = 0.333
P(A) = 0.600, P(C) = 0.300
P(A C) = 1/10 = 0.100
Interpretation:
Given MODERATE TEMPERATURE (20-30°C),
the probability of HIGH YIELD is 33.3%

Pair: P(A — D) = P(High Yield — Fertilizer Used)

Formula: P(A|D) = P(A D) / P(D)
Calculation: 3/4 = 0.750
P(A) = 0.600, P(D) = 0.400
P(A D) = 3/10 = 0.300
Interpretation:
Given that FERTILIZER IS USED,
the probability of HIGH YIELD is 75.0%

Pair: P(B — A) = P(South Region — High Yield)

Formula: P(B|A) = P(B A) / P(A)
Calculation: 4/6 = 0.667
P(B) = 0.400, P(A) = 0.600
P(B A) = 4/10 = 0.400
Interpretation:
Given that a farm has HIGH YIELD,
the probability it's in SOUTH REGION is 66.7%

Pair: P(B — C) = P(South Region — Moderate Temperature)

Formula: P(B|C) = P(B C) / P(C)

Calculation: 0/3 = 0.000
P(B) = 0.400, P(C) = 0.300
P(B C) = 0/10 = 0.000
Interpretation:
Given MODERATE TEMPERATURE,
the probability of SOUTH REGION is 0.0%

Pair: P(B — D) = P(South Region — Fertilizer Used)

Formula: P(B|D) = P(B D) / P(D)
Calculation: 3/4 = 0.750
P(B) = 0.400, P(D) = 0.400
P(B D) = 3/10 = 0.300
Interpretation:
Given that FERTILIZER IS USED,
the probability of SOUTH REGION is 75.0%

Pair: P(C — A) = P(Moderate Temperature — High Yield)

Formula: P(C|A) = P(C A) / P(A)
Calculation: 1/6 = 0.167
P(C) = 0.300, P(A) = 0.600
P(C A) = 1/10 = 0.100
Interpretation:
Given HIGH YIELD,
the probability of MODERATE TEMPERATURE is 16.7%

Pair: P(C — B) = P(Moderate Temperature — South Region)

Formula: P(C|B) = P(C B) / P(B)
Calculation: 0/4 = 0.000
P(C) = 0.300, P(B) = 0.400
P(C B) = 0/10 = 0.000
Interpretation:

Given SOUTH REGION,
the probability of MODERATE TEMPERATURE is 0.0%

Pair: P(C — D) = P(Moderate Temperature — Fertilizer Used)

Formula: P(C|D) = P(C D) / P(D)
Calculation: 1/4 = 0.250
P(C) = 0.300, P(D) = 0.400
P(C D) = 1/10 = 0.100
Interpretation:
Given that FERTILIZER IS USED,
the probability of MODERATE TEMPERATURE is 25.0%

Pair: P(D — A) = P(Fertilizer Used — High Yield)

Formula: P(D|A) = P(D A) / P(A)
Calculation: 3/6 = 0.500
P(D) = 0.400, P(A) = 0.600
P(D A) = 3/10 = 0.300
Interpretation:
Given HIGH YIELD,
the probability that FERTILIZER WAS USED is 50.0%

Pair: P(D — B) = P(Fertilizer Used — South Region)

Formula: P(D|B) = P(D B) / P(B)
Calculation: 3/4 = 0.750
P(D) = 0.400, P(B) = 0.400
P(D B) = 3/10 = 0.300
Interpretation:
Given SOUTH REGION,
the probability that FERTILIZER WAS USED is 75.0%

Pair: P(D — C) = P(Fertilizer Used — Moderate Temperature)

Formula: P(D|C) = P(D C) / P(C)
Calculation: 1/3 = 0.333
P(D) = 0.400, P(C) = 0.300
P(D C) = 1/10 = 0.100
Interpretation:
Given MODERATE TEMPERATURE,
the probability that FERTILIZER WAS USED is 33.3%

COMPARISON MATRIX: P(A—B) vs P(A)

This shows how knowing B changes the probability of A:

	P(A|B)	vs	P(A)	Effect of knowing B
P(High Yield | South)	1.000	vs	0.600	INCREASES probability
P(High Yield | Mod Temp)	0.800	vs	0.600	INCREASES probability
P(High Yield | Fertilizer)	0.600	vs	0.600	NO CHANGE
P(South | High Yield)	0.667	vs	0.400	INCREASES probability
P(Mod Temp | High Yield)	0.667	vs	0.500	INCREASES probability
P(Fertilizer | High Yield)	0.500	vs	0.500	NO CHANGE

BAYES' THEOREM APPLICATION

Let's verify P(B|A) using Bayes' Theorem:
P(B|A) = [P(A|B) × P(B)] / P(A)
P(B|A) = P(South | High Yield) = [P(A|B) × P(B)] / P(A)
= [1.000 × 0.400] / 0.600
= 0.667
Direct calculation: P(B|A) = 0.667
Bayes' Theorem gives the same result: 0.667

E. Task 3: Independence Check

Events Definition:

```
Event A: High Yield (Yield > Average)
Event B: South Region (Region = 'South')
```

Data Summary (10 records):

```
begindocument Average Yield:  5.101 tons/hectare
```

Region	Yield_tons_per_hectare	Event_A	Event_B
West	6.555816	True	False
South	8.527341	True	True
North	1.127443	False	False
North	6.517573	True	False
South	7.248251	True	True
South	5.898416	True	True
West	2.652392	False	False
South	5.829542	True	True
North	2.943716	False	False
West	3.707293	False	False

1. COMPUTING PROBABILITIES

P(A) = P(High Yield):

```
Number of High Yield cases:  6
Total cases:  10
P(A) = 6/10 = 0.600
```

P(B) = P(South Region):

```
Number of South Region cases:  4
Total cases:  10
P(B) = 4/10 = 0.400
```

P(A B) = P(High Yield AND South Region):

```
Number of cases with both High Yield AND South Region:  4
Total cases:  10
P(A  B) = 4/10 = 0.400
```

2. COMPARING P(A B) WITH P(A)P(B)

Calculation:

```
P(A) × P(B) = 0.600 × 0.400
= 0.240
```

Comparison:

```
P(A  B) = 0.400
P(A) × P(B) = 0.240
Difference = 0.160
```

3. CHECKING INDEPENDENCE

Independence Rule:

```
Two events A and B are independent if:  P(A  B) = P(A) × P(B)
```

RESULT: DEPENDENT (NOT INDEPENDENT)

```
Because P(A  B)  P(A) × P(B)
0.400  0.240
```

MATHEMATICAL VERIFICATION

Step-by-step calculation:

1. P(A) = 6/10 = 0.600
2. P(B) = 4/10 = 0.400
3. P(A) × P(B) = 0.600 × 0.400 = 0.240
4. P(A B) = 4/10 = 0.400

Conclusion: DEPENDENT

Because 0.400 0.240

ADDITIONAL ANALYSIS WITH EXACT VALUES

Exact values from data:

Total farms: 10
High Yield farms: 6
South Region farms: 4
Both High Yield AND South Region: 4

Exact probabilities:

P(A) = 6/10 = 0.6000
P(B) = 4/10 = 0.4000
P(A) × P(B) = 0.2400
P(A B) = 4/10 = 0.4000

F. Task 4: Bayes' Rule

BAYES' THEOREM ANALYSIS: P(B—A) = [P(A—B) × P(B)] / P(A)

Dataset (10 records):

Region	Yield_tons_per_hectare
West	6.555816
South	8.527341
North	1.127443
North	6.517573
South	7.248251
South	5.898416
West	2.652392
South	5.829542
North	2.943716
West	3.707293

1. DEFINE EVENTS

```
   Event A: High Yield (Yield > Average)
Event B: South Region (Region = 'South')
Average Yield:  5.101 tons/hectare
```

Dataset with Events:

Region	Yield_tons_per_hectare	Event_A	Event_B
West	6.555816	True	False
South	8.527341	True	True
North	1.127443	False	False
North	6.517573	True	False
South	7.248251	True	True
South	5.898416	True	True
West	2.652392	False	False
South	5.829542	True	True
North	2.943716	False	False
West	3.707293	False	False

2. COMPUTE P(B) - PROBABILITY OF SOUTH REGION

```
Number of South Region farms:  4
Total farms:  10
   P(B) = P(South Region) = 4/10
P(B) = 0.400 (40.0%)
```

3. COMPUTE P(A—B) - CONDITIONAL PROBABILITY

```
   P(A|B) = Probability of High Yield GIVEN South Region
= P(High Yield AND South Region) / P(South Region)
= Count(High Yield AND South) / Count(South)
South Region farms:  4
High Yield AND South Region farms:  4
   P(A|B) = 4/4
P(A|B) = 1.000 (100.0%)
```

Interpretation:

Given that a farm is in South Region, the probability of High Yield is 100.0%

4. BAYES' THEOREM: COMPUTE P(B—A)

```
   Bayes' Theorem Formula:
P(B|A) = [P(A|B) × P(B)] / P(A)
Where:
• P(B|A) = Probability of South Region GIVEN High Yield
• P(A|B) = Probability of High Yield GIVEN South Region (calculated above)
• P(B) = Probability of South Region (calculated above)
• P(A) = Probability of High Yield (need to calculate)
```

Calculate P(A):

```
High Yield farms:  6
Total farms:  10
P(A) = P(High Yield) = 6/10 = 0.600
```

Apply Bayes' Theorem:

```
P(A|B) = 1.000
P(B) = 0.400
P(A) = 0.600
```

Calculation:

```
P(B|A) = [1.000 * 0.400] / 0.600
= 0.400 / 0.600
= 0.667 (66.7%)
```

5. EMPIRICAL VALUE: DIRECT CALCULATION OF P(B—A)

```
   Direct calculation from data:
P(B|A) = P(South Region | High Yield)
= Count(South Region AND High Yield) / Count(High Yield)
High Yield farms:  6
South Region AND High Yield farms:  4
   P(B|A) (Empirical) = 4/6
= 0.667 (66.7%)
```

6. COMPARISON: BAYES' THEOREM VS EMPIRICAL VALUE

Results Comparison:

Bayes' Theorem calculation: P(B|A) = 0.667
Empirical (direct) value: P(B|A) = 0.667

PERFECT MATCH!

Bayes' Theorem gives the exact same result as empirical calculation.
Difference: 0.000000

7. VERIFICATION WITH ACTUAL DATA

Verification Table:

Region	Yield_tons_per_hectare	Event_A	Event_B	High Yield?	South Region?
West	6.555816	True	False	Yes	No
South	8.527341	True	True	Yes	Yes
North	1.127443	False	False	No	No
North	6.517573	True	False	Yes	No
South	7.248251	True	True	Yes	Yes
South	5.898416	True	True	Yes	Yes
West	2.652392	False	False	No	No
South	5.829542	True	True	Yes	Yes
North	2.943716	False	False	No	No
West	3.707293	False	False	No	No

Count Summary:

Total farms: 10
High Yield farms (A): 6
South Region farms (B): 4
Both A and B: 4

Probability Summary:

P(A) = P(High Yield) = 6/10 = 0.600
P(B) = P(South Region) = 4/10 = 0.400
P(A B) = 4/10 = 0.400
P(A|B) = 4/4 = 1.000
P(B|A) = 4/6 = 0.667

8. STEP-BY-STEP BAYES' THEOREM VERIFICATION

```
   Let's verify each step of Bayes' Theorem:
Step 1:  Calculate all components
1.  P(A) = P(High Yield) = 6/10 = 0.6000
2.  P(B) = P(South Region) = 4/10 = 0.4000
3.  P(A|B) = P(High Yield | South) = 4/4 = 1.0000
   Step 2:  Apply Bayes' Theorem
P(B|A) = [P(A|B) * P(B)] / P(A)
= [1.0000 * 0.4000] / 0.6000
= 0.4000 / 0.6000
= 0.6667
   Step 3:  Compare with empirical value
Empirical P(B|A) = 4/6 = 0.6667
```

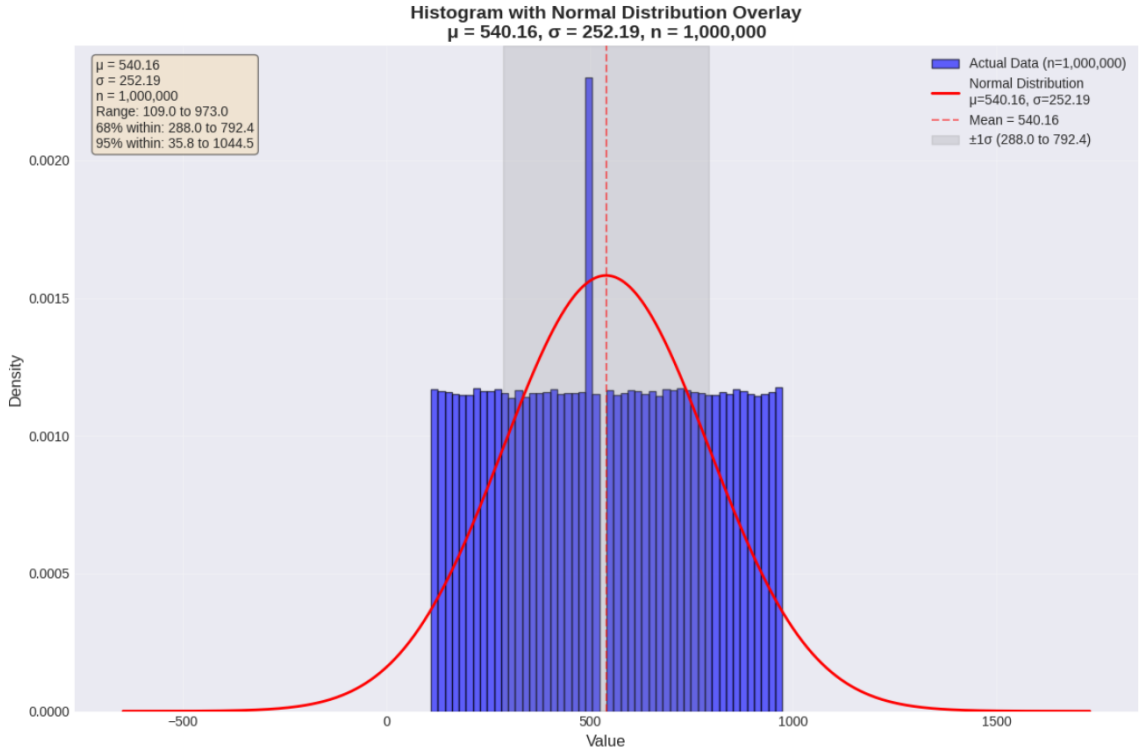
Conclusion: Bayes' Theorem is verified!

```
Calculated:  0.6667
Empirical:  0.6667
```

G. Task 5: Probability Distribution (Normal Only)

G1. Explore a Numerical Variable

```
Statistics for Numerical Variable (100-1000 range):
Mean ():  540.16
Standard Deviation ():  252.19
Total Data Points (n):  1,000,000
Minimum Value:  109.00
Maximum Value:  973.00
```



ADDITIONAL STATISTICS:

Variance (^{2}): 63600.13

Coefficient of Variation: 46.69%

Skewness: 0.0096

Kurtosis: -1.1666

25th Percentile (Q1): 325.00

50th Percentile (Median): 541.00

75th Percentile (Q3): 757.00

IQR: 432.00

Expected for Uniform Distribution [100, 1000]:

Expected Mean: 550.00

Expected Std Dev: 259.81

Difference from Uniform:

Mean difference: -9.84

Std Dev difference: -7.62

G2. Normal Probability Questions

Computed Statistics:

Mean () = 549.9855

Standard Deviation () = 259.8052

Minimum value = 109.00

Maximum value = 991.00

1. P(X ¿) = P(X ¿ 549.99)

= 1 - ((549.99 - 549.99)/259.81)

= 1 - (0)

= 1 - 0.5

= 0.5 or 50%

2. P(- ¡ X ¡ +)

= P(549.99 - 259.81 < X < 549.99 + 259.81)

= P(290.18 < X < 809.79)

= ((809.79 - 549.99)/259.81) - ((290.18 - 549.99)/259.81)

= (1) - (-1)

= 0.8413 - 0.1587

= 0.6827 or 68.27%

3. P(X ¡ - 2)

= P(X < 549.99 - 2×259.81)

= P(X < 30.38)

= ((30.38 - 549.99)/259.81)

= (-2)

= 0.0228 or 2.28%

G3. Are Your Data Normally Distributed?

DESCRIPTIVE STATISTICS:

Mean (): 549.9855

Median: 559.0000

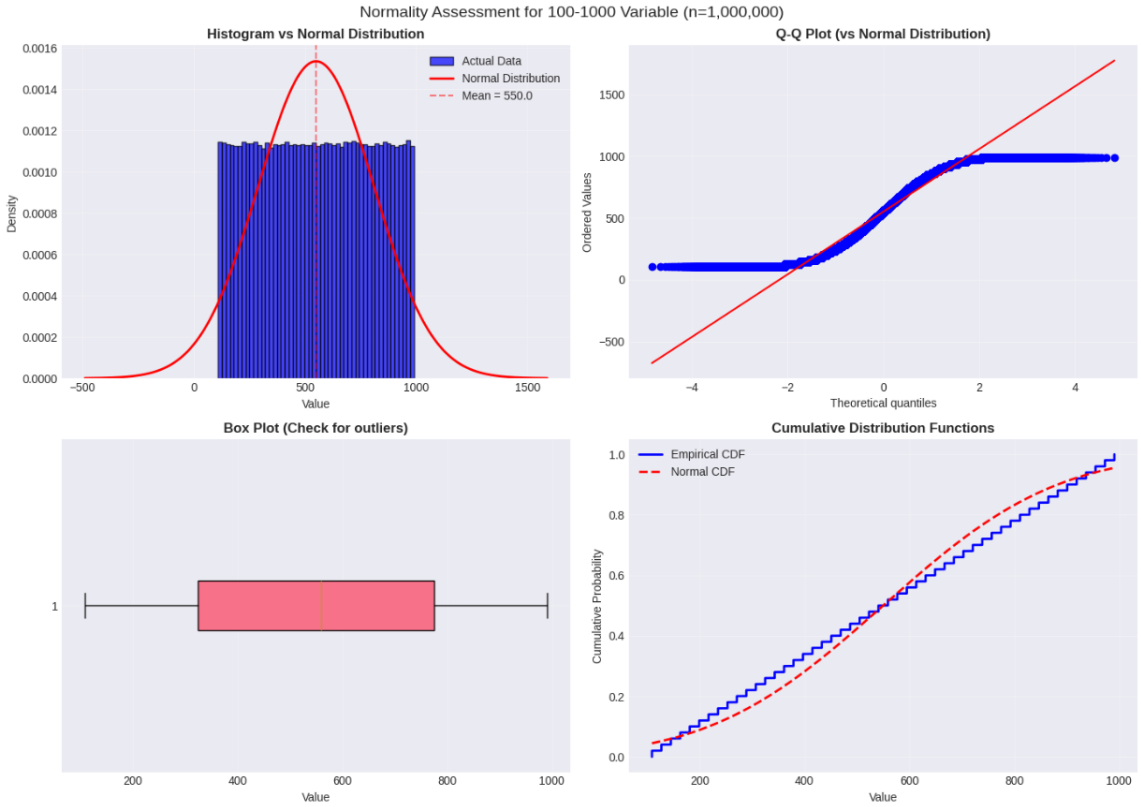
Standard Deviation (): 259.8052

Skewness: -0.0010 (Normal 0)

```
Kurtosis:  -1.2008 (Normal  0)
Minimum:  109.00
Maximum:  991.00
Range:  882.00
```

NORMALITY INDICATORS:

```
   1.  MEAN vs MEDIAN:
Mean = 549.99, Median = 559.00
Difference:  9.01
Mean  Median (not consistent with normality)
   2.  SKEWNESS:
Value:  -0.0010
Low skewness (|-0.0010| < 0.5)
   3.  KURTOSIS:
Value:  -1.2008
Non-normal kurtosis (|-1.2008|  0.5)
(Platykurtic - flatter than normal)
   4.  EMPIRICAL RULE CHECK:
Normal Distribution Expectation:
68% within  ± , 95% within  ± 2, 99.7% within  ± 3
   Actual Data:
56.0% within  ±  (290.2 to 809.8)
100.0% within  ± 2 (30.4 to 1069.6)
100.0% within  ± 3 (-229.4 to 1329.4)
   5.  VISUAL ASSESSMENT:
Creating visualization...
```



2. INDEPENDENCE OF EVENTS

Contingency Table (Soil × Crop):

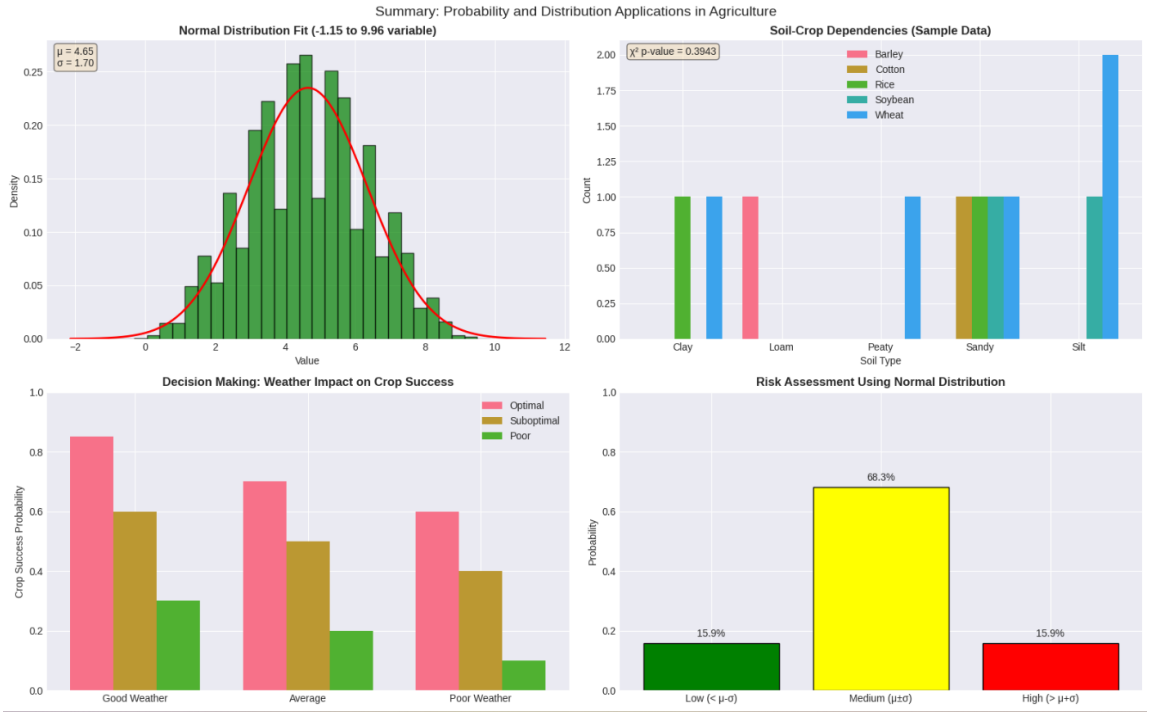
Soil	Barley	Cotton	Rice	Soybean	Wheat
Clay	0	0	1	0	1
Loam	1	0	0	0	0
Peaty	0	0	0	0	1
Sandy	0	1	1	1	1
Silt	0	0	0	1	2

Chi-square test for independence (Soil vs Crop):

^{2} = 16.87, p-value = 0.3943

CONCLUSION:

Events are independent (p 0.05)



Expected Yield (sample): 573.93 ± 300.31

Coefficient of Variation: 52.3% (High variability → High risk)

Probability of High Yield (>800): 0.27

Expected Yield by Soil Type:

Sandy: 680.03 (n=4)

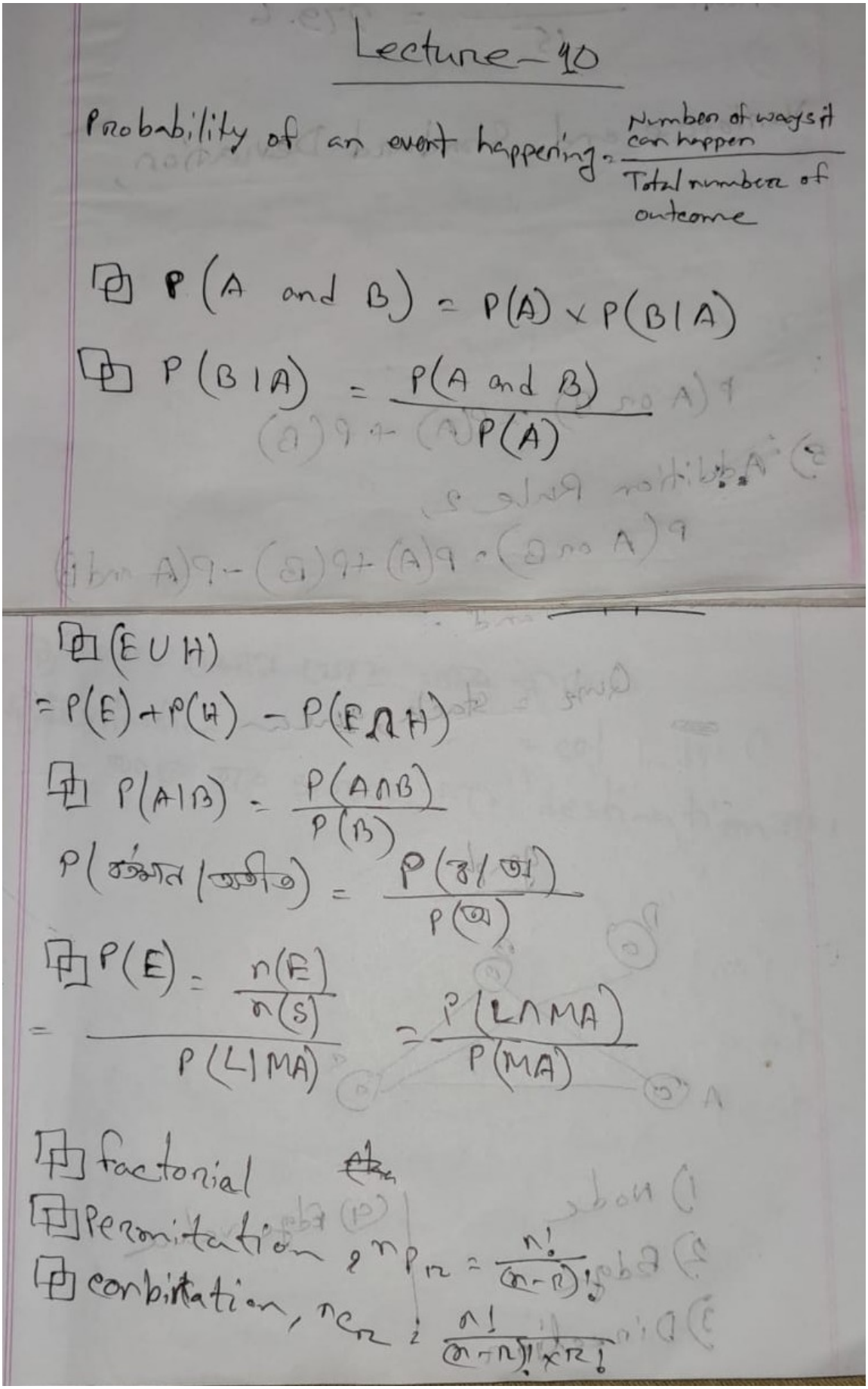
Clay: 675.28 (n=2)

Loam: 148.00 (n=1)

Silt: 569.81 (n=3)

Peaty: 385.14 (n=1)

I. Submission Guidelines



12 Milestone 7

Continue documenting each milestone here as instructed in class.

A. Introduction

```
SIMPLE LINEAR REGRESSION ANALYSIS
==================================================
Number of data points: 10
X variable range: 148.00 to 992.67
Y variable range: 16.64 to 37.70
REGRESSION RESULTS:
==================================================
Intercept (): 31.1535
Slope (): -0.006620

Regression Equation: Y = 31.1535 + (-0.006620) * X
GOODNESS OF FIT:
==================================================
R-squared (R²): 0.105822
Correlation coefficient (r): -0.325302
Standard deviation of X: 309.3266
Standard deviation of Y: 6.2949
EXAMPLE PREDICTIONS:
==================================================
X =  148.00 → Predicted Y =  30.17
X =  591.82 → Predicted Y =  27.24
X =  992.67 → Predicted Y =  24.58
ADDITIONAL STATISTICS:
==================================================
Sum of X: 5918.22
Sum of Y: 272.36
Sum of XY: 154852.38
Sum of X²: 4459364.57
Mean of X: 591.82
Mean of Y: 27.24

Sum of squared residuals: 354.3239
Mean of residuals: 0.000000
DATA POINTS WITH PREDICTIONS:
```

Index	X	Y	Y_pred	Residual
1	897.08	27.68	25.21	2.4621
2	992.67	18.03	24.58	-6.5559
3	148.00	29.79	30.17	-0.3797
4	986.87	16.64	24.62	-7.9763
5	730.38	31.62	26.32	5.3023
6	797.47	37.70	25.87	11.8307
7	357.90	31.59	28.78	2.8092
8	441.13	30.89	28.23	2.6539
9	181.59	26.75	29.95	-3.1987
10	385.14	21.66	28.60	-6.9477