



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

200
1822 - 2022

Quantitative Research

Week 8

Man Geni Addysg Uwch yng Nghymru
Where Higher Education Began in Wales

www.uwtsd.ac.uk

Topics Covered

1. Frequency Distributions
2. Measures of Central Tendency
3. Selecting a Measure of Central Tendency
4. Measures of Variability
5. Using Excel for Statistics.

FREQUENCY DISTRIBUTIONS

A frequency distribution is a summary of how often different values occur in a dataset.

In **Cyber Security** and **Digital Forensics** research it is used to analyze how frequently certain values or events appear in a dataset.

It's useful for detecting anomalies, understanding data patterns, and identifying outliers or suspicious activities.

FREQUENCY DISTRIBUTIONS

1. Network Traffic Analysis-Frequency of IP addresses, ports, or protocols used. Helps in detecting scanning attacks or unusual spikes in traffic.
2. Malware Analysis-Frequency of system calls or API usage by binaries. Reveals common vs suspicious behavior.
3. Log File Analysis-Frequency of login attempts, failed authentications, or file access. Used in brute-force detection, insider threat monitoring, etc.
4. Email/Phishing Detection-Frequency of sender domains, keywords, or links. Detect patterns in spam/phishing attempts.

Public datasets in Cyber Security and Forensics

<https://www.sei.cmu.edu/library/insider-threat-test-dataset/>

<https://www.unb.ca/cic/datasets/index.html>

How to Calculate a Frequency Distribution

Example Dataset (Logins)

Assume a log file records login attempts with this simple data:

User	Timestamp
alice	2025-07-29 09:01:12
bob	2025-07-29 09:03:45
alice	2025-07-29 09:10:02
charlie	2025-07-29 09:15:33
alice	2025-07-29 09:22:55
bob	2025-07-29 09:30:07

How to Calculate a Frequency Distribution

Step 1: Tally Occurrences

User	Frequency
alice	3
bob	2
charlie	1

How to Calculate a Frequency Distribution

Step 3: Visualization <https://python-fiddle.com/examples/matplotlib?checkpoint=1753788994>

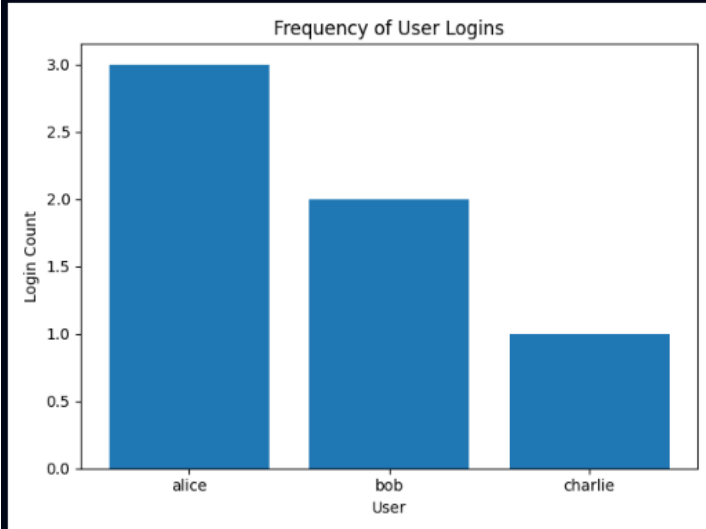
```
import matplotlib.pyplot as plt
```

```
users = ['alice', 'bob', 'charlie']  
frequencies = [3, 2, 1]
```

```
plt.bar(users, frequencies)  
plt.title('Frequency of User Logins')  
plt.xlabel('User')  
plt.ylabel('Login Count')  
plt.show()
```

```
import matplotlib.pyplot as plt  
  
users = ['alice', 'bob', 'charlie']  
frequencies = [3, 2, 1]  
  
plt.bar(users, frequencies)  
plt.title('Frequency of User Logins')  
plt.xlabel('User')  
plt.ylabel('Login Count')  
plt.show()
```

Click **Run** or press **shift + ENTER** to run code



Application in Forensics

Using frequency distribution, forensic analysts can:

- **Detect anomalies:** e.g., a user logging in 500 times a day.
- **Find patterns:** e.g., malware that always uses a certain set of DLLs.
- **Generate timelines:** of frequent file accesses or account activities.
- **Establish baselines:** for "normal" vs "abnormal" behavior.

MEASURES OF CENTRAL TENDENCY

These are numbers that describe the center of a dataset:

Mean, Median, and Mode.

Example Dataset:[4, 5, 6, 8, 10]

Calculations:-

Mean = $(4 + 5 + 6 + 8 + 10) / 5 = 6.6$

Median = 6 (middle value)

Mode = No mode (no value repeats)

MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency are statistical values that describe the **center** or **typical value** in a dataset. These are **widely used in Cyber Security and Digital Forensics** to summarize data

Example Dataset (Failed Logins Per Hour)

[5, 7, 9, 2, 4, 9, 9, 6, 4, 2]

- **Mean** = $(5+7+9+2+4+9+9+6+4+2) / 10 = 57 / 10 = 5.7$
- **Median** = Ordered: [2, 2, 4, 4, 5, 6, 7, 9, 9, 9] → Middle two: $(5+6)/2 = 5.5$
- **Mode** = Most frequent: **9**

MEASURES OF CENTRAL TENDENCY

Use in Cybersecurity & Forensics

Scenario	Description	Measure Used
Brute-force detection	Average failed logins per user per hour	Mean
User behavior analysis	Typical login time of an insider	Median
Malware analysis	Most common API call	Mode
Traffic analysis	Detect normal packet sizes	Mean / Median

MEASURES OF VARIABILITY

Dataset:

[2, 4, 6, 8, 10]

Calculate the Mean

$$\text{Mean} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

MEASURES OF VARIABILITY

Dataset:

[2, 4, 6, 8, 10]

$$\text{Mean} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

Calculate the Squared Differences from the Mean

Value	Difference from Mean	Squared Difference
2	$2 - 6 = -4$	$(-4)^2 = 16$
4	$4 - 6 = -2$	$(-2)^2 = 4$
6	$6 - 6 = 0$	$0^2 = 0$
8	$8 - 6 = 2$	$2^2 = 4$
10	$10 - 6 = 4$	$4^2 = 16$

MEASURES OF VARIABILITY

Dataset:

[2, 4, 6, 8, 10]

Calculate Variance

Variance is the **average of the squared differences**:

Value	Difference from Mean	Squared Difference
2	$2 - 6 = -4$	$(-4)^2 = 16$
4	$4 - 6 = -2$	$(-2)^2 = 4$
6	$6 - 6 = 0$	$0^2 = 0$
8	$8 - 6 = 2$	$2^2 = 4$
10	$10 - 6 = 4$	$4^2 = 16$

$$\text{Variance} = \frac{16 + 4 + 0 + 4 + 16}{5} = \frac{40}{5} = 8$$

MEASURES OF VARIABILITY

Dataset:

[2, 4, 6, 8, 10]

Calculate Standard Deviation

Standard Deviation is the square root of the variance:

$$\text{Variance} = \frac{16 + 4 + 0 + 4 + 16}{5} = \frac{40}{5} = 8$$

$$\text{Standard Deviation} = \sqrt{8} \approx 2.83$$

MEASURES OF VARIABILITY

Dataset:

[2, 4, 6, 8, 10]

```
import math
# Dataset
data = [2, 4, 6, 8, 10]
# Step 1: Calculate the mean
mean = sum(data) / len(data)
# Step 2: Calculate squared differences from the mean
squared_diffs = [(x - mean) ** 2 for x in data]
# Step 3: Calculate variance
variance = sum(squared_diffs) / len(data) # Population variance
# Step 4: Calculate standard deviation
std_dev = math.sqrt(variance)
# Output the results
print(f"Data: {data}")
print(f"Mean: {mean}")
print(f"Squared Differences: {squared_diffs}")
print(f"Variance: {variance}")
print(f"Standard Deviation: {std_dev}")
```

USING EXCEL FOR STATISTICS

- Enter your data in a column.
- Use functions:
 - =AVERAGE(A1:A5) → Mean
 - =MEDIAN(A1:A5) → Median
 - =MODE.SNGL(A1:A5) → Mode
 - =STDEV.P(A1:A5) → Standard Deviation
 - =VAR.P(A1:A5) → Variance

Activity - Analyzing Cybersecurity Incident Data

- Calculate key statistical measures (Mean, Median, Mode, Variance, Standard Deviation)
- Interpret and compare statistical results
- Discuss implications for cybersecurity decision-making based on data

Activity - Analyzing Cybersecurity Incident Data

You are part of a Security Operations Center (SOC) team. The team tracks the number of detected cybersecurity incidents across five departments of an organization over the past month. You've been asked to analyze this data and summarize trends to help prioritize future security efforts.

Activity - Analyzing Cybersecurity Incident Data

Sample Dataset (To be entered into Excel)

Enter the incident numbers into cells A1:A5 in Excel.

Department	Number of Detected Incidents
HR	12
Finance	18
IT	25
Legal	15
Marketing	10

Activity - Analyzing Cybersecurity Incident Data

Sample Dataset (To be entered into Excel)

Department	Number of Detected Incidents
HR	12
Finance	18
IT	25
Legal	15
Marketing	10

Activity - Analyzing Cybersecurity Incident Data

Sample Dataset (To be entered into Excel)

Use Excel formulas to calculate:

=AVERAGE(B2:B6)

=MEDIAN(B2:B6)

=MODE.SNGL(B2:B6)

=STDEV.P(B2:B6)

=VAR.P(B2:B6)

=MAX(B2:B6)-MIN(B2:B6)

Department	Number of Detected Incidents
HR	12
Finance	18
IT	25
Legal	15
Marketing	10

Activity - Analyzing Cybersecurity Incident Data

Sample Dataset (To be entered into Excel)

Use Excel formulas to calculate:

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Activity - Analyzing Cybersecurity Incident Data

Discussion

Which department has the highest number of incidents?
What could explain this?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Activity - Analyzing Cybersecurity Incident Data

Discussion

Which department has the highest number of incidents?
What could explain this?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

•IT has the highest number of incidents (**25**).

Possible explanations:

- IT systems are often more complex and exposed to internal and external threats.
- They manage core infrastructure, making them a frequent target.
- Higher visibility and logging may result in more detected incidents.

Activity - Analyzing Cybersecurity Incident Data

Discussion

How do the Mean and Median compare? What does this tell you about the distribution?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Mean = 16

Median = 15

Since the **mean is slightly higher than the median**, this suggests a **slight positive skew** — possibly caused by the high value from the IT department.

The data is **fairly symmetric**, with only one high outlier (25) pulling the mean upward.
Mean > Median

Activity - Analyzing Cybersecurity Incident Data

Discussion

How do the Mean and Median compare? What does this tell you about the distribution?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Inference:

The average (mean) of 16 is slightly higher than the middle value (median) of 15, indicating a slight positive skew in the data. This skew is likely due to a higher value of 25 coming from the IT department, which pulls the overall average upward.

Most departments have values close to the median, showing fairly balanced contributions. However, the IT department's unusually high number suggests it stands out compared to others, affecting the overall distribution.

Activity - Analyzing Cybersecurity Incident Data

Discussion

Why is there no mode in this dataset? What does that suggest?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

There is **no mode** because **all values are unique**.

This suggests that incident counts are spread out and there's **no recurring frequency**, indicating **diverse threat levels** across departments.

Activity - Analyzing Cybersecurity Incident Data

Discussion

Why is there no mode in this dataset? What does that suggest?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

If there is a **mode** in the dataset, then one or more incident counts repeat across departments. This suggests that some departments face **similar threat levels or incident frequencies**, highlighting common areas of concern that may require focused attention

Activity - Analyzing Cybersecurity Incident Data

Discussion

Is the spread (Standard Deviation) of incidents high or low? Should certain departments be prioritized?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Standard Deviation ≈ 5.02 , which is moderately high considering the mean is 16.

Indicates a noticeable spread in incident frequency.

IT (25) and Finance (18) should be prioritized for:

Detailed investigation, Policy reviews, Awareness training

Marketing and HR may still face risk but show fewer incidents — possibly due to lower exposure or under-reporting.

Activity - Analyzing Cybersecurity Incident Data

Discussion

How would these insights help you recommend resource allocation for the next quarter?

Department	Number of Detected Incidents			Measure	Formula (Assuming B2:B6)
HR	12			Mean	16
Finance	18			Median	15
IT	25			Mode	#N/A
Legal	15			Standard Deviation	5.25357021
Marketing	10			Variance	27.6
				Range	15

Allocate more security resources to IT and Finance (e.g., monitoring tools, analyst time, staff training).

Review controls in departments with low incidents to ensure they're not missing detections.

Consider tailored awareness campaigns based on department-specific risk profiles.

Use data-driven approaches to justify budget requests and optimize team efforts.

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

1. Introduction

This section presents a statistical analysis of detected cybersecurity incidents across five departments: HR, Finance, IT, Legal, and Marketing. The purpose of this analysis is to identify which departments experience higher or lower threat levels and assess the consistency of incident detection. Basic descriptive statistics such as mean, median, mode, standard deviation, and range were used to interpret the data and support data-driven decision-making in security planning.

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

2. Dataset Summary

Department	Number of Detected Incidents
HR	12
Finance	18
IT	25
Legal	15
Marketing	10

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

3. Statistical Measures

Measure	Value
Mean	16
Median	15
Mode	N/A
Standard Deviation (SD)	≈ 5.02
Variance	25.2
Range	15

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

4. Interpretation of Results

•Highest Incident Count:

The **IT department** reported the most incidents (25). This is expected due to its responsibility for managing critical infrastructure and exposure to various systems and networks. High incident numbers may reflect either increased attack surface or better detection capabilities.

•Mean vs. Median:

The **mean (16)** is slightly higher than the **median (15)**, suggesting a **slight positive skew**. This indicates that most departments have incident counts near the middle, with IT's high count pulling the average upward.

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

4. Interpretation of Results

- **No Mode Present:**

Since all values are unique, there is **no mode**, indicating a **wide distribution** without repetition. This reflects varied threat exposure or detection efficiency across departments.

- **Spread of Data (Standard Deviation):**

The **standard deviation of ≈ 5.02** is moderately high relative to the mean. This implies **noticeable variation** among department incident counts. The **range of 15** (from 10 to 25) further reinforces this spread.

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

5. Implications for Cybersecurity Strategy

Based on the analysis:

IT and Finance departments should be prioritized for additional security measures, including:

- Continuous monitoring

- Patch and configuration audits

- Department-specific awareness training

Marketing and HR, despite lower incident numbers, should be evaluated for possible **under-reporting** or lack of visibility.

Data insights justify resource reallocation and the development of **risk-informed policies**.

The use of descriptive statistics supports evidence-based decisions and promotes a **proactive security posture**.

Activity - Analyzing Cybersecurity Incident Data

- Descriptive Analysis of Detected Cybersecurity Incidents Across Departments

6. Conclusion

Descriptive statistical analysis offers critical insights into incident patterns across departments. By identifying departments with higher risk and variability, organizations can strategically allocate security resources, enhance monitoring, and develop tailored risk mitigation plans. This simple dataset illustrates the power of foundational data analysis in real-world cybersecurity operations.

Activity - Analysis of Weekly Cybersecurity Alerts

- **SOC Alert Volume Analysis**
- You are working as a junior analyst in a national **Cybersecurity Operations Center (SOC)**. Over the last 8 weeks, your team has logged the number of cybersecurity alerts triggered by suspicious activities in 10 different client networks. You've been asked to analyze the alert patterns to understand system vulnerabilities and plan proactive defense strategies.

<https://forms.office.com/e/BJFyAwkLLA>

