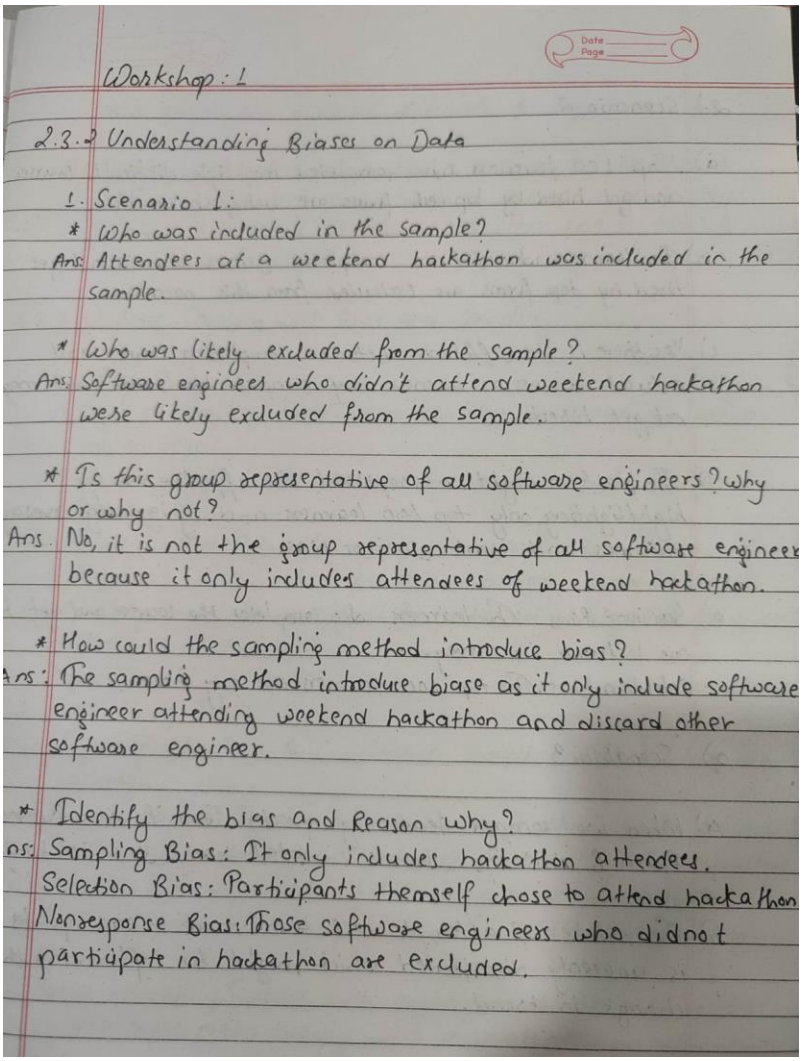
Worksheet:1



Workshop : 1

2.3.2 Understanding Biases on Data

1. Scenario 1:

* Who was included in the sample?

Ans: Attendees of a weekend hackathon was included in the sample.

* Who was likely excluded from the sample?

Ans: Software engineers who didn't attend weekend hackathon were likely excluded from the sample.

* Is this group representative of all software engineers? why or why not?

Ans: No, it is not the group representative of all software engineer because it only includes attendees of weekend hackathon.

* How could the sampling method introduce bias?

Ans: The sampling method introduce biase as it only include software engineer attending weekend hackathon and discard other software engineer.

* Identify the bias and Reason why?

ns: Sampling Bias: It only includes hackathon attendees.
Selection Bias: Participants themself chose to attend hackathon
Nonresponse Bias: Those software engineers who did not participate in hackathon are excluded.

2.) Scenario 2

a) Top 100 learners who completed multiple difficult courses and got hired by top tech firms are being shown here.

b) The learners who did not completed courses and did not get hired by top firms are excluded from this narrative.

c) Yes, there is the difference between completing course and actual success as learner may complete the course but may not get hired.

d) The platform might misrepresent the effectiveness by highlighting only top 100 learners and by ignoring average outcomes and create survivor bias.

e) Survivor Bias: The learners who completes the course and get hired are included in survey.
Selection Bias: The top 100 learners are only chosen.

3) Scenario 3

a) When we compare teams by quater. Team A wins Q_1 and Team B wins Q_2 whereas in total team B wins in total.

b) The number of attempts each team made in each quater is unevenly distributed which might be influencing the change in trend.

c) This apparent reversal may happen due to uneven

each quaters and different difficulty of bugs.

d) Simpson's paradox occurs when a trend appears in several difficult groups are combined. So, in this case also, in each quaters, one team outperforms the other but in total B seems to perform better.

2.3.3 Critical Thinking with Bias and sampling:

a. Case Study 1

i) Type of Sampling: Convenience sampling as users who renewed are surveyed.

ii) No, the sample is not representative as it excludes all non-renewing users.

iii) Selection Bias: Only the active users are chosen.

iv) Survivour Bias: user who renewed are only ~~survived~~ surveyed.

v) The study could be redesigned for better reliability: including both active and inactive users.

b. Case Study 2

i) Type of Sampling: Self-selection sampling via social media hashtags.

ii) No, the sample is not representative as it only includes people who use twitter and post about their jobs with positive hashtags.

iii) Selection Bias: only people who are using twitter are included
Positive Recording Bias: people who post with positive hashtags are included.

iv) Better reliability: Include both negative and positive response.
Combine social media analysis with more objective data.

c) Case study 3

i) Type of Sampling: Voluntary response sampling (only limited to fitness app users)

ii) No, the sample is not representative as it only includes health conscious individual

iii) Selection Bias: Only people using fitness app are included
Survivor Bias: People who maintain fitness habits and stay on app

iv) Redesigned for better reliability: collect the data from general public.

2.3.4 Survey Design, Bias and Reflection

a) Spot the flaws:-

→ Three flaws in the sampling.

i) Survey only include student who attended mindfulness workshop

ii) Leading question as it lack neutrality and doesnot allow for negative or no-effect response

iii) It doesnot include the data of student who did not attend workshop.

→ Propose improvement.

i) Sampling Strategy: use random sampling across the entire student population

ii) Question Neutrality: replacing lead question with a neutral, open-ended and Likert questions. For eg: "Have you noticed any change in your mental health after attending workshop."

iii) No, this survey's findings are not generalizable as it only included data of student who attended workshop.

b) Redesign a Biased Survey

i) Response Bias: employee may fear being identified or judged.
Question Bias: leading language which encourage positive affirmation and ~~surp~~ supress honest criticism.
Non-response Bias: Employee less engaged or skeptical of HR may not participate.

ii) "How ~~likely~~ satisfied are you with our current flexible working policies?" of likert-scale?

ii) More inclusive and anonymous sampling method.
* Send survey directly through emails.
* Enable anonymous mode internally.

iii) * Ensure highes response rate and representation
* Make survey shorter and neutral
* Send reminders to fill up the survey
* Take action as per the response

c) Survey Ethics and Consent:

i) You are invited to participate in a voluntary survey about student's use of AI tools in academic assessments. Your responses will be used for research purposes only.
Your response will be kept confidential. No personally identifying information will be collected. If you agree to participate, please proceed with the survey. If you have any queries, feel free to ask us. By continuing, you indicate that you understand this information and consent to participate.

ii) Ethical Consideration

a) Confidentiality and Privacy
b) Voluntary participation
c) Minimizing risk and sensitivity

iii) For this survey, anonymization is more appropriate, because it protects student privacy fully and encourage honest response without fear.

iv) Your identity/role may affect participation as they might fear consequence and judgement.

→ To mitigate

i) Use anonymized survey tools

ii) Consider having neutral facilitator send out the survey

d) Designing a survey for policy insights:

i. sampling strategy: multi model and inclusive sampling approach

ii) Questions

a) How do you primarily commute on weekdays?

→ Walk, private car, public bus, bicycle, Sharing ride others

b) How satisfied are you with the current public transport options in your area?

→ (1-5) likert scale (very dissatisfied - Very satisfied)

c) Rate your ~~agg~~ agreement with this statement.

"I feel safe cycling in my city"

→ Strongly Disagree → Strongly Agree (1-5)

d) Please rank the following in order of priority for improvement (1 = Most Important).

* Safer Bicycle lane
* More frequent Buses
* Better night time service
* lower fares
* Bike sharing availability

e) What is one thing you would change to improve your daily commute? (open-ended)

iii) Minimize and selection Bias

a) Using offline and online method

b) Ensuring sample across different areas

c) Ensure anonymity and no consiquence
d) Keep survey short and avoid loading question.

iv) Analyzing Result
a. Quantative
b Qualitative
c. Geographical analysis
- Policy impact use
a. Seperate budget allocation on transportation as need.
b Track changes over time
c. Promote low-cost transportation options to students

3. Describing the Data Descriptive Statistic A Review.

1. Sports Performance

Players Heights (ft)

5.1 | 4.8 | 5.0 | 4.9 | 5.3 | 4.7 | 4.9 | 4.8 | 5.2 | 5.0 |

Mean ($\bar{x}$) = $\frac{\Sigma x_i}{n}$

$= \frac{5.1 + 4.8 + 5.0 + 4.9 + 5.3 + 4.7 + 4.9 + 4.8 + 5.2 + 5.0}{10}$

$= 49.7/10$

$= 4.97$

x	$(x-\bar{x})$	$(x_i-\bar{x})^2$
5.1	0.13	0.0169
4.8	-0.17	0.0289
5.0	0.03	0.0009
4.9	-0.07	0.0049
5.3	0.33	0.1089
4.7	-0.27	0.0729
4.9	-0.07	0.0049
4.8	-0.17	0.0289
5.2	0.23	0.0529
5.0	0.03	0.0009

$\Sigma(x_i-\bar{x})^2 = 0.321$

Standard Deviation $(s) = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$

$= \sqrt{\frac{0.321}{9}}$

$= 0.189.$

coefficient of variation (CV)

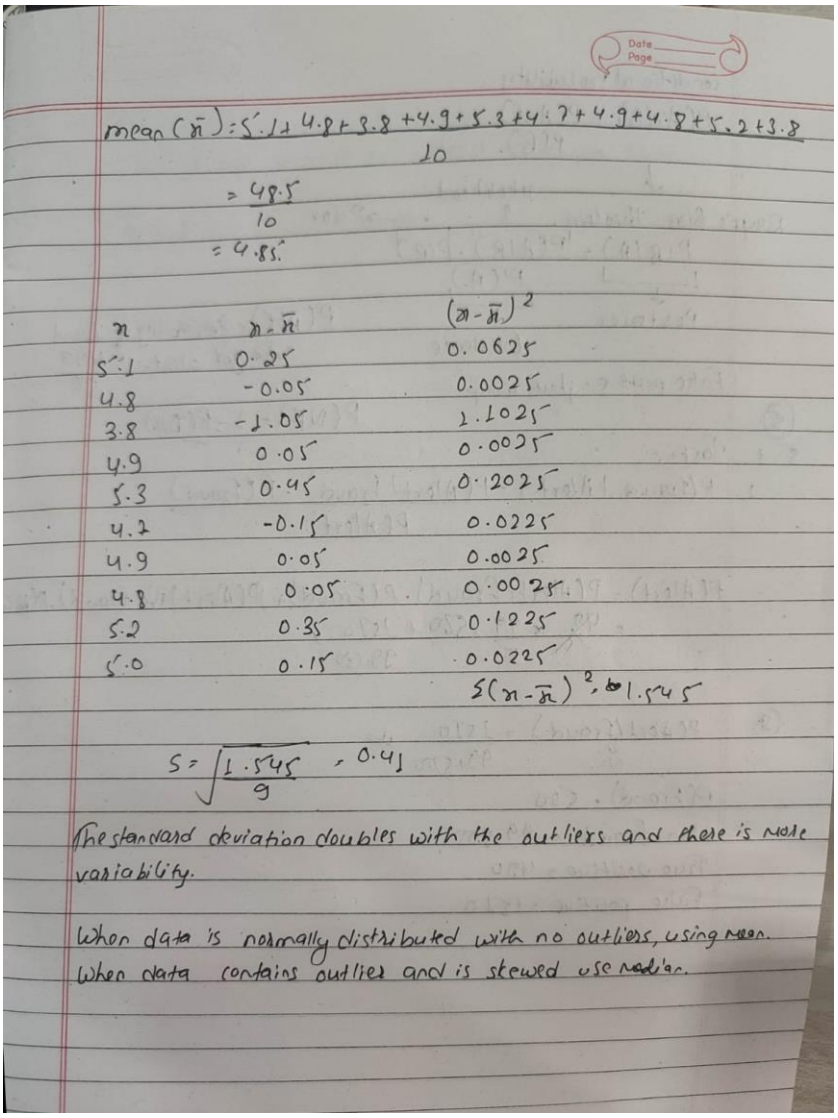
$(CV) = \frac{s}{\bar{x}} \times 100$

$= \frac{0.189}{4.97} \times 100$

$= 3.8\%.$

Since CV is low (3.8%), so times are tightly clustered.

Replacing 5.0 with 3.8.



$$\text{mean}(\bar{x}) = \frac{5.1 + 4.8 + 3.8 + 4.9 + 5.3 + 4.7 + 4.9 + 4.8 + 5.2 + 3.8}{10}$$

$$= \frac{48.5}{10}$$

$$= 4.85$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
5.1	0.25	0.0625
4.8	-0.05	0.0025
3.8	-1.05	1.1025
4.9	0.05	0.0025
5.3	0.45	0.2025
4.7	-0.15	0.0225
4.9	0.05	0.0025
4.8	0.05	0.0025
5.2	0.35	0.1225
5.0	0.15	0.0225
		$\sum(x - \bar{x})^2 = 1.545$

$$s = \sqrt{\frac{1.545}{9}} = 0.41$$

The standard deviation doubles with the outliers and there is more variability.

When data is normally distributed with no outliers, using mean.
When data contains outliers and is skewed use median.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

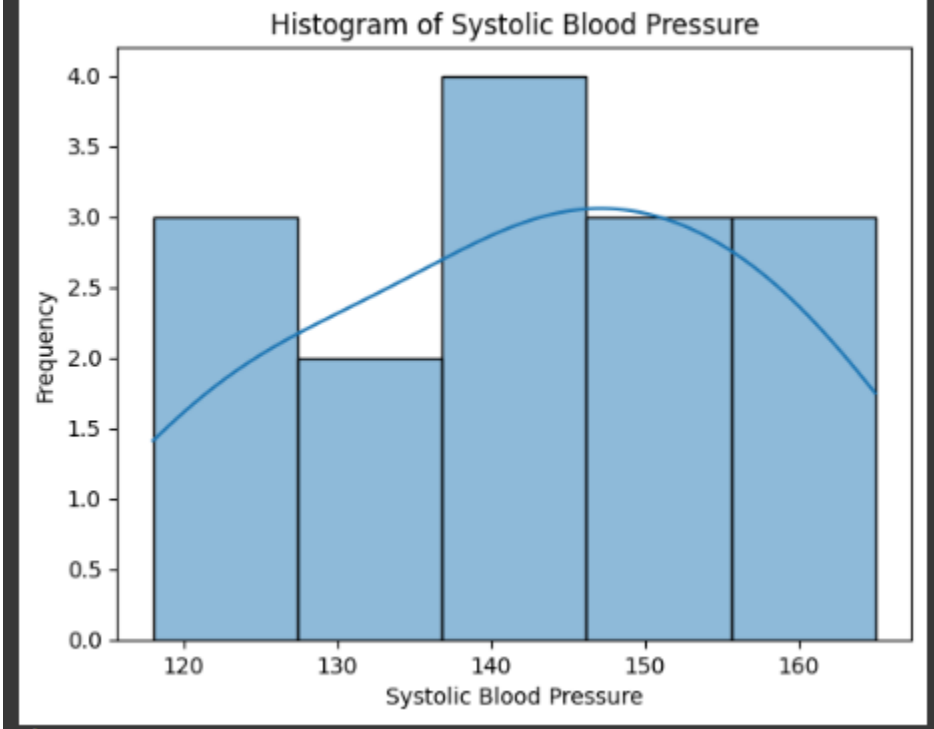
Mean, Median and Standard Deviation

```
num=np.array([118,122,125,130,135,138,142,144,146,150,152,155,160,162,165])
mean=np.mean(num)
median=np.median(num)
std=np.std(num)
print(f"mean:{mean:.2f}")
print(f"median:{median}")
print(f"standard deviation: {std:.2f}")
```

```
mean:142.93
median:144.0
standard deviation: 14.30
```

Plot a histogram. Is the distribution skewed?

```
[ ] sns.histplot(num,kde=True)
    plt.title("Histogram of Systolic Blood Pressure")
    plt.xlabel("Systolic Blood Pressure")
    plt.ylabel("Frequency")
    plt.show()
    skewness=stats.skew(num)
    print(f"Skewness: {skewness:.2f}")
```




```
Skewness: -0.18
```

Compute the IQR. Identify any patients with unusually high pressure

```
q1=np.percentile(num,25)
q3=np.percentile(num,75)
IQR=q3-q1
upp_bound=q3 + 1.5 * IQR
low_bound=q1 - 1.5 * IQR
#outliers
outliers=num[(num>upp_bound)| (num<low_bound)]
print(f"IQR:{IQR:.2f}")
print(f"outliers:{outliers}")
```

```
IQR:21.00
outliers:[]
```

Explain whether these statistics support that the clinic population has a normal range of BP levels.

The distribution appears to be symmetrical, as the mean is approximately equal to the median. The skewness is close to zero, indicating minimal skewness. Additionally, the histogram suggests a normal distribution, as there is no evidence of extreme skew. The absence of outliers further supports the assumption of normality.

Mean, Median, range, mode , variance, standard deviation

```
[ ] sales=[212,198,245,210,230,185,270,205,190,250,260,225,215,195]
    mean=np.mean(sales)
    median=np.median(sales)
    sales_range=np.max(sales)-np.min(sales)
    sales_mode=stats.mode(sales).mode
    variance=np.var(sales)
    std=np.std(sales)
    print(f"mean:{mean:.2f}")
    print(f"median:{median}")
    print(f"mode:{sales_mode}")
    print(f"range:{sales_range}")
    print(f"variance:{variance:.2f}")
    print(f"standard deviation:{std:.2f}")
```

```
mean:220.71
median:213.5
mode:185
range:85
variance:670.78
standard deviation:25.90
```

```python
days=[f'Day{i+1}'for i in range(len(sales))]
#Bar Chart
plt.figure(figsize=(10, 5))
bars=plt.bar(days,sales)
plt.title("Daily Sales")
plt.xlabel("Days")
plt.ylabel("Sales")

#Annote highest and lowest sales
max_idx=np.argmax(sales)
min_idx=np.argmin(sales)
plt.text(max_idx,sales[max_idx]+5,f"High:{sales[max_idx]}",ha='center',color='green')
plt.text(min_idx,sales[min_idx]-10,f"Low:{sales[min_idx]}",ha='center',color='red')
bars[max_idx].set_color('green')
bars[min_idx].set_color('red')

plt.tight_layout()
plt.show()
```

Comment on consistency of sales - do the spread and measures indicate a steady flow?

```python
#Sample lower days
actual_sundays=[185,190]
#Reverse the 20% dip
adjusted_sundays=[185/0.8,190/0.8]

sales_adj=sales.copy()
sales_adj[5]=adjusted_sundays[0]
sales_adj[8]=adjusted_sundays[1]
print("Adjusted Mean:",np.mean(sales_adj))
print("Adjusted Std Dev:",np.std(sales_adj,ddof=1))
```

```
Adjusted Mean: 227.41071428571428
Adjusted Std Dev: 23.100034786615737
```

4 Descriptive - {Numerical and Graphical} Analysis of Data.

4.1 Advanced Case Studies - Numerical Summary: For the following task, please feel free to use Python programming and any library that you find suitable.

1. Case 1 - Dropout Risk Assessment:

- **Context:** A University program is concerned about students dropping out in their first year. You are given GPA scores of 120 first - year students and their dropout status.

Status	GPA Mean	GPA Std. Dev	n
Dropped Out	2.1	0.6	30
Retained	3.1	0.5	90

Table 5: GPA Statistics with Decimal Alignment

Compute and compare the coefficient of variation (CV) for both groups

```python
#Coefficient of variation
dropout_cv=(0.6/2.1)*100
retained_cv=(0.5/3.1)*100
print(f"Dropout CV:{dropout_cv:.2f}%")
print(f"Retained CV:{retained_cv:.2f}%")
```

```
Dropout CV:28.57%
Retained CV:16.13%
```