Foundations of Statistical Thinking: Understanding Data. Applying Statistical Thinking to Real - World Data.

Prepared By: Siman Giri, Instructor: Ronit and Shiv for Herald Center for AI.

Summer, 2025

1 Learning Objectives.

- Focus: Data Collection Methods, Sampling Techniques, and Statistical Biases
- Skills Targeted: Critical Evaluation, Real - World Application, Ethical Considerations.
- By the end of this worksheet, you will be able to:
  - Distinguish between qualitative vs. quantitative and discrete vs. continuous data.
  - Compare primary and secondary data collection methods and their uses.
  - Identify common biases in real - world examples.
  - Evaluate sampling techniques for potential flaws.
  - Design a bias resistant data collection plan for a given scenario.

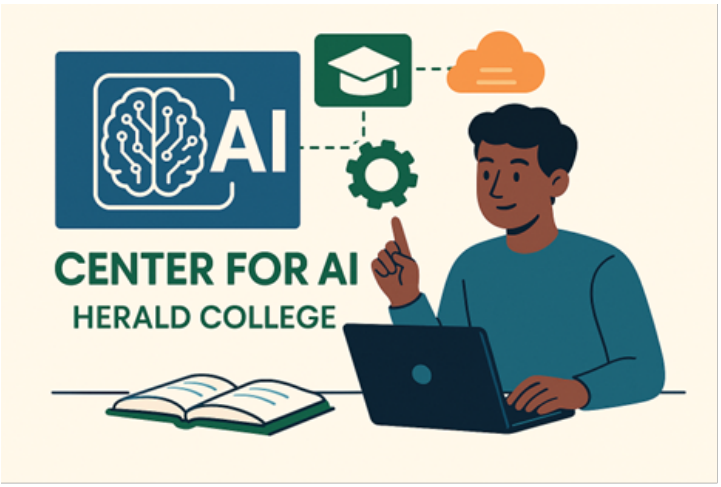

image generated via copilot.

2 Methods of Data Collection.

2.1 Exploring Observational Studies:

Instructions:
Read each case below. Decide whether it's a form of observational study. If it seems like an experiment, just write "Experimental – to be discussed later." For observational cases, explain what type it is and any possible limitations or biases.

To - Do:

- **Select** the better method {Survey or Secondary Sources}.
- **Justify** your choice with 2 - 3 specific reasons.
- **Identify** one potential limitation of your chosen approach.

Case 1: Healthcare Policy Evaluation:

The Ministry of Health wants to understand why vaccination rates for measles dropped 15% last year in rural areas.

- **Available resources:**
  - 5 researchers for 3 months.
  - Access to national health records.
  - Budget for 500 in - person interviews.
- **Task:** Which method would provide more actionable insights? What key variables would you prioritize?

Case 2: Corporate Diversity Audit:

A tech company needs to analyze gender representation in leadership roles across its 20 global offices over 5 years.

- **Constraints:**
  - Legal restrictions on collecting new demographic data in 3 countries.
  - Existing HR databases with promotion records.
  - Employee resistance to internal surveys.
- **Task:** How would you balance data completeness with legal/ethical constraints?

Case 3: Urban Planning Challenge:

A city wants to assess public transportation usage patterns after introducing free subway passes for seniors.

- **Available Data:**
  - Smart card swipe records (age - group tagged).
  - Previous year's rider satisfaction survey.
  - Budget for follow - up focus groups.
- **Task:** What blended approach could validate findings while minimizing bias?

Case 4: Education Reform:

A school district needs to compare STEM enrollment trends (2015-2025) with local job market demands.

- **Limitations:**
  - Student record lack career intent data.
  - Province labor statistics have 2 - year lag.
  - Parent surveys have 30% response rate historically.
- **Task:** How would you address the temporal mismatch between education and labor data?

Case 5: GitHub Productivity Study:

A tech research team analyzes developer productivity patterns using public commit data.

- **Available Data:**
  - Timestamped GitHub commit logs.
  - Project metadata (programming languages, team size).
  - Optional: Developer demographic survey budget.
- **Task:** What blended approach could validate productivity findings while minimizing bias?

2.2 Designing your Own Survey.

✅ Step	Description
🎯 **Clear Objective**	Do you have a focused research question?
🧍 **Target Audience**	Do you know who the survey is meant for?
✍️ **Simple Language**	Are your questions free of jargon and bias?
⚖️ **Balanced Options**	Are all possible responses fairly represented?
🔢 **Appropriate Scales**	Are you using the right scale (e.g., Likert, ratio)?
🧪 **Pilot Tested**	Have you tested your survey on a small group?
🔐 **Ethical Considerations**	Are you transparent about anonymity, consent, and usage?
📱 **Accessible Format**	Is it easy to access on phones/laptops?
🎯 **Relevant & Short**	Will it take <5 mins to complete and stay on topic?
📈 **Easy to Analyze**	Will the answers be easy to quantify or interpret?

Figure 1: Survey Design Checklist.

A Sample Survey:

Case Study: Programming Habits of First-Year Students.

You are part of a team studying how first-year CS students learn programming.

- **Background Scenario:** Your department is redesigning the introductory programming course. The team wants to know:
  - What tools and resources do students use?
  - How confident are they?
  - What are their challenges?
- A Sample Survey
  `https://docs.google.com/forms/d/1V7h_NZGmdkKR9r41DaDWfEze4eAu25cMh7pVkyYrVBM/edit`

Task 1: Survey Design Review:

Based on the earlier review, answer the following question:

1. What is the objective of this survey?
2. Who is the target population?
3. List 2 closed-ended and 1 open-ended question from the survey.
4. Do the questions avoid bias or leading phrasing? Explain.
5. How would you improve this survey (at least 1 suggestion)?

Task 2: Discuss - Is is this a Good Survey?

Some Discussion Points:

- Why is this a Good Survey?
  1. Clear Target and Purpose:
     - The objective is well-defined – to understand first-year programming experiences.
     - Instructor Note: Emphasize the importance of aligning questions with the objective.
  2. Mix of Question Types:
     - Includes categorical, ordinal, Likert-scale, and open-ended questions.
     - Instructor Note: Discuss how this allows both quantitative and qualitative analysis.
  3. Logical Flow and Grouping:
     - Questions are grouped into thematic sections – background, experience, challenges, and feedback.
     - Instructor Note: Ask Students why flow matters for response quality and clarity.
  4. Student-Centric Language:
     - Language is simple and relatable to the target audience.
     - Instructor Note: Explain how tone and wording can affect response rates and accuracy.

Task 3: Design Your Own Survey:

Design a 5 - 7 question survey around one of the topics below:

- **Choose a Topic:**
  1. Screen time vs. grades.
  2. Online vs. offline learning preferences.
  3. Use of AI tools (e.g. ChatGPT) in coursework.
  4. Coding habits during final exams.
  5. Social Media's impact on focus.

Hint - Step by Step Guide:

1. **Ask a Good Question.**
2. **Define What to Measure.**
3. **Plan Your Survey.**
   - Format: Google Form.
   - Target: Second year CS students.
   - Time to Complete: < 5 minutes.
4. **Pilot Your Survey.**

- Ask 3 peers to take it and give feedback on clarity.

5. **Ethical Considerations.**
   - Ensure responses are anonymous.
   - Ask for Consent.

Task 4: Collect and Reflect:

Instructions:

- Send your survey to at least 10 respondents.
- Collect responses and prepare a short reflection with basic data analysis.
- Hint for Reflection.
  1. How did your respondents react?
  2. Were any questions confusing or skipped.
  3. Did the data surprise you?
  4. What would you change if you re - ran this survey?
  5. Use Descriptive Statistics and Basic Visualization to support your reflection.

2.3 Decode the Dataset:

Objective:

You are expected to critically evaluate the open - source datasets to identify:

- The population and sample
- Sampling method used or likely used
- Potential biases and limitations

2.3.1 In class Poll or Quiz:

1. A dataset of 1000 tweets collected to study hate speech on Twitter. What is the population in this scenario?
   (a) All social media platforms.
   (b) Twitter users.
   (c) 1000 tweets.
   (d) Users of Facebook and Twitter.
2. A Kaggle dataset includes job salaries submitted by data scientists voluntarily. What kind of bias might this dataset suffer from?
   (a) Selection bias
   (b) Non-response bias
   (c) Self-selection bias
   (d) Sampling bias
3. Which sampling technique is most appropriate if you want to survey CS students from each academic year?
   (a) Simple Random Sampling.
   (b) Stratified Sampling.
   (c) Cluster Sampling.
   (d) Systematics Sampling.

4. Which of the following is NOT a challenge in using open - source datasets?
   (a) Lack of documentation
   (b) Free access
   (c) Sampling bias
   (d) Missing values

5. You want to study online learning behavior by emailing a questionnaire to 200 randomly chosen students from a university database. What kind of sampling is this?
   (a) Convenience
   (b) Systematic
   (c) Simple Random
   (d) Stratified

2.3.2 Understanding Biases in Data:

Think, Reflect and Discuss:

1. **Scenario 1:** A tech company surveys attendees at a weekend hackathon to understand how often software engineers work on open-source projects. Based on the results, they report that "most software engineers regularly contribute to open-source software."
   Think & Reflect:
   - **Think & Reflect:**
     - Who was included in the sample?
     - Who was likely excluded from this sample?
     - Is this group representative of all software engineers? Why or why not?
     - How could the sampling method introduce bias?
     - Identify the Biases and Reason why?

2. **Scenario 2:** An online course platform highlights its "Top 100 Learners" who completed multiple difficult courses and got hired by top tech firms. They claim their platform is the fastest way to get into a high-paying tech job.
   - **Think & Reflect:**
     - Who are being showcased here?
     - Who might have been excluded from this narrative?
     - Is there a difference between completing courses and actual success?
     - How might the platform be misrepresenting the overall effectiveness?
     - Identify the Biases and Reason why?

3. **Scenario 3:** Two teams of developers are evaluated on their bug-fixing performance over two quarters. Individually, each team had a higher success rate in one quarter. But when their performances were aggregated, the team that underperformed in both quarters suddenly appeared to outperform the other.

Table 1: Team Performance by Quarter

	Team A (Fixed/Total)	**Team B (Fixed/Total)**
Q1	45/50	30/35
Q2	20/80	60/90
Total	65/130	90/125

- **Think & Reflect:**
  - What happens when you compare teams by quarter vs in total?
  - What other factors might be influencing the change in trend?
  - Why does this apparent reversal happen?
  - Can you explain why this situation is a classic Simpson's Paradox?

2.3.3 Critical Thinking with Bias and Sampling:

Instructions: For each of the following case studies, identify:

- **The type of sampling used:**
- **Whether the sample is representative:**
- **Any biases present (Selection, Survivor, Simpson's Paradox, etc.):**
- **How the study could be redesigned for better reliability:**

1. **Case Study 1:** A tech company sends a feedback form only to users who have renewed their subscriptions in the past year. Based on responses, they claim 90% of their users are satisfied.
2. **Case Study 2:** A journalist analyzes job satisfaction from Twitter posts using hashtags like #lovemyjob and concludes job satisfaction is increasing in 2025.
3. **Case Study 3:** An insurance agency concludes that people who exercise regularly are less likely to make claims. However, the data used only comes from a fitness-tracking app's users.

2.3.4 Survey Design, Biases and Reflection:

1. **Spot the Flaws:**
   - **Scenario:** A university wants to assess mental health among students. They send out a survey only to students who attended a recent mindfulness workshop, asking:
     - *"How much has mindfulness helped improve your mental health?"*
     - Responses were **overwhelmingly** positive.
   - **Tasks:**
     - Identify at least three flaws in the sampling and question framing.
     - Propose improvements to the survey's:
       * Sampling Strategy
       * Question Neutrality
     - Discuss whether this survey's findings are generalizable.
2. **Redesign a Biased Survey:**
   - **Scenario:** A tech company asks employees:
     - *" Do you agree that our flexible working policies are excellent?"*
     - The Survey is voluntary and only visible on the internal HR portal.
   - **Tasks:**
     - Identify sources of response bias, question bias, and non-response bias.
     - Rewrite the question to remove leading language.
     - Propose a more inclusive and anonymous sampling method.
     - Describe how you'd ensure higher response rate and representation.
3. **Survey Ethics and Consent:**
   - **Scenario:** You're conducting a survey on students' use of AI tools in assessments.
   - **Tasks:**
     - Draft a short informed consent statement to appear at the start of the survey.
     - Identify 2 - 3 ethical considerations in collecting and storing this data.
     - Discuss whether anonymization or pseudonymization is more appropriate, and why?
     - Reflect: How might your own identity/role influence survey participation or response honesty?

4. **Designing a Survey for Policy Insight:**

   - **Scenario:** You are hired by a city government to survey young adults (age 18–25) about their views on public transport and cycling infrastructure.
   - **Tasks:**
     - Define your sampling strategy: how will you ensure you reach working youth, students, and those without regular internet access?
     - Draft at least 5 survey questions that:
       * Include different formats (multiple choice, Likert, open-ended)
       * Are non-leading, clear, and accessible
     - Outline steps to minimize selection and response bias.
     - Propose how results should be analyzed and used to shape policy.

5. **Peer Review Survey Evaluation:**

   - **Tasks:**
     - Exchange your survey draft from above with a peer.
     - Evaluate their survey for{Provide Rating from 1(low) - 5(high)}:
       * Clarity of questions:
       * Potential Biases:
       * Appropriateness of sampling:
       * Ethical safeguards:
     - Write a short feedback note with 2 strengths and 2 suggestions.

3 Describing The Data - Descriptive Statistics A Review.

Complete the Tasks with **Pen** and on **Paper**:

1. **Sports Performance:**
   - **Context:** A football coach is analyzing players' sprint times (in seconds) over 40 meters.

Player Heights (ft)									
5.1	4.8	5.0	4.9	5.3	4.7	4.9	4.8	5.2	5.0

Table 2: Player Heights (in feet) with Full Borders

   - **Tasks:**
     - Compute:
       - Mean and Standard Deviation (sample):
       - Coefficient fo variation.
     - Are the player's times tightly clustered or highly variable?
     - If one player was later found to have mistakenly reported 3.8 seconds, recalculate and explain the impact of outliers.
     - Recommend whether mean or median should be used for performance reporting.
2. **Patient Blood Pressure Readings:**
   - **Context:** A health researcher is analyzing systolic blood pressure levels from a clinic.

Systolic Blood Pressure														
118	122	125	130	135	138	142	144	146	150	152	155	160	162	165

Table 3: Systolic Blood Pressure of 15 patients.

   - **Tasks:**
     - Compute the mean, median, and standard deviation.
     - Plot a histogram. Is the distribution skewed?
     - Compute the IQR. Identify any patients with unusually high blood pressure.
     - Explain whether these statistics support that the clinic population has a normal range of BP levels.

3. **Retail Sales Summary:**

- **Context:** A store tracks daily sales (Nrs.) over 2 weeks:

Stores Daily Sales													
212	198	245	210	230	185	270	205	190	250	260	225	215	195

Table 4: Daily Sales.

- **Tasks:**
  - Calculate: Mean, median, mode, range, variance, standard deviation
  - Draw a bar chart of daily sales and annotate any highs/lows
  - Comment on consistency of sales - do the spread and measures indicate a steady flow?
  - Suppose Sunday sales are usually 20% lower than the other days. How would this affect interpretation?

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

   - **Tasks:**
     - Compute and compare the coefficient of variation (CV) for both groups.
     - Interpret: Which group shows greater relative variability in GPA?
     - Suppose 5 of the 30 dropout GPAs were missing. How would that affect your analysis?
     - Discuss limitations of only using mean and standard deviation here—what's missing?
     - Sketch boxplots to compare GPA distributions between groups.

2. **Case 2 - Gender Pay Gap Investigation:**

   - **Context:** A tech company releases salary data (in $1000s) for 100 male and 80 female employees.

Gender	Mean	Median	SD
Male	105	98	18
Female	92	90	25

Table 6: Statistical Comparison by Gender

   - **Tasks:**
     - Interpret the difference between mean and median in both groups—are outliers likely?
     - Compute and compare IQRs if Q1 and Q3 for males are 90 and 110, and for females are 85 and 105.
     - Discuss which measure of central tendency best reflects typical salary for each group.
     - Suggest a visualization that could reveal more about potential pay gaps and justify your choice.
     - Consider the Simpson's Paradox—how might departmental breakdowns affect these results?

3. **Case 3 - Fitness Tracker Accuracy:**

   - **Context:** Two fitness trackers (Brand A and Brand B) record the number of steps per day for 15 users over 7 days. Below is the aggregated average and standard deviation per brand:

Tracker	Mean Steps	SD	Median	IQR
A	8050	310	8000	430
B	8250	800	8100	1100

   - **Tasks:**
     - Which tracker has more consistent measurements? Use CV to justify.
     - Why might Tracker B have higher mean but lower median?
     - Interpret how IQR and SD together inform the nature of variability.
     - A user claims Tracker B is "more optimistic." How would you statistically evaluate this claim?
     - Propose a statistical report layout for a consumer review site.

4. **Case 4 - Vaccine Response Data:**

   - **Context:** In a vaccine trial, antibody levels were measured in two age groups post-vaccination.

Age Group	Mean (AU/mL)	Median (AU/mL)	SD (AU/mL)	Min (AU/mL)	Max (AU/mL)
Under 40	840	800	140	500	1100
Over 60	620	610	85	480	800

Table 7: Antibody Levels by Age Group

   - **Tasks:**
     - Discuss which group shows more spread relative to their center.
     - Plot overlapping histograms or boxplots to visualize distribution differences.
     - Identify which summary statistics might be misleading if data is bimodal.
     - Suggest how the five-number summary could help identify immune outliers.
     - Consider ethical implications: Should dosage adjustments be made? Justify using descriptive analysis.

5. **Case 5 - E - Commerce Performance:**

   - **Context:** An e-commerce company tracks daily cart values (in $) for two types of users over 30 days:
   - **Tasks:**
     - Calculate and compare CVs.
     - Determine which group is more skewed and explain the implication.
     - Suppose a campaign increased new user spending on a few days. How would that affect measures?
     - Propose at least 2 descriptive statistics techniques for business insights here.

4.2 Advanced Case Studies - Graphical Summary:

1. **Case 1 - Daily Sales Trends in Two Product Categories:**

   - **Context:** A retail company is analyzing daily sales data (in USD) over 60 days for two product categories:
     - **Category A:** Higher volume but lower average price items
     - **Category B:** Premium items with fewer but larger transactions

     Each day has recorded total sales, average basket size, and number of transactions for both categories.
   - You are provided with a dataset: daily_sales.csv with columns:
     - `date, category, total_sales, avg_basket, num_transactions.`
   - **Tasks:**
     - Visualize the distribution of `total_sales` for each category using boxplots and histograms. Interpret central tendency and spread.
     - Create a time series plot of daily total sales. Are there visible trends or outliers?
     - Calculate Coefficient of Variation (CV) for `total_sales` and avg_basket for each category. Which category is more variable?
     - Compute and visualize a 7-day moving average for both categories. Discuss stability and implications for business planning.
     - **Discuss:** How would a flash sale or promotional campaign distort the mean? How can you account for it in visuals or summaries?

2. **Case 2 - Customer Spending by Segment:**

   - **Context:** You are analyzing customer transaction data over the last quarter to understand the behavior of two customer segments:
     - **Student.**
     - **Working Professionals.**
   - Each customer has multiple transactions. You are given `customer_segments.csv` dataset with columns:
     - `customer_id, segment, transaction_amount, date`
   - **Tasks:**
     - Create a violin plot and strip plot of `transaction_amount` for each segment. What can you infer about skewness, modality, and outliers?
     - Plot a density plot (KDE) of `transaction_amount` per segment. Which segment shows more variability? Which one has higher spending tendencies?
     - Compute mean, median, IQR, and CV for each group and display them in a summary table.
     - Construct a boxplot with swarm overlay to visualize typical vs. exceptional spending behavior. Do the visuals agree with summary statistics?
     - Create a scatter plot of number of transactions vs. total spending for each customer, color-coded by segment. What does the relationship suggest about loyalty or value?
     - **Discuss:** If you were to target one of the groups for a premium membership offer, what would guide your decision based on this analysis?

The - End