# The Art of Data Analysis and Reporting - Online

## Session 2: Introduction to Inferential Statistics (90 minutes, Synchronous)

| | |
|---|---|
| **Session Objectives** | ● Understand the concept of inferential statistics and explain how it differs from descriptive statistics.<br>● Explore the various sampling methods.<br>● Analyse the advantages and limitations of different sampling methods.<br>● Calculate and interpret confidence intervals. |
| **Key Points** | ● **Inferential Statistics:** Predicts population trends from a sample.<br>● **Descriptive Statistics:** Summarises and presents data.<br>● **Population:** The total group being studied.<br>● **Sample:** A subset of the population.<br>● **Sample Size:** The number in the sample.<br>● **Simple Random Sampling:** Everyone has an equal chance of selection.<br>● **Systematic Sampling:** Selects every nth individual after a random start.<br>● **Stratified Sampling:** Samples subgroups proportionally.<br>● **Cluster Sampling:** Surveys entire randomly selected clusters.<br>● **Convenience Sampling:** Chooses participants based on ease of access.<br>● **Confidence Interval:** Likely range for the true population value.<br>● **Bootstrapping:** Resampling with replacement to estimate variability. |
| **Assessment** | ● Learners are assessed through active questioning, chat responses, and hands-on coding activities to test their understanding of confidence intervals and sampling techniques. |
| **Instructor Prep** | ● Familiarise yourself with the following websites:<br>○ Introduction to Statistics (1.1)<br>○ What Are Is Descriptive And Inferential Statistics - What Are The Different Branches Of Statistics<br>○ What Are The Types Of Sampling Techniques In Statistics - Random, Stratified, Cluster, Systematic<br>○ Confidence Intervals, Clearly Explained!!!<br>○ How To Calculate the Confidence Interval (With Examples)<br>○ Bootstrapping Confidence Intervals: the basics |
| **Materials** | ● **HYP-02- LAB - Identifying Sampling Method - Learner Instructions**<br>● **HYP-02- LAB - Identifying Sampling Method - Instructor Instructions**<br>● **confidence_intervals.zip**<br>● **HYP-02 - Introduction to Inferential Statistics - Slide Deck** |

**The Art of Data Analysis and Reporting - Online**
**Session 2: Introduction to Inferential Statistics** (90 minutes, Synchronous)

| Time | Activity |
|---|---|
| 5 minutes | **Opening**<br>**Slides 1-12**<br>● **Introduce** inferential statistics<br>○ *Today, we're diving into the fascinating world of inferential statistics and exploring how it connects to decision-making.*<br>● **Explain** what statistics is.<br>○ *First, let's define statistics.* ***Statistics*** *is the collection and interpretation of data.*<br>○ *To understand statistics you have to know some basic definitions.*<br>○ ***Population****: Refers to the total amount of "things" you are interested in studying. This can refer to the total amount of people, dogs, houses, etc.*<br>■ *If you're researching students' test scores in a country, the population is all the students in that country.*<br>■ *If you're researching students' test scores in a school, the population is all the students in that school.*<br>○ ***Sample****: Refers to a small part of the population that is used for study.*<br>○ ***Sample Size****: The total amount of "things" is the sample size.*<br>● **Introduce** the activity.<br>○ *I'll show you two pie charts.*<br>○ *The first represents* ***descriptive statistics****—what we already know about customer spending in a store.*<br>○ *The second is a mystery and will be revealed soon!*<br>● **Present** the first chart. Refer to the image on slide 6.<br>○ *Here's the first pie chart, showing the average spending of all customers in the store.*<br>○ *This data is based on records from the entire customer population.*<br>○ *This chart is an example of* ***descriptive statistics****, where we* **summarise** *data for an entire population.*<br>○ *The spending is divided into three main categories:* ***Food****,* ***Clothing****, and* ***Electronics****.*<br>○ *Based on the data:*<br>■ ***40%*** *of spending is on food.*<br>■ ***30%*** *is on clothing.*<br>■ ***30%*** *is on electronics.*<br>○ *This chart gives us a clear picture of how the entire group of customers spends their money.*<br>● **Explain** the relevance of the previous chart.<br>○ *Charts like this are essential for understanding trends and patterns.* |

| | |
|---|---|
| | ○ *For example, a store manager could use this information to decide which products to stock more of.*<br>○ *But what if we only have a small sample of customers? That's where **inferential statistics** come in.*<br>● **Introduce** the second chart. Refer to the image on slide 9.<br>○ *Now, here's a blank pie chart. Imagine this chart is based on a sample of **100 customers**.*<br>**Instructor Note:** If using a poll, have learners enter their guesses for each category: Food, Clothing, and Electronics.<br>● **Ask** learners to give answers to the following question in the **Chat Room**[FT]<br>○ *In the chat, share your guesses for the spending proportions in the sample. Think about whether you expect the proportions to stay the same, change slightly, or vary a lot.*<br>○ **Possible Responses***:* I think the sample might have 45% for Food, 25% for Clothing, and 30% for Electronics, slightly different from the population chart.<br>● **Read** a variety of responses in the chat to notice different perspectives or patterns.<br>● **Encourage** learners to reflect on the concept of sampling.<br>● **Ask** learners to respond to the following question through an **interactive poll**.<br>○ *Do you think a sample of 100 customers can represent the spending habits of the entire population accurately?*<br>○ *Options:*<br>○ *A. Yes, it can be representative.*<br>○ *B. No, it's not large enough to be representative.*<br>○ *C. It depends on how the sample is selected.*<br>● **Share** objectives. |
| 5 minutes | **Understanding Descriptive and Inferential Statistics: Summarising Data and Making Prediction**<br>**Slides 13-16**<br>● **Explain** what descriptive statistics is.<br>○ ***Descriptive Statistics** involves **summarising, organising** and **presenting** data to understand its main features.*<br>○ *For example, in a class of 30 students, calculating the average test score (mean), identifying the most common score (mode), and plotting a bar graph of score distributions are all descriptive statistics.*<br>○ *These methods provide a clear overview of the data but don't predict how students in other classes might perform.*<br>● **Explain** what inferential statistics is.<br>○ ***Inferential Statistics** involves **making predictions** or **generalisations** about a population based on a sample.*<br>○ *For instance, if you survey **100 customers** from a store to estimate the average monthly spending for all customers, you're using inferential statistics.*<br>○ *This extends beyond the immediate data to inform broader decisions.*<br>● **Open** and **share** the Introduction to Statistics (1.1) and What Are Is Descriptive And Inferential Statistics - What Are The Different Branches Of Statistics video. |

| | |
|---|---|
| | ● **Tell** learners the video explains descriptive statistics (organising and summarising data) and inferential statistics (drawing conclusions about a population from a sample). |
| 20 minutes | **Types Of Sampling Techniques In Statistics**<br>**Slides 17-30**<br><br>**Instructor Note:** The project for this activity is in **confidence_intervals.zip**.<br>● This section addresses the **Sampling_Activity.ipynb** notebook.<br>● It uses a Jupyter notebook to show how to calculate the confidence interval.<br><br>● **Introduce** the activity.<br>○ *We'll explore different* ***sampling methods****—techniques used to select a smaller group from a larger population for analysis.*<br>○ *Understanding these methods is crucial for collecting reliable data, and we'll also explore some code examples to illustrate how they work in practice.*<br>● **Share** the link to the project **confidence_intervals.zip.**<br>● **Ask** learners to download the project.<br>● **Give Directions:**<br><br>● **Step 1: Open the Project Folder**<br>○ **Say -** *Navigate to the directory where you downloaded the project.*<br>○ *Open the project folder with Visual Studio.*<br>● **Step 2: Setup a Virtual Environment**<br>○ **Say** - *Create a virtual environment to keep project dependencies separate. In your terminal, navigate to your project folder and type:* ***python -m venv .venv***<br>■ `Code:`<br>`python -m venv .venv`<br>○ *To activate it:*<br>○ ***Windows:*** *.venv\Scripts\activate*<br>○ ***macOS/Linux:*** *source .venv/bin/activate*<br>■ `Code:`<br>`#Windows: .venv\Scripts\activate`<br>`#macOS/Linux: source .venv/bin/activate`<br>○ *Once activated, you'll see the environment name in your terminal prompt.*<br>● **Step 3: Install Required Libraries**<br>○ **Say** - *With the environment activated, install the required libraries.*<br>■ `Code:`<br>`pip install numpy pandas matplotlib ipykernel` |

- 
    - ***NumPy (Numerical Python)*** *is a powerful Python library used for numerical and scientific computing.*
    - *It provides tools for working with arrays, mathematical operations, and data manipulation, making it essential for data analysis, machine learning, and computational tasks.*
- **Explain** what Simple Random Sampling is.
    - *In **Simple Random Sampling**, every individual in the population has an equal chance of being selected. Selection is entirely random, akin to drawing names from a hat.*
    - *This approach minimises bias and ensures each subset is equally likely to be chosen.*
    - ***Example**: Assigning numbers to each member of a population and using a random number generator to select participants.*
- **Step 4**: **Simple Random Sampling**
    - **Say -** *Open the **notebooks > Sampling_Activity.ipynb** notebook.*
        - Code:

          ```
          # Random Sampling
          import random
          population = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
          # Sample size of 4
          sample = random.sample(population, 4) # Sample size of 4
          print("Random Sample:", sample)
          ```

    - **Say** - *Select the **.venv** kernel.*
    - Run the cell
    - **Explanation**:
    - *The random module is imported to use its built-in functions for generating random selections.*
    - ***Defining the Population**: The population is defined as a list of integers from 1 to 10, representing the full set of individuals or items to choose from.*
    - ***Specifying the Sample Size:** A sample size of 4 is specified, meaning 4 elements will be randomly selected from the population.*
    - ***Random Sampling Using random.sample():** The function **random.sample(population, 4)** is used to select 4 elements randomly from the population list.*
- **Explain** what Systematic Sampling is.
    - ***Systematic sampling** involves selecting every nth individual from a list after a random starting point.*
    - *This method is straightforward and ensures a spread across the population.*
    - ***Fixed Interval (n):***
        - *After selecting the **starting point**, every nth individual is chosen based on a fixed interval.*
        - *The interval **n** is determined by dividing the population size by the desired sample size.*
        - *For example, if you want to sample 10 individuals from a population of 100, the interval n would be **100/10=10.***
    - ***Selection Process:***

    - *Starting from the random point, you add the interval $n$ repeatedly to select the remaining individuals.*
    - *For example: If the starting point is 7 and $n$ = 10, the selected individuals are 7, 17, 27, 37, ..., 97.*
- **Step 4**: **Systematic Sampling**
    - Code:

```
# Systematic Sampling
population = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
sample_size = 3
n = len(population) // sample_size  # Calculate the interval
start = random.randint(0, n - 1)   # Random starting point within the interval
sample = [population[i] for i in range(start, len(population), n)]

print("Interval:", n)
print("Sarting Point:", start)
print("Systematic Sample:", sample)
```

  - Run the cell
  - **Explanation**:
  - *The population is represented as a list of integers from 1 to 10.*
  - ***sample_size = 3:*** *This indicates that 3 individuals will be selected from the population.*
  - ***Calculate the Sampling Interval (n):***
    - ***n = len(population)*** *// sample_size*
    - *The interval $n$ is calculated by dividing the population size by the desired sample size.*
    - *For a population of 10 and a sample size of 3:* ***$n$ = 10/3 = 3***
    - *This means every 3rd individual will be selected after the starting point.*
  - ***range(start, len(population), n):*** *Generates a sequence of indices starting from the value* ***start****, ending at the length of the population* ***(len(population))****, and incrementing by the interval* ***n****.*
- **Explain** what Stratified Sampling is.
  - *In* ***Stratified Sampling****, the population is divided into distinct subgroups, or* ***strata****, based on specific characteristics (e.g., age, gender, income level).*
  - *Random samples are then taken proportionally from each stratum. This technique ensures representation from all key subgroups.*
  - ***Example****: Surveying equal numbers of students from different academic years to understand study habits across all levels.*
- **Step 5**: **Stratified Sampling**
    - Code:

```
#Stratified Sampling
population = {"Group1": [1, 2, 3], "Group2": [4, 5, 6], "Group3": [7, 8, 9]}
sample = {group: random.sample(values, 1) for group, values in population.items()}
```

```
print("Stratified Sample:", sample)
```

- Run the cell.
- **Explanation**:
- ***Define the Population:***
- *The population is a dictionary, where:*
  - *Each key represents a subgroup (e.g., Group1, Group2, Group3).*
  - *The values are lists of individuals belonging to each subgroup.*
- ***Perform Sampling:***
- *The code uses a dictionary comprehension to randomly select one individual from each subgroup.*
- *For each subgroup (group), the function:*
- *Accesses the list of individuals (values).*
- *Uses **random.sample(values, 1)** to randomly select one individual to keep the explanation simple.*
- ***Create the sample:***
- *The output is a new dictionary (sample) where:*
- *The keys are the subgroup names.*
- *The values are the randomly selected individuals from each subgroup.*

- **Explain** what Cluster Sampling is.
  - *In **Cluster Sampling**, the population is divided into **clusters**, often based on geography or other natural groupings.*
  - *Entire clusters are randomly selected, and all individuals within chosen clusters are surveyed. This method is cost-effective, especially when populations are widespread.*
  - ***Example**: Selecting specific neighbourhoods in a city and surveying every household within those neighbourhoods.*
- **Step 6**: **Clustered Sampling**
  - Code:

```
#Cluster Sampling
clusters = [["A1", "A2", "A3"], ["B1", "B2", "B3"], ["C1", "C2", "C3"]]
selected_cluster = random.choice(clusters)
print("Cluster Sample:", selected_cluster)
```

- Run the cell.
- **Explanation**:
- ***Population Organisation:***
- *The population is grouped into predefined clusters.*
- *Clusters could represent geographic areas, teams, classrooms, etc.*
- ***Random Selection of a Cluster:***
- *Only one cluster is selected, and all individuals in that cluster become part of the sample.*

| | |
|---|---|
| | ○ *The code is kept simple to illustrate the basic idea of randomly selecting a cluster from a list of clusters.*<br>○ *Often, multiple clusters are selected to ensure the sample captures diversity across the population.*<br>● **Highlight** the main difference between Stratified Sampling and Cluster Sampling.<br>○ **Stratified Sampling**:<br>■ *Focuses on **representation.***<br>■ *Samples **within subgroups** to ensure **representation**.*<br>■ *Members within each **stratum** are **similar** in characteristics (e.g., age, gender, income level), and the goal is to sample a proportion from each stratum to ensure representation.*<br>○ ***Cluster Sampling**:*<br>■ *Emphasises **efficiency.***<br>■ *Samples **entire groups** for convenience, and **cost-efficiency** and entire clusters are sampled.*<br>■ *Members within each **cluster** are **not necessarily similar** but instead represent a mix of the population.*<br>● **Explain** what Convenience Sampling is.<br>○ ***Convenience Sampling** is a **non-probability** sampling method where participants are selected based on how easy they are to reach or access.*<br>○ *It's often used when time, resources, or access to the full population is limited.*<br>○ ***Example**: Asking friends for opinions.*<br>○ ***Bias**: Your friends might have similar preferences, backgrounds, or experiences as you, therefore their opinions may not reflect those of the larger population.*<br>● **Ask** 1 learner to answer the following question. **Cold Call**[FT]<br>○ *Why is there no code in the Jupyter notebook to demonstrate convenience sampling?*<br>■ **Possible Response***:* Convenience sampling is typically based on selecting individuals who are easiest to access, rather than following a specific algorithm or process that can be coded. It often involves manual selection or ad-hoc decisions, which may not require code to illustrate.<br>● **Conclude** the explanation.<br>○ *Each sampling technique has its advantages, and the choice depends on the objectives, the organisation of the population, and the resources available.*<br>● **Open** and **share** the What Are The Types Of Sampling Techniques In Statistics - Random, Stratified, Cluster, Systematic video.<br>● **Tell** learners this video explains key statistical sampling techniques, including random, stratified, cluster, and systematic sampling, with practical examples. |
| 20 minutes | **Group Discussion: Identify the Sampling Method**<br>**Slides 31-32**<br><br>**Instructor Note:** This group activity allows learners to identify the sampling method applied in a closely related real-world scenario, emphasising the reasoning behind selecting that method.<br>● This section uses the **Identify_Sample_Activity.ipynb** notebook from the **confidence_intervals.zip** project. |

- Divide the learners into groups of 2 members randomly.
- The learners’ instructions are in the **HYP-02- LAB - Identifying Sampling Method - Learner Instructions** handout.
- The instructor's instructions are in the **HYP-02- LAB - Identifying Sampling Method - Instructor Instructions** handout.

- **Guide** learners through a collaborative group activity where they will identify a sampling method used and justify their answers.
  - *You will collaborate in groups to determine the sampling method used, based on the provided description and the outcomes of the code execution.*
- **Share** the link to the **HYP-02- LAB - Identifying Sampling Method - Learner Instructions** handout.
- **Ask** learners to make copies of the handout.
- **Give Directions**
  - **30 seconds -** Go into your breakout rooms (2 members per room).
  - **10 mins -** Work in groups to analyse the code, identify the sampling method, and justify your reasoning.
    - Run the provided code and review its output.
    - Discuss the structure of the code and its implementation with the help of the guiding questions.
    - Use the provided table to support your analysis and conclusions.
    - Write a concise justification for your identified method in 1–2 sentences.
    - Prepare to present your findings.
- **Monitor** time.
- **Send a One Minute Warning**FT via voice and chat.
- **Ask** 1 group to give their answers to the following question. **Cold Call**FT
  - *Based on the evidence in the code, determine which sampling method is being used.*
  - *Explain why this specific sampling method was used with evidence from the table.*
    - **Expected Response:** The method used is stratified sampling. The code splits the population into subgroups (faculties) to ensure all groups are included.
      It calculates the sample size for each group based on its proportion of the total population and selects individuals randomly within each group using **random.sample()**.
      **Stratified sampling** is a method that divides a population into distinct subgroups and ensures each subgroup is represented in the sample, often proportionally to its size.
- **Facilitate** discussion.
- **Invite** the non-presenting groups to share their perspectives by asking:
  - *Did you identify the same method as the presenting groups? Why or why not?*
- **Share** the correct answers.
- **Clarify** any misunderstandings or provide further insights about the sampling method.

15 minutes

**Confidence Intervals Using Bootstrapping**
**Slides 33-38**

- **Introduce** *the activity.*
  - *We'll learn about **confidence intervals**, a statistical tool that helps us estimate an unknown population parameter, like an average, based on a sample.*
  - *To make this concept easier to understand, we'll use the **bootstrapping method**, which involves resampling data to calculate confidence intervals intuitively.*
  - *First, let's watch a short video to introduce this concept visually.*
- **Ask** learners to focus on these key points while watching the video**:**
  - *Understand how bootstrapping works—resampling from the original dataset.*
  - *Observe how confidence intervals are constructed from the range of resampled statistics.*
  - *Learn what a 95% confidence interval represents.*
- **Open** and **share** the Confidence Intervals, Clearly Explained!!! Video. Refer to the slide 35.
- **Play** the video starting at **00:57** and ending at **05:25.**
- **Ask** 2 learners to give answers to the following question. **Warm Call**$^{FT}$
  - *Why is resampling important for making predictions about a population from a small sample?*
    - **Possible Response***:* Resampling helps us understand how much a statistic, like the mean, might vary by creating many new samples from the original data. This shows us the range where the true population value is likely to fall, even if the sample is small or we don't know the population's distribution.
- **Recap** the video.
  - *As you saw, bootstrapping creates a confidence interval by taking many random resamples from the data, calculating a statistic like the mean for each, and identifying the range where most of those statistics fall.*
- **Explain** what a **95%** confidence interval means.
  - *A **95% confidence interval** means we're 95% confident that the true average lies within this range.*
- **Explain** why the 95% confidence interval is commonly chosen.
  - *The **95% confidence interval i**s commonly chosen because it represents a widely accepted balance between confidence and precision. Here's why:*
  - ***Standard in Statistical Practice:** A 95% confidence level means that if we repeated the sampling process many times, approximately 95% of the calculated intervals would contain the true population parameter.*
  - ***Balancing Confidence and Precision:***
  - ***Higher confidence levels** (e.g., 99%) result in wider intervals, which may reduce precision.*
    - *A wider interval means we are less specific about where the true value lies.*
    - *For instance, saying "the average is between 40 and 60" (wider interval) is less precise than "the average is between 45 and 55" (narrower interval).*
    - *In a medical study, knowing that a treatment is effective within a precise range (e.g., reducing symptoms by 5–10%) is more useful than a wide range (e.g., 2–20%).*
  - ***Lower confidence levels** (e.g., 90%) produce narrower intervals with less certainty.*

| | |
|---|---|
| | ■ *The trade-off for the narrower interval is less certainty—there's a 10% chance the true value is not in the interval (compared to 5% for a 95% confidence interval).*<br>■ *In practical terms, you are less confident that the narrower interval accurately represents the population.*<br>○ *A 95% interval strikes a **practical balance**, offering high confidence without overly broad intervals.* |
| 15 minutes | **Code-Along: Calculating Confidence Intervals Using Bootstrapping**<br>**Slides 39-41**<br><br>**Instructor Note:** The project for this activity is in **confidence_intervals.zip.** It uses a Jupyter notebook to show how to calculate the confidence interval.<br>● It uses the bootstrapping percentile method to calculate the confidence interval.<br>● Demonstrate the process first, then allow learners time to code along and replicate the steps independently.<br>● Address questions as they arise.<br>Answer to possible questions learners might have:<br>● The **mean** is typically used for calculating confidence intervals because it is a standard measure of central tendency and is directly tied to many statistical methods, such as hypothesis testing and regression analysis.<br>● Here's why the mean is preferred over the **median**:<br>● **Why Use the Mean for Confidence Intervals?**<br>○ **Statistical Properties:**<br>■ The mean has well-established theoretical properties in statistics, such as its relationship to the normal distribution (Central Limit Theorem).<br>■ This makes it easier to estimate variability (e.g., standard error) and construct confidence intervals.<br>○ **Ease of Calculation:** The formula for confidence intervals, such as **Mean±Margin of Error**, relies on the mean because the margin of error is often calculated using the mean's standard error.<br>○ **Applicability in Most Situations:** The mean is widely used in practical applications where data are expected to represent an average tendency, such as in surveys or experiments.<br>● **Why Not the Median?**<br>○ **Median Lacks a Direct Standard Error Formula:** Unlike the mean, the median does not have a straightforward formula for its standard error, making it harder to calculate confidence intervals.<br>○ **Less Sensitive to Data Changes:** While the median is more robust to outliers, this same robustness makes it less sensitive to subtle variations in the data, which confidence intervals aim to capture.<br>● **When to Use the Median**<br>○ The median might be used for confidence intervals in cases where the data are heavily skewed, contain extreme outliers, or the median is more representative of the "typical" value in the dataset (e.g., income data).<br>● The **t-distribution** (also called Student's t-distribution) is a probability distribution used in statistics when the sample size is small, or the population standard deviation is unknown. It is similar to the normal distribution but has heavier tails, meaning it accounts for more variability in the data. |

- The **z-distribution,** also known as the standard normal distribution, is a specific type of normal distribution where the mean is 0 and the standard deviation is 1. It is widely used in statistics for standardising data, calculating probabilities, and performing hypothesis tests.
- **Bootstrapping as a Method:**
    - No normality assumption.
    - Resamples the observed data to approximate the sampling distribution of a statistic empirically.
- **Percentile Method:**
    - Relies on the quality of the bootstrap distribution as an approximation of the true sampling distribution.
    - This can fail for small samples or skewed distributions, but this is not about normality—it's about how well the bootstrap distribution mirrors the true variability of the statistic.

- **Introduce** the activity.
    - *Using the bootstrapping method, we will calculate a 95% confidence interval for weekly grocery spending.*
    - *This interval represents the range within which most possible values are likely to fall.*
- **Recap** what bootstrapping is.
    - *Bootstrapping is a technique for estimating a population parameter, such as the mean, using repeated random sampling with replacement.*
- **Give Directions.**

- **Step 1: Dataset Setup**
    - **Say** - Open **notebooks > Confidence_Interval_Activity.ipynb.**
    - *Our dataset includes weekly grocery spending for 10 households. It’s small but will help us understand spending patterns.*
        - Code:

          ```
          # Import necessary libraries
          import numpy as np
          import pandas as pd
          import matplotlib.pyplot as plt

          # Define the dataset
          data = [50, 60, 55, 70, 65, 80, 75, 60, 85, 90]
          print("Original Dataset:", data)
          ```

    - Select the virtual environment **.venv**.
    - Run the cell.
    - **Say** - *We aim to calculate a 95% confidence interval for the true average weekly grocery spending in the population.*
- **Step 2: Resample the Data**

- **Say** - *Resampling involves creating new datasets by randomly selecting data points from the original sample, with replacement.*
- *This mimics the variability we might see if we collected multiple samples.*
- *For example, if we resample from the data [50, 60, 55], one resample might be [60, 50, 60], and another might be [55, 55, 60].*
- *Let's generate 1,000 resamples.*
    - Code:

      ```
      num_resamples = 1000  # Number of bootstrap samples
      np.random.seed(42)
      resamples = [np.random.choice(data, size=len(data), replace=True) for _ in range(num_resamples)]

      # Display the first 5 resamples as examples
      print("\nFirst 5 Resamples:")
      for i, resample in enumerate(resamples[:5], start=1):
          print(f"Resample {i}: {resample}")
      ```

- Run the cell.
- **Explanation:**
- ***np.random.seed(42):*** *It is used to set the random number generator's seed in NumPy.*
- *It ensures that whenever you run your code, you get the same random results. It is very useful in debugging and teaching.*
    - *42 is just a number. You can use any number.*
- ***num_resamples = 1000:*** *This specifies the number of bootstrap samples you want to generate. In this case, 1,000 resamples will be created.*
- ***np.random.choice:*** *This function selects elements from the data array.*
- ***data****: This is the original dataset you're working with.*
- ***size=len(data):*** *The size of each resample is the same as the size of the original dataset.*
- ***replace=True****: Sampling is done with replacement, meaning that the same data point can appear multiple times in a single resample.*
- ***[... for _ in range(num_resamples)]:*** *This is a list comprehension that repeats the resampling process* ***num_resamples*** *times, creating a list of 1,000 resampled datasets.*
- ***What Does It Do?***
- *It generates 1,000 bootstrap samples, each of which is a randomly sampled version of the original dataset (data) with replacement.*
- *Each resample is stored as an element in the* ***resamples*** *list.*
- *These resamples can be used to estimate the distribution of a statistic (e.g., mean, median, standard deviation) by calculating that statistic for each resample.*

- **Step 3: Calculate Means for Resamples**
    - **Say** - *We calculate the mean for each resample to estimate the variability in sample means.*
    - *For the resample [60, 50, 60], the mean is (60+50+60)/3 = 56.7*
        - Code:

          ```
          # Calculate the mean for each resample
          resampled_means = [np.mean(resample) for resample in resamples]

          # Display the first 10 resampled means
          print("\nFirst 10 Resampled Means:")
          for mean in resampled_means[:10]:
              print(f"{float(mean):.2f}")
          ```

    - Run the cell.
    - **Explanation**:
    - *Each value in the list is the mean of one bootstrap resample.*
- **Step 4: Visualise the Distribution**
    - **Say** - *Next, We'll plot a histogram of the 1,000 resampled means to see their distribution. This distribution shows the range of possible sample means.*
        - Code:

          ```
          plt.hist(resampled_means, bins=30, alpha=0.7, edgecolor='black')
          plt.title('Distribution of Resampled Means')
          plt.xlabel('Mean Value')
          plt.ylabel('Frequency')
          plt.axvline(np.percentile(resampled_means, 2.5), color='red', linestyle='dashed',
          label='2.5th Percentile')
          plt.axvline(np.percentile(resampled_means, 97.5), color='blue', linestyle='dashed',
          label='97.5th Percentile')
          plt.legend()
          # plt.savefig('Bootstrap_Confidence_Interval_Plot.png')  # Save the plot
          plt.show()

          # print("\nHistogram saved as 'Bootstrap_Confidence_Interval_Plot.png'.")
          ```

    - Run the cell.
    - **Explanation**:
    - *The **x-axis** represents the range of values (the resampled means in your case).*
    - *The **y-axis (height of bars)** represents the **frequency**: how many data points fall into each range (bin).*
    - *For example:*

  - *If there's a tall bar at "70" on the x-axis, it means that many resampled means are **close to 70**.*
  - ***Shorter bars** indicate that fewer resampled means fall within those value ranges.*
  - ***What Does the Spread (Width of the Distribution) Mean?***
  - *The **spread (width)** of the histogram reflects variability in the resampled means:*
  - *A **narrow distribution (steep peak)** indicates low variability—the data is consistent.*
  - *A **wide distribution (flatter peak)** indicates high variability—the data has more spread.*
  - *In the histogram, the resampled means are tightly clustered, indicating that the original data is consistent and doesn't have extreme values.*
  - ***Causes of High Variability***
    - ***Large Differences in Data:** If the original dataset has values that are very different (e.g., [10, 100, 200]), resampled means will vary greatly.*
    - ***Small Sample Size**: If the original dataset has very few data points (e.g., 5 instead of 100), each resample will be more sensitive to the specific data points chosen, leading to more variability.*
    - ***Outliers**: Extreme values in the data (like a value much higher or lower than others) can increase the variability.*
  - ***Implications of High Variability***
  - *High variability can mean:*
    - ***Uncertainty**: It's harder to estimate the **true mean** of the population because the data is inconsistent.*
    - ***Wide Confidence Intervals:** With high variability, the 95% confidence interval will be wider, reflecting more uncertainty about the true mean.*

- **Step 5: Find the 95% Confidence Interval**
  - **Say** - *To calculate the 95% confidence interval, we find the 2.5th percentile (lower bound) and the 97.5th percentile (upper bound) of the resampled means.*
    - Code:

      ```
      # Calculate the 95% confidence interval
      lower_bound = np.percentile(resampled_means, 2.5)
      upper_bound = np.percentile(resampled_means, 97.5)
      confidence_interval = (lower_bound, upper_bound)

      print("\n95% Confidence Interval:")
      print(f"Lower Bound: {lower_bound}")
      print(f"Upper Bound: {upper_bound}")
      ```

  - Run the cell.
  - **Explanation:**
  - *If the lower bound is 61.5 and the upper bound is 77.0, the confidence interval is (61.5, 77.0). This means we're 95% confident the true average lies within this range.*
  - You can also see it in the histogram plot.

- **Ask** 2 learners to give answers to the following question. **Cold Call**$^{FT}$
  - *If your confidence interval for weekly grocery spending is (**60, 80**) dollars, you are fairly certain the true average lies within this range. Now, imagine the confidence interval widens to (**50, 90**) dollars.*
  - *What does the wider interval tell you about how certain we are about the true average?*
    - **Possible Response***:* The wider interval indicates less certainty about the exact value of the mean.
- **Expand** on the answer.
  - *A **narrow interval** like (**60, 80**) suggests you have **high precision** and can estimate the mean more accurately.*
  - *A **wider interval**, like (**50, 90**), tells us that we are **less certain** about the true average. The wider the interval, the more variability or uncertainty there is in our estimate, meaning we can't pinpoint the true average as precisely.*
- **Ask** 2 learners to give answers to the following question. **Warm Call**$^{FT}$
  - *Why do you think we take many samples (1,000) when using the bootstrap method? What might happen if we took only a few?*
    - **Possible Response***:* We take many samples in the bootstrap method to make our results more reliable. If we only took a few samples, the estimates might not represent the true behaviour of the data well, and our conclusions could be less accurate. Taking many samples helps us get a clearer picture of the possible outcomes.
- **Explain** why 1,000 resamples were used.
  - *Bootstrapping works by generating multiple resamples to estimate the sampling distribution of a statistic. The more resamples you take, the closer your estimate will be to the true distribution.*
  - *In practice, 1,000 resamples usually provide a good balance: it's large enough to produce reliable estimates, but not so large that it becomes computationally expensive for most modern systems.*
  - *Other resampling or iterative techniques in statistics and machine learning also use a similar idea of 1,000 resamples to ensure robust estimates.*
  - ***Flexibility***
  - *Depending on the complexity of your problem and computational resources, you can adjust this number.*
  - *For instance:*
  - *Use **500** resamples for a quicker, less precise estimate.*
  - *Use **10,000** or more if you need extremely high accuracy and have the resources to handle it.*
- **Show** how the mean of the resamples stabilises as the number of samples increases.
  - **Say** - *The following code visualises how increasing the number of resamples affects the stability of the mean.*
    - Code:

```
# The mean of the resamples stabilises as the number of samples increases
sample_sizes = [100, 500, 1000, 5000]

for n in sample_sizes:
    resamples = [np.random.choice(data, size=len(data), replace=True) for _ in range(n)]
    resample_means = [np.mean(resample) for resample in resamples]
    print(f"Mean for {n} resamples: {np.mean(resample_means):.2f}")
```

    - ○ Run the cell.
    - ○ **Explanation**:
    - ○ *With a small number of resamples (e.g., 100), there is more variability in the estimate.*
    - ○ *As the number of resamples increases, the variability decreases, and the mean estimate becomes more consistent (stabilizes).*
    - ○ *By looking at these results, you can see that increasing the number of resamples beyond 500 doesn't change the mean much.*
    - ○ *This suggests that using 500 or more resamples is likely sufficient for this dataset to get a reliable estimate of the mean.*

- **Explain** the confidence interval formula.
    - ***Confidence Interval = Sample Mean ± Margin of Error***
    - *The **sample mean** is your best guess for the true population mean based on your sample.*
    - *The **margin of error** accounts for uncertainty by adding and subtracting a range around the sample mean.*
    - *It depends on how much your **data varies** and how **confident** you want to be (e.g., 95%).*
- **Explain** the formula for confidence intervals used in bootstrapping.
    - ***Sample Mean:** Bootstrapping still uses the sample mean as the best estimate of the true mean.*
    - ***Margin of Error:** Instead of relying on formulas, bootstrapping generates many new datasets by resampling your data with replacement.*
        - ■ *For each resample, calculate the mean (or another statistic).*
        - ■ *The variation in these resampled means gives a practical way to estimate the margin of error.*
    - ***No Normality Assumption**: Unlike traditional methods, bootstrapping doesn't assume the data is **normally distributed**. It works directly with the data, making it flexible for non-normal or unknown distributions.*
    - ***Other Resampling Techniques:** Bootstrapping is one type of resampling method, but there are others.*
- **Emphasise** key points.
    - *Bootstrapping uses **sampling with replacement** to mimic population variability.*
    - *The **confidence interval** is built using the percentiles of the resampled statistics (e.g., the 2.5th and 97.5th percentiles for a 95% confidence interval).*
    - *This method is data-driven and doesn't rely on assumptions about the population.*
- **Open** and **share** the How To Calculate the Confidence Interval (With Examples) web page.
- **Tell** learners this resource provides a clear explanation of confidence intervals, their importance, and step-by-step instructions on how to calculate them, supplemented with practical examples.

| | |
|---|---|
| 5 minutes | **Closing**<br>**Slides 42-45**<br>● **Ask** 2-3 learners to share *What's one concept from today that changed how you think about data or decision-making?* |