Lecture 1 Notes

Module: Mathematics II
Topic: Descriptive Statistics
Name: Timothy Chia
Date: 05/01/2025

1. Introduction to Statistics

Definition: Statistics is not just about calculation; it is the art of collecting, analyzing, and interpreting data. It allows us to discover hidden patterns and make informed decisions even when we face uncertainty in the real world.

Two Main Branches:

- **Descriptive Statistics:** This branch focuses on the "here and now." It involves collecting, organizing, summarizing, and presenting the data you actually have.
  - *Goal:* To turn raw data into understandable information (e.g., calculating an average or creating a chart).
- **Inferential Statistics:** This branch focuses on the "what if." It involves making predictions, testing hypotheses, and drawing conclusions about a larger group based on a smaller sample.
  - *Goal:* To generalize findings from a sample to a population.

Data Sources:

- **Population:** The "whole picture." This is the entire collection of all individuals or items you are interested in. Measuring every single item in a population is called a *Census* (which is often expensive and time-consuming).
- **Sample:** A "snapshot." This is a smaller subset selected from the population. We study the sample to estimate properties of the population. The process of selecting this subset is called *Sampling*.

2. Types of Data

It is crucial to distinguish between data types because the statistical method you choose (like which graph to draw or which average to use) depends entirely on the type of data you have.

- **Qualitative (Categorical):** This data describes a quality, characteristic, or category. It is non-numeric and generally answers "which one?" rather than "how many?".
  - **Nominal:** Categories that are just names or labels with no inherent order or ranking.
  - **Ordinal:** Categories that have a meaningful order or ranking, but the difference between the ranks is not necessarily equal.
- **Quantitative (Numeric):** This data is expressed numerically and represents amounts or quantities. It answers "how much?" or "how many?".
  - **Discrete:** Data that can only take specific, countable values. There are "gaps" between possible values.
  - **Continuous:** Data that can take any value within a range. There are no gaps; it can be measured with increasing precision.

3. Data Presentation

Categorical Data:

- **Frequency Table:** A simple summary that lists each category alongside its frequency (f_i, the count) or relative frequency ($p_i = f_i/N$, the proportion).
- **Graphical:**
  - **Bar Charts:** Use separate bars (with gaps between them) to compare different categories.
  - **Pie Charts:** Show how a whole is divided into different categories (best for showing proportions).

Quantitative Data:

- **Frequency Distribution:** Because numeric data can be vast, we group it into "classes" or "intervals" to see the pattern.
- **Histogram:** The most common way to visualize numeric distributions. Unlike bar charts, the bars in a histogram *touch each other* to indicate that the data is continuous number-line data.
- **Unequal Class Widths:** Sometimes, grouping intervals have different sizes. If we simply plotted frequency, wider bars would look visually "heavier" or more important than they should. To correct this, we use **Density** for the height.

$$\text{Density} = \frac{\text{Relative Frequency}}{\text{Class Width}}$$

Note: In a density histogram, the *area* of the bar represents the frequency, not just the height.

Bivariate Data (Two Variables):

- **Crosstabulation (Contingency Table):** A table that displays the frequency distribution of two variables simultaneously (e.g., "Gender" vs. "Subject Choice").
- **Charts: Stacked** or **Clustered Bar Charts** allow you to compare subgroups within the data side-by-side.

4. Measures of Location (Central Tendency)

These measures try to identify the "center" or "typical value" of your dataset using a single number.

Mean ($\mu, \bar{x}$): Also known as the arithmetic average. It is the balancing point of the data. However, it is *sensitive to outliers*—one extremely high value can pull the mean upward significantly.

- **Sample Mean:** Sum all values and divide by the number of observations.

$$\bar{x} = \frac{\sum x_i}{n}$$

- **Grouped Data Mean:** When data is grouped, we don't know exact values, so we use the midpoint (x_i) of each class to estimate.

$$\bar{x} = \frac{\sum (x_i \cdot f_i)}{n}$$

Median (Q2): The exact middle value when the data is sorted from smallest to largest. It splits the data into two equal halves (50% above, 50% below). Unlike the mean, it is *robust* (not affected) by outliers.

- **Grouped Data Median:** Since we only have intervals, we use interpolation to estimate exactly where the middle value falls inside the median class.

$$Q = L_m + \left[\frac{\frac{n}{2} - cf_{m-1}}{f_m} \right] \cdot w_m$$

Key: L_m (Lower bound of median class), f_m (frequency of that class), w_m (width of that class), cf_{m-1} (cumulative frequency counted *before* that class).

Mode: The most popular value (the one with the highest frequency). A dataset is *unimodal* if it has one clear peak, or *bimodal* if it has two peaks.

Skewness: This describes the symmetry of the distribution.

- **Symmetrical:** The left and right sides look roughly the same (Mean $\approx$ Median).
- **Positively Skewed:** The "tail" of the graph drags out to the right (positive numbers). The mean is pulled to the right of the median.
- **Negatively Skewed:** The "tail" drags out to the left (negative numbers). The mean is pulled to the left of the median.

5. Measures of Spread (Variability)

Knowing the center isn't enough; we also need to know how spread out or consistent the data is.

Range: The simplest measure of spread. It only looks at the extremes.

$$\text{Range} = \text{Highest Value} - \text{Lowest Value}$$

Interquartile Range (IQR): Measures the spread of the middle 50% of the data. It ignores the extremes, making it a robust measure of spread.

$$\text{IQR} = Q_3(\text{Upper Quartile}) - Q_1(\text{Lower Quartile})$$

- **Outliers:** We use the IQR to mathematically detect unusual values. A data point is an outlier if it falls below $Q_1 - 1.5(\text{IQR})$ or above $Q_3 + 1.5(\text{IQR})$.

Variance (s^2): Ideally, we want the average distance of data points from the mean. Variance calculates the average *squared* distance (squaring makes all negatives positive).

- **Sample Variance Formula:** We divide by $n-1$ instead of n to get a more accurate estimate for the population (this is called Bessel's correction).

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- **Alternative Formula (Computational):** A shortcut formula often used in manual calculations to avoid rounding errors.

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$$

Standard Deviation (s): Because variance is in "squared units" (e.g., dollars squared), it is hard to interpret. We take the square root to return to the original units.

$$s = \sqrt{s^2}$$

Coefficient of Variation (CV): Standard deviation is an absolute measure. The CV is a *relative* measure (percentage). It is very useful for comparing **risk** or volatility between two datasets that have very different averages (e.g., stock prices of \$10 vs \$1000).

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