MTL108
Descriptive Statistics

Rahul Singh

We trust in God!

All others must bring data!

Probability vs. Statistics

We start by formally distinguishing probability from statistics. They are inverse processes.

- **Probability:** Operates from a *known population* to predict an *unknown sample. Example (Coin Toss):* If we know we hold a perfectly fair coin (the population parameter is known: $p = 0.5$), probability asks: What is the exact chance of getting 7 heads in 10 flips?
- **Statistics:** Operates from a *known sample* to infer the *unknown population. Example (Coin Toss):* We find a coin on the street, flip it 10 times, and observe 7 heads (the sample is known).

Statistics asks: Based on this sample, is the coin fair ($p = 0.5$), or is it biased?

Example 1 (The German Tank Problem)**.** Before we define our terms, let us look at a historical example that demonstrates why we study statistics: the Allied effort to estimate German tank production during World War II.

The Setup: The Allies needed to know how many Panzer V (Panther) tanks the Germans were producing. They had two methods for estimating this unknown *population size* (N):

1. **Conventional Intelligence:** Spies, intercepted communications, etc.
2. **Statistical Analysis:** Using the sequential serial numbers found on captured or destroyed tanks (our *sample*, n).

The Statistical Approach: Suppose the Allies captured $n = 4$ tanks with the serial numbers: $19, 40, 42,$ and 60. The highest observed serial number (the sample maximum) is $m = 60$. Statisticians developed an estimator for the total population size (N):

$$\hat{N} = m + \frac{m}{n} - 1$$

Using our small sample:

$$\hat{N} = 60 + \frac{60}{4} - 1 = 60 + 15 - 1 = 74 \text{ tanks}$$

The Historical Reality: In August 1942, conventional intelligence estimated the Germans were producing $1,550$ tanks per month. The statisticians, using formulas similar to the one above on serial numbers, estimated 327 tanks per month.

After the war, internal German records were captured. The actual production number for that month was 342.

Takeaway: A small, properly analyzed sample completely outperformed a massive intelligence-gathering operation.

Changing definition of statistics

- Statistics has then for its object that of presenting a faithful representation of a state at a determined epoch. (Quetelet, 1849)
- Statistics are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (Galton, 1889)
- Statistics may be regarded (i) as the study of populations, (ii) as the study of variation, and (iii) as the study of methods of the reduction of data. (Fisher, 1925)
- Statistics is a scientific discipline concerned with collection, analysis, and interpretation of data obtained from observation or experiment. The subject has a coherent structure based on the theory of Probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology. (E. Pearson, 1936)
- Statistics is the name for that science and art

which deals with uncertain inferences — which uses numbers to find out something about nature and experience. (Weaver, 1952)

- Statistics has become known in the 20th century as the mathematical tool for analyzing experimental and observational data. (Porter, 1986)
- Statistics is the art of learning from data. (Ross's book, 2014)

Population vs. Sample

As seen in the tank problem, we are usually trying to understand a large group based on a smaller subset.

- **Population (N):** The complete collection of all elements or items under study. *Example:* Every single tank produced by Germany in a given month.
- **Sample (n):** A subset of the population selected for analysis. *Example:* The handful of tanks the Allies managed to capture.

Parameters vs. Statistics

- **Parameter:** A fixed, but often unknown, numerical value summarizing a characteristic of the *population* (e.g., true total production N, population mean μ).
- **Statistic:** A known, fluctuating numerical value computed entirely from the *sample* data (e.g., observed maximum m, sample mean $\bar{x}$).

Types of Data

Before calculating descriptive statistics, we must classify the type of data we are analyzing, as this dictates which statistical tools are appropriate.

- **Qualitative (Categorical) Data:** Describes categories or attributes.
  - *Nominal:* Categories with no inherent order. *Example:* Blood types (A, B, AB, O) or coin toss outcomes (Heads, Tails).
  - *Ordinal:* Categories with a meaningful ranking or order, but the intervals between them are not equal. *Example:* Survey responses

(Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) or finishing positions in a race (1st, 2nd, 3rd).

- **Quantitative (Numerical) Data:** Represents measurable quantities.
  - *Discrete:* Data that can only take specific, countable values. *Example:* The number of heads in 10 coin flips, or the number of students in a classroom.
  - *Continuous:* Data that can take any value within a range; it is measured rather than counted. *Example:* The exact weight of an apple, or the time it takes to run a mile (6.24 minutes).

Measures of Central Tendency

Measures of central tendency aim to identify the center, or typical value, of a dataset.

1. The Mean (Arithmetic Average)

$$\text{Population Mean: } \mu = \mathbb{E}(X)$$

$$\text{Sample Mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. The Median

The median is the exact middle value of a dataset when ordered from smallest to largest. Let the order statistics be $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

$$\text{Median} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

Example 2 (Robustness to Outliers)**.** Consider the salaries of 5 entry-level employees: $\$40k, \$45k, \$50k, \$55k, \$60k$. The median is $\$50k$. If we replace the top earner with the CEO making $\$1,200k$, the sorted data is $\$40k, \$45k, \$50k, \$55k, \$1,200k$. The new mean shifts drastically to $\$278k$, but the median remains exactly $\$50k$. The median is highly resistant to extreme outliers.

3. The Mode

The mode is the value that occurs most frequently in the dataset. It is the only measure of central tendency that can be used for nominal data. A dataset can have one mode (unimodal), more than one mode (bimodal/multimodal), or no mode at all.

- *Example 1 (Unimodal):* Consider the die rolls: $2, 3, 3, 4, 5$. The mode is 3.
- *Example 2 (Bimodal):* Consider the dataset: $2, 2, 3, 4, 4, 5$. Both 2 and 4 appear twice. The dataset is bimodal with modes 2 and 4.
- *Example 3 (Nominal Data):* If a sample of 10 cars contains 6 red cars, 3 blue cars, and 1 black car, the mode is "Red". We cannot calculate a mean or median for colors.

Position Measures: Quantiles and Percentiles

While the median divides a dataset into two equal halves, quantiles are values that divide a ranked dataset into k equal-sized subsets.

1. Percentiles ($k = 100$)

Percentiles divide the dataset into 100 equal parts. The p-th percentile ($\mathcal{P}_p$) is the value below which $p\%$ of the observations fall.

Example 3. If a student scores in the 90th percentile on a standardized math exam, it means they scored higher than 90% of all students who took the exam. Only 10% of students scored higher than them.

2. Quartiles ($k = 4$)

Quartiles are specific percentiles that divide the data into four equal quarters.

- **First Quartile (Q_1):** The 25th percentile (P_{25}). 25% of the data lies below it.

- **Second Quartile (Q_2):** The 50th percentile (P_{50}). This is exactly the **Median**.
- **Third Quartile (Q_3):** The 75th percentile (P_{75}). 75% of the data lies below it.

Calculating a Percentile (Index Method): To find the p-th percentile of a sorted dataset of size n, compute the index position i:

$$i = \frac{p}{100} \times n$$

If i is not an integer, round up to the next integer to find the position. If i is an integer, the percentile is the average of the values at positions i and $i + 1$.

Example 4. Find the 75th percentile (Q_3) of the following 8 ordered test scores: $50, 60, 65, 70, 75, 80, 85, 90$.

$$i = \frac{75}{100} \times 8 = 0.75 \times 8 = 6$$

Since 6 is an integer, we average the 6th and 7th values:

$$Q_3 = \frac{80 + 85}{2} = 82.5$$

Measures of Dispersion

While central tendency tells us where the data is centered, dispersion tells us how spread out the data is around that center.

A. The Range and IQR

- **Range:** The simplest measure of spread. Range $= x_{(n)} - x_{(1)}$ (Maximum - Minimum). It is highly sensitive to outliers.
- **Interquartile Range (IQR):** Measures the spread of the middle 50% of the data. IQR $= Q_3 - Q_1$, where Q_3 is the 75th percentile and Q_1 is the 25th percentile.

B. Variance

Variance measures the average squared deviation of each data point from the mean.

$$\text{Population Variance: } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\text{Sample Variance: } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

C. Standard Deviation

Because variance is measured in squared units (e.g., ”squared runs”), we take the square root to return to the original units of measurement.

$$\text{Population Standard Deviation: } \sigma = \sqrt{\sigma^2}$$

$$\text{Sample Standard Deviation: } s = \sqrt{s^2}$$

Practical Example: Evaluating Consistency

Consider two cricket batsmen over a 5-match series.

- **Batsman A runs:** $40, 50, 45, 55, 60$ $(\bar{x} = 50)$
- **Batsman B runs:** $0, 100, 10, 140, 0$ $(\bar{x} = 50)$

Both have the exact same mean (50 runs). However, calculating their sample standard deviations reveals a stark difference in dispersion. Batsman A has a very low s (high consistency), while Batsman B has a very high s (high volatility/heavy-tailed performance). Measures of dispersion are required to capture this reality.

March 10

Let X has PMF

$$p_X(x) = \begin{cases} \frac{c}{x^2}, & x \in \{1, 2, 3, \dots\} \\ 0, & \text{otherwise} \end{cases}$$

where

$$c^{-1} = \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$$

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} x \cdot p_X(x)$$

$$= c \underbrace{\sum_{x=1}^{\infty} \frac{1}{x}} = \infty$$

Measure of dispersion

Data 1:

95, 100, 105.

Data 2:

5, 100, 195