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)
- Statistics is the science and art of learning from data. (We conclude!)

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 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

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

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 90^{th} 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 25^{th} percentile (P_{25}). 25% of the data lies below it.
- **Second Quartile (Q_2):** The 50^{th} percentile (P_{50}). This is exactly the **Median**.
- **Third Quartile (Q_3):** The 75^{th} 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 75^{th} 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 6^{th} and 7^{th} values:

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

Formal Definitions of Population Median and Mode

While we previously defined the median and mode for a finite sample, in statistical theory, these parameters are formally defined using the underlying probability distributions.

A. Population Median

The population median, often denoted as M or $\tilde{\mu}$, is formally defined using the Cumulative Distribution Function (CDF), $F(x) = P(X \leq x)$.

For a Continuous Random Variable: The median is the value m that divides the area under the probability curve perfectly in half. It is the solution to the equation:

$$F(m) = \int_{-\infty}^{m} f(x)\,dx = 0.5$$

where $f(x)$ is the Probability Density Function (PDF). If the CDF $F(x)$ is strictly increasing, the median is uniquely defined as $m = F^{-1}(0.5)$.

For a Discrete or General Random Variable: Because the CDF of a discrete variable is a step function, an exact value where $F(x) = 0.5$ might not exist. Therefore, the median is rigorously defined as any value m that satisfies both of the following inequalities:

$$P(X \leq m) \geq 0.5 \quad \text{and} \quad P(X \geq m) \geq 0.5$$

B. Population Mode

The population mode is the value at which the probability mass or density is at its absolute highest.

For a Continuous Random Variable: The mode is the value x that maximizes the Probability Density Function (PDF), $f(x)$.

$$\text{Mode} = \arg\max_{x} f(x)$$

If $f(x)$ is continuously differentiable, modes can often be found by locating the critical points where the first derivative is zero and the second derivative is negative:

$$f'(x) = 0 \quad \text{and} \quad f''(x) < 0$$

For a Discrete Random Variable: The mode is the value x that maximizes the Probability Mass Function (PMF), $p(x) = P(X = x)$.

$$\text{Mode} = \arg\max_x p(x)$$

Note: A distribution can have a single global maximum (unimodal), multiple peaks (bimodal/multimodal), or no defined mode (such as a continuous uniform distribution where $f(x)$ is entirely constant).

C. Population Quantile

The population quantile is a direct generalization of the population median. While the median divides the probability distribution exactly in half, a quantile divides the distribution such that a specified proportion of the probability lies below it.

Let p be a probability such that $0 < p < 1$. The p-th population quantile (often corresponding to the $100p$-th percentile), denoted as x_p, is defined using the Cumulative Distribution Function (CDF), $F(x) = P(X \leq x)$.

For a Continuous Random Variable: If the random variable X is continuous with Probability Density Function (PDF) $f(x)$, the p-th quantile is the value x_p such that the area under the PDF curve to its left is exactly p:

$$F(x_p) = \int_{-\infty}^{x_p} f(x)\,dx = p$$

If the CDF is strictly increasing, the quantile can be found using the inverse CDF:

$$x_p = F^{-1}(p)$$

For a General or Discrete Random Variable: For discrete distributions, an exact value where the CDF equals p may not exist due to the step-like nature of the function. Therefore, the general and rigorous definition states that x_p is any value satisfying both of the following conditions:

$$P(X \leq x_p) \geq p \quad \text{and} \quad P(X \geq x_p) \geq 1 - p$$

Note: The population median is simply the special case where $p = 0.5$. Similarly, the first and third population quartiles correspond to $p = 0.25$ and $p = 0.75$, respectively.

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.

$$\text{Population Variance: } \sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

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

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:** $0, 100, 10, 140, 0$ ($\bar{x} = 50$)

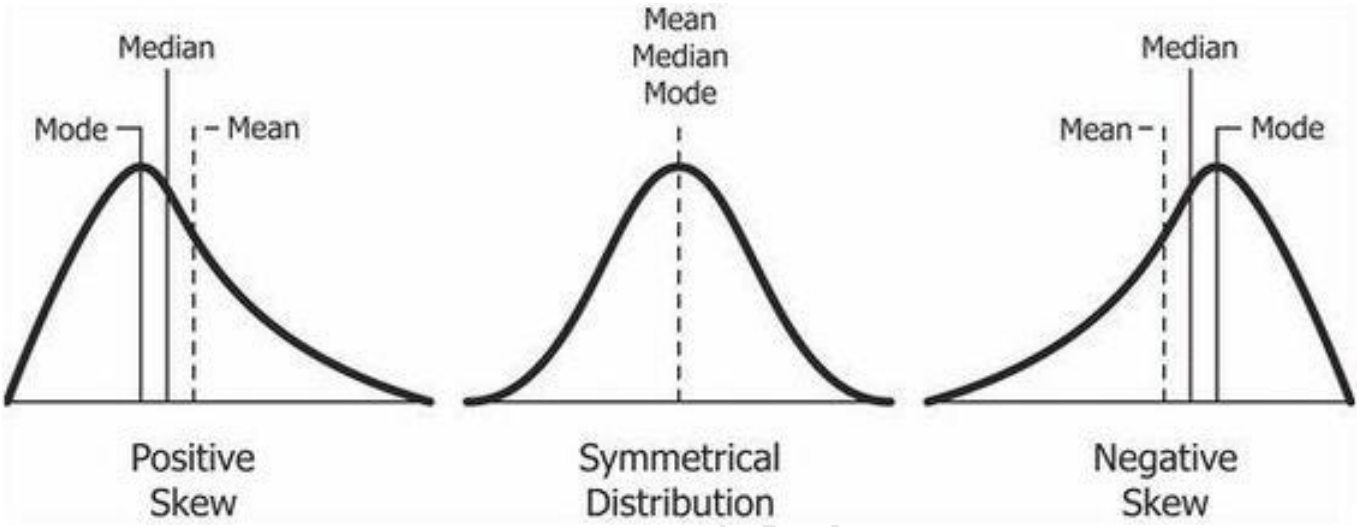
Both have the exact same mean (50 runs). However, calculating their sample standard deviations reveals a stark difference in dispersion. Batsman A has a very low s (high consistency), while Batsman B has a very high s (high volatility/heavy-tailed performance). Measures of dispersion are required to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

8. Higher-Order Moments: Skewness and Kurtosis

While the mean and variance describe the center and spread of a distribution, they do not describe its shape. To understand the asymmetry and the weight of the tails, we use the third and fourth standardized central moments: skewness and kurtosis.

A. Skewness

Skewness measures the degree of asymmetry of a distribution around its mean.

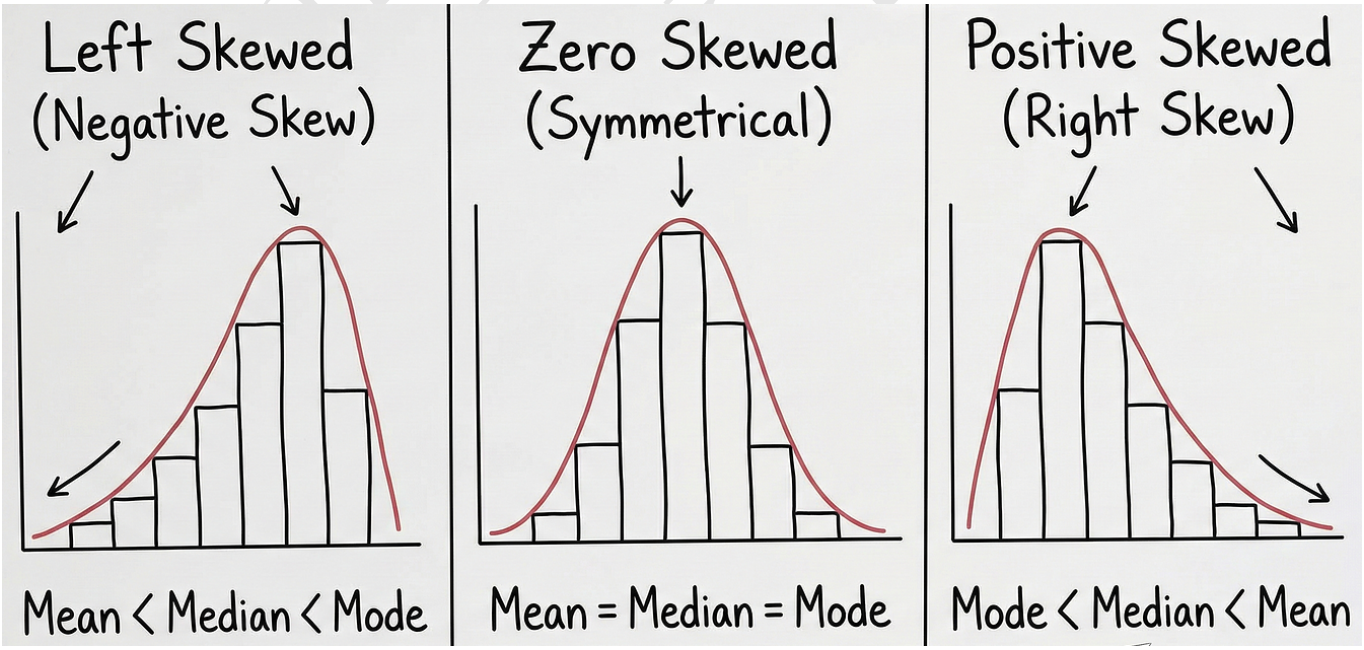


Population Skewness (γ_1): Let $\mu_3 = \mathbb{E}[(X-\mu)^3]$ be the third central moment of the population. The population skewness is the third central moment standardized by the population standard deviation (σ) cubed:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(X-\mu)^3]}{(\mathbb{E}[(X-\mu)^2])^{3/2}}$$

Sample Skewness (g_1): For a sample of size n, let $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$ be the k-th sample central moment. The sample skewness is estimated as:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}}$$



Interpretation and Examples:

- **Symmetric ($\gamma_1 = 0$):** The tails on both sides of the mean balance out perfectly. *Example:* A standard Normal distribution or a fair coin toss distribution.

- **Right-Skewed / Positive Skew ($\gamma_1 > 0$):** The right tail is longer or fatter than the left. The mass of the distribution is concentrated on the left. *Example:* Household income. Most people earn a low-to-medium salary, but a few billionaires create a long right tail.
- **Left-Skewed / Negative Skew ($\gamma_1 < 0$):** The left tail is longer or fatter. *Example:* Scores on a very easy exam. Most students score close to 100%, but a few very low scores drag the left tail out.

B. Kurtosis

Kurtosis measures the "tailedness" of a distribution—specifically, how often extreme outliers occur compared to a normal distribution.

Population Kurtosis (β_2): Let $\mu_4 = \mathbb{E}[(X-\mu)^4]$ be the fourth central moment. The population kurtosis is:

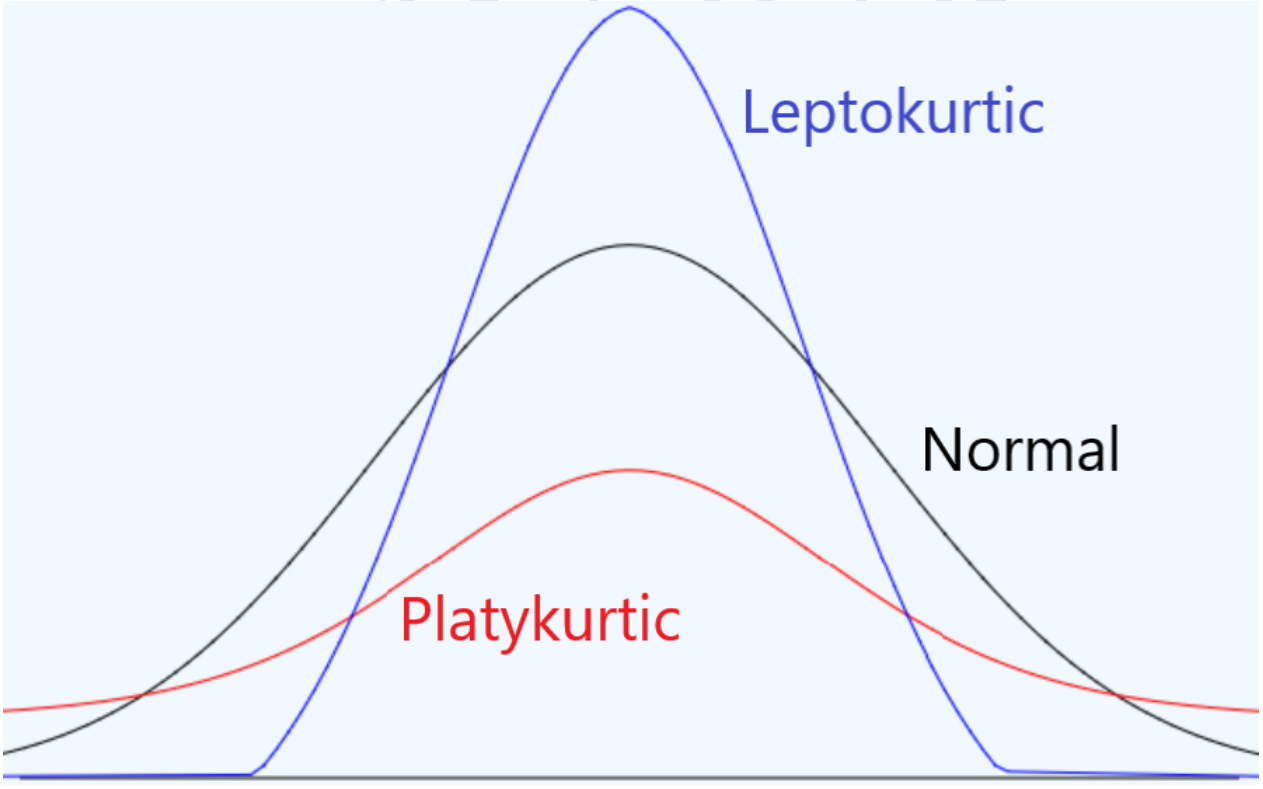
$$\beta_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X-\mu)^4]}{(\mathbb{E}[(X-\mu)^2])^2}$$

Because a normal distribution has a kurtosis of 3, we often define **Excess Kurtosis** as $\gamma_2 = \beta_2 - 3$.

Sample Kurtosis (g_2): Using the sample central moments m_k:

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$$

Sample excess kurtosis is simply $g_2 - 3$.



Interpretation and Examples:

- **Mesokurtic ($\beta_2 = 3$, Excess $= 0$):** The baseline shape of tails. *Example:* The Standard Normal distribution.

- **Leptokurtic ($\beta_2 > 3$, Excess > 0):** Distributions with fatter, heavier tails and a sharper peak than the normal distribution. This indicates a higher probability of extreme outlier events. *Example:* Daily returns of stock market indices often exhibit leptokurtic behavior, experiencing "black swan" extreme crashes or surges more often than a normal distribution predicts.

- **Platykurtic ($\beta_2 < 3$, Excess < 0):** Distributions with thinner tails and a flatter peak. Outliers are highly infrequent. *Example:* A Continuous Uniform distribution, where values are strictly bounded and cannot produce extreme outliers.

9. Why Should Engineers Care About Skewness and Kurtosis?

Mean and variance assume that systems behave perfectly symmetrically and predictably (like a Gaussian curve). However, real-world engineering systems are messy, constrained by physical limits, and prone to extreme, unexpected failures. Skewness and kurtosis are the diagnostic tools for this reality.

Applications of Skewness (Asymmetry)

Skewness tells an engineer that a physical process is pushing up against a boundary or that degradation is not symmetric.

- **Reliability and Maintenance Engineering:** The "time-to-failure" for mechanical components (like gears or structural beams) is almost never normally distributed; it is heavily **right-skewed** (often modeled by Weibull or Lognormal distributions). A component cannot fail in negative time (a hard lower boundary of zero), most components fail around their rated lifespan, but a few resilient ones survive much longer, dragging out the right tail.

- **Civil and Spatial Engineering:** Consider the modeling of traffic delays at major urban intersections. Delay times cannot be less than zero, creating a hard left boundary. Most vehicles experience a standard delay, but accidents or severe congestion create extreme, asymmetric delays on the right. An engineer designing a traffic flow model based purely on "mean delay" will drastically underestimate the severity of these right-skewed bottleneck events.

Applications of Kurtosis

Kurtosis is fundamentally a measure of risk and anomaly detection. A high kurtosis (leptokurtic) distribution warns an engineer that extreme, system-breaking outliers will happen much more frequently than a normal distribution would predict.

- **Signal Processing and Imagery Analysis:** When developing algorithms for object detection in satellite imagery or radar, engineers must filter out background noise. If the noise is assumed to be Gaussian (mesokurtic) but is actually leptokurtic (heavy-tailed), the frequent extreme noise spikes will be misclassified by the algorithm as actual objects (false positives). Understanding kurtosis is vital for designing robust statistical filters.

- **Mechanical Vibration Analysis:** Kurtosis is a primary metric for early fault detection in rotating machinery. A perfectly healthy motor produces continuous, normally distributed

vibration (Kurtosis ≈ 3). If a microscopic crack forms in a ball bearing, it will cause sharp, very infrequent "clicking" impacts. These sudden spikes will barely change the overall mean or variance of the vibration, but they will cause the kurtosis of the signal to skyrocket. By monitoring kurtosis, engineers can detect impending failures long before a machine destroys itself.

Key Takeaway: If you assume a system is perfectly normal (symmetric, mesokurtic) when it is actually heavily skewed or leptokurtic, you will drastically underestimate the probability of catastrophic failures or extreme events.

Correlation: Measuring Linear Relationships

While variance and standard deviation describe the behavior of a single variable, we often need to understand how two different variables interact. Correlation measures the strength and direction of the **linear** relationship between two quantitative variables.

A. Population Correlation Coefficient (ρ)

The population Pearson correlation coefficient, denoted by the Greek letter ρ (rho), describes the true linear relationship between two random variables X and Y in the entire population. It is defined as the covariance of the variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}}$$

Properties of ρ:

- **Boundedness:** $-1 \leq \rho \leq 1$.
- **Direction:** A positive ρ means that as X increases, Y tends to increase. A negative ρ means that as X increases, Y tends to decrease.
- **Strength:** A value of 1 or -1 indicates a perfect linear relationship. A value of 0 indicates no *linear* relationship (though a non-linear relationship might still exist).
- **Scale Invariance:** ρ is unitless. Changing the units of X or Y (e.g., from meters to centimeters) does not change the correlation.

B. Sample Correlation Coefficient (r)

In practice, we rarely know the population parameters. Instead, we estimate ρ using paired sample data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The sample correlation coefficient, r, is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

C. Examples of Correlation

- **Strong Positive Correlation ($r \approx 0.8$ to 1.0):** Engine horsepower and a car's top speed. As horsepower goes up, top speed consistently goes up.
- **Strong Negative Correlation ($r \approx -0.8$ to -1.0):** A vehicle's weight and its fuel efficiency (miles per gallon). As a car gets heavier, its fuel efficiency consistently drops.
- **Zero Correlation ($r \approx 0$):** A student's shoe size and their score on a statistics exam. Knowing one provides no information about the other.

D. Practical Uses of Correlation

- **Exploratory Data Analysis (EDA):** Before building complex models, engineers and statisticians compute a correlation matrix to quickly identify which variables are strongly related to the target outcome.
- **Feature Selection & Multicollinearity:** In machine learning and multiple regression, if two predictor variables are highly correlated with each other (e.g., measuring the same object in both inches and centimeters), it causes redundancy and model instability. Correlation helps identify and remove these redundant features.
- **Sensor Calibration:** An engineer might correlate the readings of a new, low-cost temperature sensor with those of a highly accurate, expensive laboratory reference sensor. A correlation near 1.0 proves the new sensor behaves linearly and can be reliably calibrated.

Remark 1. Correlation does not imply causation. Observing a high correlation between ice cream sales and shark attacks does not mean eating ice cream causes shark attacks; both are simply driven by a lurking third variable (summer weather).

Exploratory data analysis

See chapter 2 in Ross's book.

References

[1] Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability.* Chapman and Hall/CRC.

[2] Ross, Sheldon M. (2020). *Introduction to probability and statistics for engineers and scientists.* Academic press.

Disclaimer

This lecture note is prepared solely for teaching and academic purposes. Some parts of the material, including definitions, examples, and explanations, have been adapted or reproduced from the references. These notes are not intended for commercial distribution or publication, and all rights remain with the respective copyright holders.