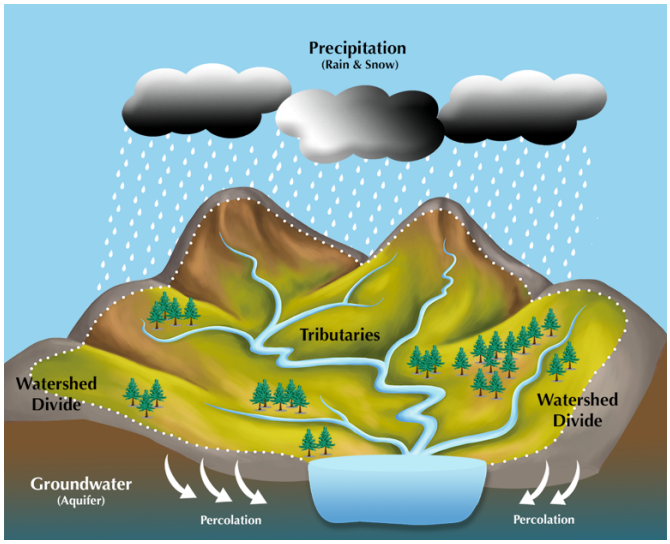
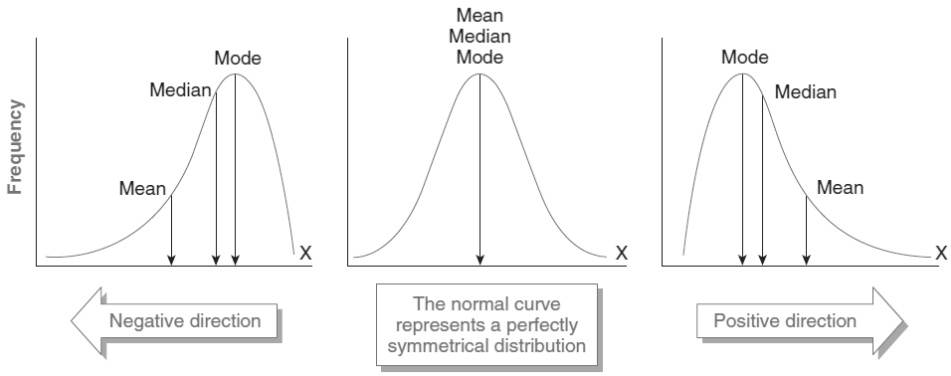
Basics of Statistics

EN5423 | Spring 2024

w02_statistics_basics.pdf
(Week 2)

Contents


Precipitation
(Rain & Snow)
Tributaries
Watershed Divide
Watershed Divide
Groundwater
(Aquifer)
Percolation
Percolation


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Summarizing Univariate Data

• In environmental science, data summarization is not merely about reducing data volume. It is about extracting meaningful insights from complex datasets, such as those related to water quality, hydrology, and aquatic ecosystems.

• The essence of summarizing univariate data lies in its ability to help scientists and decision-makers understand the underlying patterns, trends, and anomalies within a vast array of environmental data points.

• Summarization techniques enable the extraction of critical insights necessary for ecological assessments and policy formulation, highlighting significant patterns and anomalies.

• **Example Use**: Calculating the mean temperature of a lake over a year to monitor climate impacts.

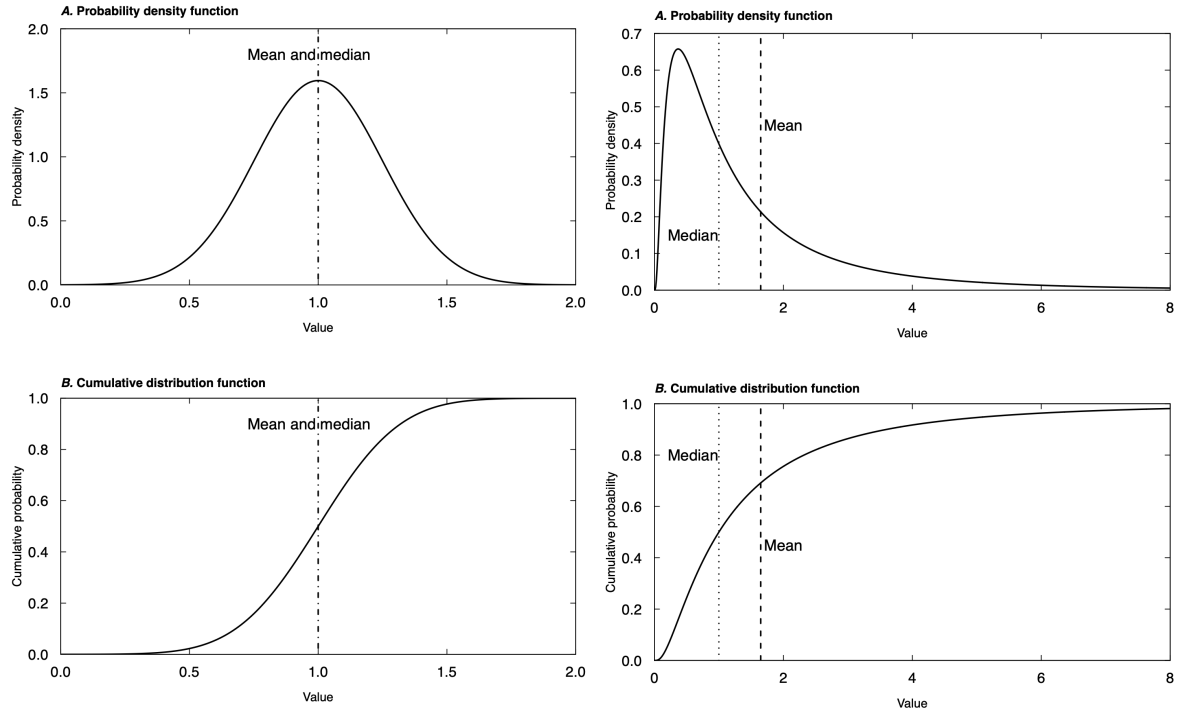


Figure 1. Graphs showing a normal distribution. A, the probability density function (pdf) of a normal (right: lognormal) distribution showing the mean and median of the distribution, which are identical. B, the cumulative distribution function (cdf) of the same distribution.

1 Measure of Location

Measures of location, or central tendency, provide essential insights into the typical conditions within an environmental dataset. These measures establish baselines for assessing changes, impacts, and the quality of resources.

1.1 A Classical Measure of Central Tendency—The Arithmetic Mean

The arithmetic mean, often referred to simply as the mean, is a fundamental measure of central tendency in statistics, particularly relevant in the analysis of water resources data.

1.1.1 Definition and Calculation:Classical Measure: The Mean (*Exercise 1*)

• The arithmetic mean (X) is calculated as the sum of all data values (X_i) divided by the sample size (n):

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

• This formula (Equation 1.1) provides a straightforward method for determining the average value of a dataset, which is essential for understanding the general behavior of water-related variables such as flow rates, precipitation, and contaminant levels.

1.1.2 Grouped Data

• For data categorized into k groups, the overall mean is determined by the mean of each group (X_i), weighted by the number of observations (n_i) in each group:

$$\overline{X} = \frac{\sum_{i=1}^{n} (X_i \cdot n_i)}{n} \tag{2}$$

• This adaptation allows for the analysis of data segmented by different criteria, such as time periods or geographical locations, providing a nuanced understanding of environmental data.

1.1.3 Influence of Individual Observations

• The influence of a single observation (X_j) on the mean can be significant, especially if it is an outlier. The overall mean can be disproportionately affected by observations that deviate markedly from the rest of the dataset.

$$\bar{X} = \bar{X}_{(j)} \frac{(n-1)}{n} + X_j \cdot \frac{1}{n} = \bar{X}_{(j)} + \left(X_j - \bar{X}_{(j)}\right) \cdot \frac{1}{n} \tag{3}$$

• This sensitivity highlights the mean's vulnerability to extreme values, underscoring the importance of considering other measures of central tendency in environmental data analysis, where outliers may represent extraordinary but important events like floods or droughts.

1.1.3.1 Visual Representation

• Imagining the mean as the balance point of data on a number line illustrates how data points further from the center exert more influence, akin to a lever.

• Removal of an outlier can significantly shift this balance point, reflecting the mean's lack of resistance to outliers.

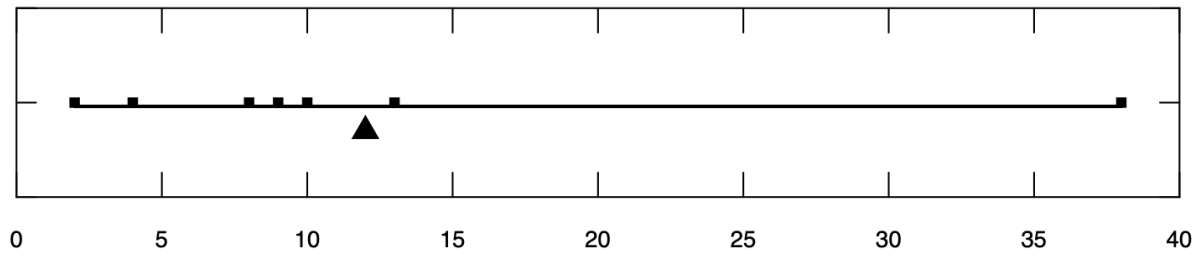


Figure 2. Graph showing the arithmetic mean (triangle) as the balance point of a dataset. The mean is 12.

Figure 3. Graph showing the shift of the arithmetic mean (triangle) downward after removal of an outlier. The mean is 7.67.

1.1.4 Application and Considerations

• While the mean provides a valuable summary statistic, its lack of resistance to outliers may not always make it the most reliable measure for central tendency in water resources data, which often contain extreme values due to natural variability or anthropogenic impacts.

• In cases where the total sum of a variable is of interest (e.g., total nutrient flux into a water body), the mean remains an essential measure. However, for characterizing typical conditions or values, more resistant measures like the median or mode might offer more robust insights.

Keypoints:

1) Reflects the average value, suitable for symmetric distributions without outliers.
2) Sensitive to extreme values, which can skew the results.

1.2 A Resistant Measure of Central Tendency—The Median (*Exercise 1*)

1.2.1 Definition and Calculation

• The median, or 50th percentile ($P_{0.50}$), is the central value of a dataset when sorted by magnitude.

$$\text{median} = P_{0.50} = \begin{cases} X\left(\frac{n+1}{2}\right) & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X\left(\frac{n}{2}\right) + X\left(\frac{n}{2}+1\right)\right) & \text{when } n \text{ is even} \end{cases} \quad (4)$$

• For an odd number of observations, the median is the middle data point.

• For an even number of observations, it is the average of the two middle data points.

• The median minimizes the impact of outliers by focusing on the central position within the sorted dataset.

1.2.2 Resistance to Outliers

• The median is notably resistant to changes in the value or presence of outlying observations. This property makes it a desirable measure of central tendency when data include extreme values.

Example 1
• Dataset (a) [2, 4, 8, 9, 11, 11, 12] has a mean of [ ] and a median of [ ]. • Dataset (b) [2, 4, 8, 9, 11, 11, 120] has a mean of [ ] and a median of [ ]. This shows the mean significantly increased to [ ] due to the outlier (120), while the median remains unchanged at [ ], showcasing its resistance to the outlier.

1.2.3 Preference Over Mean in Certain Situations

• When summarizing data that might be influenced by extreme observations, the median provides a more stable measure of central tendency than the mean.

• For instance, in assessing chemical concentrations across various streams, the median ensures that a single stream with an unusually high or low concentration does not disproportionately influence the overall estimate. This makes the median a more reliable indicator of typical conditions across the sampled locations.

Keypoints:

1) The median is the middle value in a sorted dataset, balancing the dataset by minimizing outlier impact.
2) Unlike the mean, the median is unaffected by extreme values, offering a stable central tendency in outlier-prone data.

1.3 Other Measures of Central Tendency

Three less commonly used measures of central tendency include the mode, geometric mean, and trimmed mean. The mode, identifying the most frequent value, suits discrete data well but requires binning for continuous data, making its value bin-dependent and less reliable for continuous datasets.

1.3.1 Mode

• **Definition**: The mode is the most frequently occurring value in a dataset. It is the only measure of central tendency that can be used with nominal data.

• **Application**: Useful in understanding common behaviors or preferences in a population, such as the most common land cover types.

• **Consideration**: A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all if all values are unique.

1.3.2 Geometric Mean (*Exercise 2*)

• **Definition**: The geometric mean is the nth root of the product of n numbers and is used for datasets that are multiplicative or have exponential growth.

• **Equation**:

$$GM = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \tag{5}$$

A simple way to calculate it is to take the mean of the logarithms of the data and then transform that value back to the original units.

$$GM = \exp(\bar{Y}) \tag{6}$$

where $Y_i = \ln(X_i)$ and $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

• **Application**: The GM is often reported for positively skewed datasets. It is only defined in cases where all data values are positive.

• **Consideration**: All data points must be positive as the geometric mean is not defined for negative or zero values.

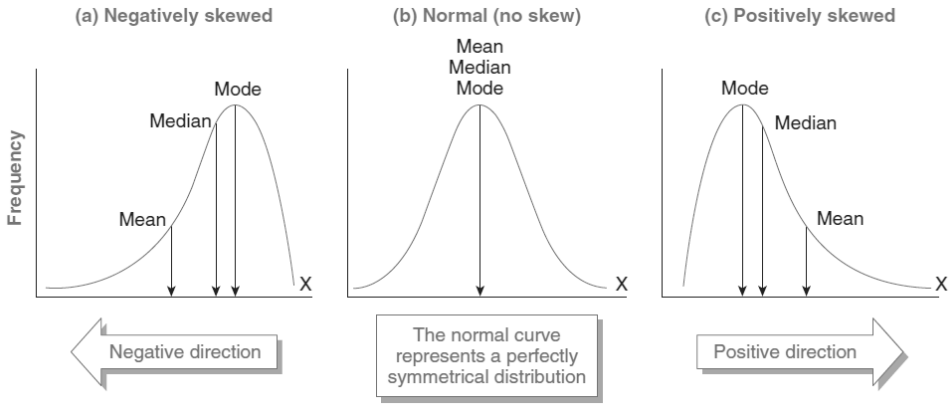


Figure 4. Negatively, Normal, and Posively skewed data sets.

1.3.3 Weighted Mean (*Exercise 3*) (*HW2 #2*)

• **Definition**: The weighted mean takes into account the relative importance or frequency of each data point, assigning weights that reflect each point's contribution to the overall dataset. Particularly useful when some data points contribute more to the dataset than others, such as in weighted grading systems or in watershed data where different sectors have different sizes.

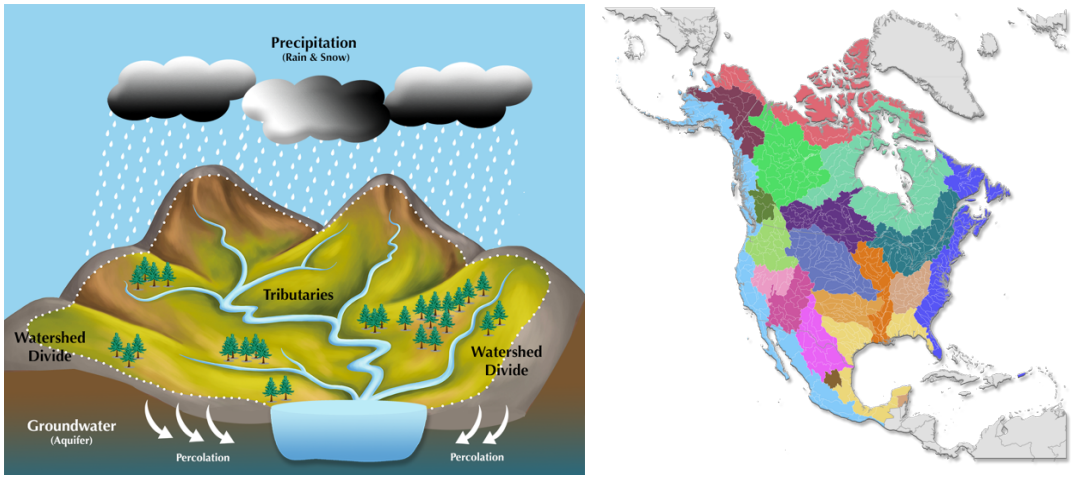


Figure 5. Watersehd and watershed map of North America

• **Equation**:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{7}$$

• **Application**: In environmental research, the weighted mean is crucial for analyzing data where different samples or variables have varying levels of significance, such as pollution levels from various sources contributing differently to overall water quality, or species observations weighted by their ecological importance.

• **Consideration**: Choosing appropriate weights is crucial for the weighted mean to accurately reflect the dataset's characteristics.

2 Measure of Variability (*Exercise 4*)

• Measures of variability, also known as measures of dispersion, quantify the spread or dispersion of data points within a dataset. These measures provide insight into the degree of variation or consistency among data points.

• Understanding the spread within environmental data is key to interpreting the natural variability of ecosystems, the impact of human activities, and the effectiveness of conservation efforts. Measures of variability give us a quantitative basis for these interpretations.

2.1 Classical Measures of Variability

Classical measures provide insights into the overall spread of data, crucial for evaluating the homogeneity or heterogeneity within environmental datasets.

2.1.1 Range

• Captures the simplest form of variability, highlighting the extent between extremes in datasets such as temperature ranges within a season.

• **Equation**:

$$Range = Max(X) - Min(X) \tag{8}$$

• **Application**: Assessing temperature fluctuations within a specific period in climatology.

• **Consideration**: Might overemphasize extreme events like wildfires or floods, affecting interpretation.

2.1.2 Variance

• Provides a detailed measure of dispersion by considering the squared deviations from the mean, useful in assessing variability in continuous environmental data like soil moisture levels.

• **Equation (Sample Variance)**:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{9}$$

• **Application**: Used for analyzing the variability in river discharge rates to manage water resources.

• **Consideration**: Effectiveness relies on normal distribution; skewed distributions common in environmental data might require data transformation.

2.1.3 Standard Deviation

• Translates variance into more interpretable units, aiding in understanding the spread of pollution levels across different water bodies.

• **Equation (Sample Standard Deviation)**:

$$s = \sqrt{s^2} \tag{10}$$

Keypoint (Range, Variance, & Standard Deviation):
Offer a straightforward assessment of data spread but vary in sensitivity to outliers and data distribution shapes. Their application ranges from understanding climate variability to analyzing spatial patterns of soil moisture and vegetation cover.

2.2 Resistant Measures of Variability

Resistant measures offer robust insights into data spread, reducing the influence of outliers, which are common in environmental data due to natural and anthropogenic factors.

2.2.1 Interquartile Range (IQR)

• Focuses on the middle 50% of data, offering a clearer view of central dispersion in datasets like annual rainfall amounts, unaffected by extreme events.

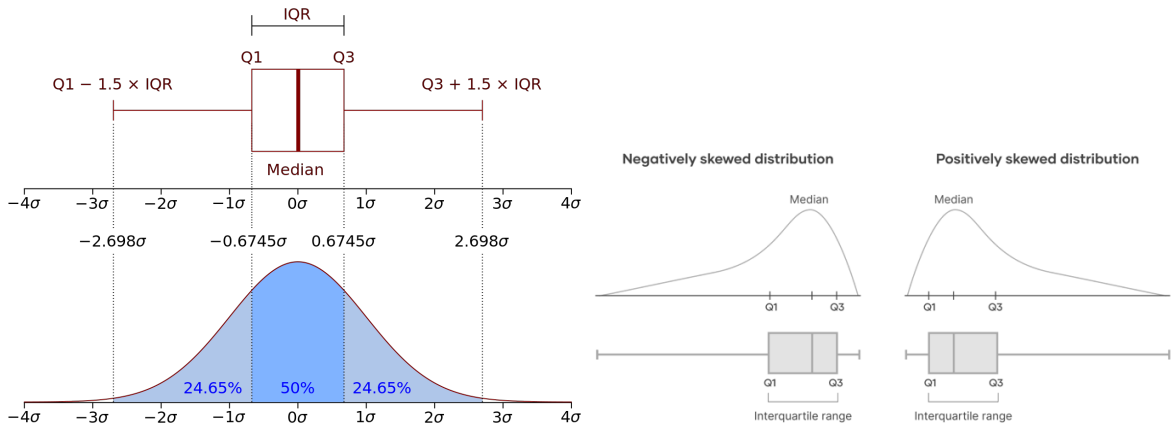


Figure 6. (left) Noraml distribution's IQR (right) skewed distributions' IQR

• **Application**: Identifying areas of concern by evaluating pollutant concentrations in water without influence from extreme contamination events.

• **Consideration**: Robust against outliers, but understanding data distribution is crucial for interpretation in heterogeneous landscapes.

2.2.2 Median Absolute Deviation (MAD)

• Provides a measure of variability centered around the median, useful for datasets with skewed distributions, such as the distribution of certain plant or animal species in fragmented habitats.

• **Equation**:

$$\text{MAD} = \text{median}\left(|x_i - \text{median}\,(X)|\right) \tag{11}$$

• **Application**: Assessing the consistency of air quality indices across urban areas.

• **Consideration**: Useful for skewed datasets or where outliers represent rare but critical events (e.g., peak pollution levels).

Keypoint (IQR & MAD):

These measures are invaluable for analyzing datasets characterized by skewness or extreme values, such as precipitation events, pollutant concentrations, and ecological data, ensuring that the analysis is not unduly influenced by anomalies.

2.2.3 The Coefficient of Variation—A Nondimensional Measure of Variability

• The CV allows for the comparison of variability across datasets with different scales, essential for cross-study comparisons in environmental science.

• Facilitates the comparison of relative variability, such as comparing the variability in temperature across different climatic zones or the variability in nutrient concentrations across different soil types.

• **Equation**:

$$\mathrm{CV} = \frac{s}{x}\ 100\% \tag{12}$$

• **Application**: Comparing the relative variability of precipitation across different climatic regions.

• **Consideration**: Assumes the mean is non-zero and meaningful, which might not hold for all environmental variables (e.g., days with no rainfall).

Keypoint:
The CV's ability to standardize variability relative to the mean makes it a powerful tool for comparing the consistency of environmental processes across different contexts, such as comparing water use efficiency among different plant species or variability in air quality measurements across urban and rural settings.

3 Measures of Distribution Symmetry (*Exercise 4*) (*HW2 #3*)

Symmetry in distribution is a crucial aspect of data analysis in environmental sciences, revealing the balance or imbalance in datasets. These measures help in understanding whether the data are skewed towards higher or lower values, indicating potential biases or underlying processes affecting the environment.

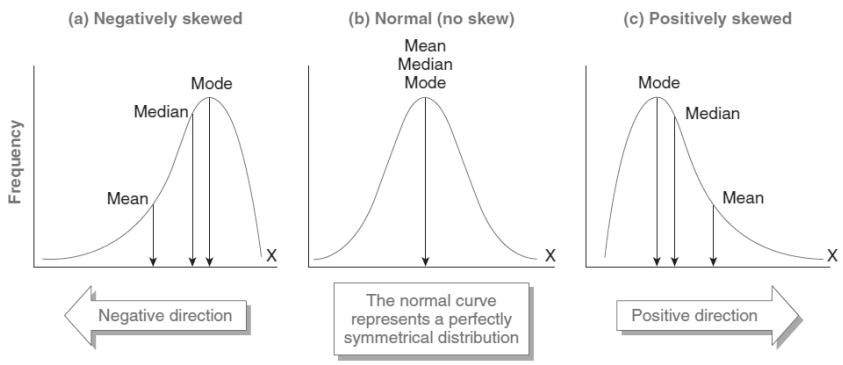


Figure 4 (again). Negatively, Normal, and Posively skewed data sets.

3.1 A Classical Measure of Symmetry—The Coefficient of Skewness

• Skewness quantifies the degree of asymmetry from the normal distribution in a dataset, with positive values indicating a tail to the right (more higher values) and negative values indicating a tail to the left (more lower values).

• Right-skewed distributions are indicated by a positive skewness coefficient (***g***), while left-skewed distributions have a negative ***g***. The presence of even a single outlier can significantly alter the skewness measure, potentially leading to misleading conclusions.

• For instance, altering one data point in a sample dataset from 12 to 120 changed its skewness coefficient from -0.84 to 2.6. Studies, including Monte Carlo tests, have shown that skewness coefficients can be highly biased in small sample sizes, typical in hydrology (often less than 100 observations), deviating notably from the population skewness. This bias and the sampling variability suggest caution in interpreting skewness in small datasets.

• An alternative, the L-moment method, offers a less biased and more outlier-resistant approach to assessing skewness, though it is noted that skewness measures, unless the sample size is significantly large, provide limited insight beyond distinguishing between right and left-skewed distributions.

• **Equation**:

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{13}$$

• **Application**: Analyzing skewness in precipitation data can help identify biases towards more frequent heavy rainfall events or more common light rain, crucial for flood risk assessment and water resource management.

• **Consideration**: While skewness provides insights into distribution shape, it is sensitive to outliers, which can significantly affect its value. Environmental data, often exhibiting extreme values due to natural events, require careful interpretation of skewness.

3.2 A Resistant Measure of Symmetry—The Quartile Skew

• The Quartile Skew, or Bowley's skewness, offers a resistant measure of distribution symmetry using quartiles, reducing the influence of outliers by focusing on the middle 50% of the data.

• **Equation**:

$$qs = \frac{(P_{0.75} - P_{0.5}) - (P_{0.5} - P_{0.25})}{P_{0.75} - P_{0.25}} = \frac{(Q3 - Q2) - (Q2 - Q1)}{Q3 - Q1} \tag{14}$$

• **Application**: Useful in evaluating the symmetry of ecological data, such as species richness across different habitats, where extreme values (very high or low species counts) might skew

traditional measures. The Quartile Skew provides a clearer picture of the central tendency and distribution shape.

• **Consideration**: Quartile Skew is particularly beneficial for skewed environmental datasets, ensuring that the analysis reflects more typical conditions rather than being dominated by extremes. However, it gives less weight to the tails of the distribution, which might contain ecologically significant information in certain studies.

Keypoint (Skweness):
Even if sample sizes are large (well above 100 samples), skewness coefficients computed using equation (13) are not very informative except to the extent that they may distinguish between right-skewed and left-skewed populations.

4 Ourliers (*Exercise 4*)

Outliers are data points that significantly deviate from the rest of the data in a dataset. They are important in statistical analyses because they can have a profound impact on the results, sometimes leading to misleading conclusions if not appropriately handled. However, outliers may be the most important points in the dataset, and should be investigated further. If outliers are deleted, it creates the risk that those who use the dataset will only see what they expected to see and may miss gaining important new information. The graphical methods are very helpful for identifying outliers. Outliers typically have one of these three causes:

1. A measurement or recording error;
2. An observation from a different population than most of the data, such as a flood caused by a dam break rather than by precipitation or a concentration resulting from a brief chemical spill into a river; or
3. A rare event from a single population; for example, if floods are always caused by rainfall events, the outlier may arise simply because the rainfall was extreme.

• **Definition**: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or statistical method) to determine what will be considered abnormal.

• **Applications**:

1. **Extreme Weather Events**: Outliers can indicate rare but significant weather events, such as unprecedented rainfall or drought conditions. Analyzing these outliers helps in understanding and preparing for extreme weather impacts.
2. **Pollution Incident Detection**: Outliers in data on air or water quality can signal pollution incidents, such as chemical spills or unauthorized discharges, necessitating immediate attention and remediation efforts.
3. **Ecological Anomalies**: Unusual observations in species counts or biodiversity metrics may point to ecological anomalies, including the sudden appearance or disappearance of species, often linked to environmental changes or threats.
4. **Climate Change Indicators**: Outliers in long-term temperature, precipitation, or sea-level data can serve as early indicators of climate change, highlighting shifts from historical patterns.

• **Consideratons**:

1. **Validation Before Exclusion**: It is essential to investigate outliers for potential errors or data quality issues. However, valid outliers should be retained as they may contain valuable environmental insights.
2. **Impact on Statistical Analyses**: Outliers can significantly affect the results of statistical analyses, skewing measures of central tendency and variability. Choosing robust statistical methods that are less sensitive to outliers is often necessary.
3. **Representing Natural Variability**: In environmental science, outliers may represent natural variability rather than anomalies. The decision to exclude outliers should consider the ecological or physical processes that could produce such values.
4. **Ethical and Scientific Integrity**: The decision to remove or retain outliers should be based on scientific rationale rather than convenience or the desire to achieve a specific outcome, ensuring the integrity of the research findings.

5 Transformations (*Exercise 5*) (*HW2 #4*)

Transformations in data analysis are used to modify the scale or distribution of data, making it more suitable for statistical analysis, especially when the original data do not meet the assumptions of statistical tests or models. There are three common reasons to consider transformations of the data (and often more than one of them are involved):

1. To make data distributions more symmetric,

2. To make relations between variables more linear, and

3. To make variability more constant.

• Purpose and Types of Transformations

- Normalization: Making the data conform to a normal distribution if skewed or containing outliers. Common methods include logarithmic, square root, and Box-Cox transformations.

- Stabilizing Variance: When the variance of a dataset increases with the mean (heteroscedasticity), transformations can help stabilize the variance across the range of data. This is often the case in environmental data, such as pollutant concentrations or meteorological measurements.

- Linearizing Relationships: Transforming variables can help linearize relationships between variables, making linear regression models applicable and improving model fit and interpretation.

• Common Transformations

- Logarithmic Transformation: Used to reduce right skewness; particularly effective for data with exponential growth or multiplicative error structures.

$$y' = \log(y) \tag{15}$$

> Useful for data that exhibit exponential growth or are highly skewed right. It can help stabilize variance and make the data more normally distributed.

- Square Root Transformation: Often applied to count data or rates, reducing right skewness and stabilizing variance.

$$y' = \sqrt{y} \tag{16}$$

> Often applied to count data or rates. It reduces right skewness and is particularly helpful when dealing with data that follow a Poisson distribution.

- Box-Cox Transformation: A parametric transformation that generalizes log and square root transformations, systematically varying a parameter to find the transformation that best normalizes the data.

$$y'(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \tag{17}$$

> A family of power transformations that includes both logarithmic and square root transformations as special cases. It systematically varies λλ to find the best approximation to normality.

- Arcsine Square Root Transformation

$$y' = \arcsin(\sqrt{y}) \tag{18}$$

> Typically used for proportions or percentages. It helps normalize data bounded by 0 and 1, like binomial proportions.

- Reciprocal Transformation

$$y' = \frac{1}{\mathrm{y}} \tag{19}$$

> Effective for dealing with certain types of skewness and can be useful for rate data.

- Reciprocal Transformation

$$y' = \mathrm{e}^{\mathrm{y}} \; or \; 10^{y} \tag{20}$$

> Useful when dealing with multiplicative effects and to counteract logarithmic transformations.

• Applications:

1) **Pollution Data Analysis**: Transformations can help normalize data on pollutant concentrations, facilitating the use of parametric tests to assess differences or trends.
2) **Ecological Indices**: Transforming species abundance data can make them more amenable to multivariate analysis, revealing underlying patterns in biodiversity.
3) **Climate Model Outputs**: Transforming skewed outputs from climate models can improve the accuracy of statistical downscaling or trend analysis.

• Considerations:

1) **Interpretation**: Transformations can complicate the interpretation of results, as effects sizes, differences, and relationships are in the transformed scale. Back-transforming results for interpretation requires caution.
2) **Choice of Transformation**: The choice depends on the data distribution and the research question. No one-size-fits-all; it requires exploratory data analysis and possibly trying multiple transformations.
3) **Data Zeroes**: Some transformations (e.g., log) are undefined for zero values. Adding a small constant to the data before transformation is a common workaround, but this can influence results and interpretations.

Table 1.1. Ladder of powers as modified from Velleman and Hoaglin (1981).
[θ, the power exponent; -, not applicable]

θ	Transformation	Name	Comment
		Used for negatively skewed distributions	
i	x^i	ith power	-
3	x^3	Cube	-
2	x^2	Square	-
		Original units	
1	x	Original units	No transformation.
		Used for positively skewed distributions	
1/2	$\sqrt{x}$	Square root	Commonly used.
1/3	$\sqrt[3]{x}$	Cube root	Commonly used. Approximates a gamma distribution.
0	$\log(x)$	Logarithm	Very commonly used. Holds the place of x^0.
−1/2	$-1/\sqrt{x}$	Negative square root	The minus sign preserves the order of observations.
−1	$-1/x$	Negative reciprocal	-
−2	$-1/x^2$	Negative squared reciprocal	-
$-i$	$-1/x^i$	Negative ith reciprocal	-