Descriptive Statistics

A warm guide to variables, distributions, and numerical summaries

Ratiomera Statistics

Table of contents

Course-Page Summary

Descriptive statistics connect measurement to analysis. Begin by identifying what each case, variable, and value represents, then use the measurement level to choose defensible frequencies, numerical summaries, and graphs. Center, spread, and shape must be interpreted together because none provides a complete description alone. These ideas support probability, inference, correlation, regression, and every later topic in this learning sequence.

Notation and Decision Checklist

Table 1: Core notation used in descriptive statistics.

Symbol	Meaning	Question to ask
x_i	Value of observation i	What does one recorded value represent?
n and N	Sample size and population size	Am I describing a sample or the complete population?
$\bar{x}$ and μ	Sample mean and population mean	Is the symbol consistent with the target?
s^2 and σ^2	Sample variance and population variance	Does the denominator match the target?
s and σ	Sample SD and population SD	What is the original measurement unit?
$\sum$	Add the stated expression across its index range	What is the summand, and what are the bounds?

Before reporting a result, identify the cases and variables, inspect missing or impossible values, determine the measurement level, examine the distribution, select a compatible center and spread, and state what the summaries cannot establish about the population or causality.

Levels of Measurement

Five scale levels determine which operations and summaries are valid. Each level adds properties to the one above it.

Table 2: Measurement properties, admissible transformations, and core summaries.

Scale	Defining property	Admissible recoding	Core summaries
Nominal	Categories without order	One-to-one relabeling	Counts, proportions, mode
Ordinal	Ordered categories, unequal spacing	Strictly increasing recoding	Counts, proportions, median, percentiles, mode
Interval	Equal spacing, assigned zero	$y = a + bx,\ b > 0$	Mean, SD, differences, correlation
Ratio	Equal spacing, true zero	$y = bx,\ b > 0$	Interval summaries plus meaningful ratios
Absolute	True zero and fixed natural unit	$y = x$	Ratio summaries and counts

Interval, ratio, and absolute scales are collectively called **metric** scales.

Frequency Distributions

Table 3: Compact reference for frequency distributions.

Quantity	Formula	Interpretation
Absolute frequency	n_j	Number of observations in category j
Relative frequency	$f_j = n_j/n$	Proportion in category j
Cumulative relative frequency	$F_j = \sum_{k=1}^{j} f_k$	Proportion at or below ordered category j

Absolute frequencies sum to n and relative frequencies sum to 1, apart from rounding. Cumulative frequencies require an ordered variable and are not meaningful for nominal categories.

Central Tendency

Three measures describe the typical value:

Table 4: Measures of central tendency and their uses.

Measure	Formula	When to use
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Symmetric distributions; metric data
Median	Middle value when sorted	Skewed distributions; robust to outliers
Mode	Most frequent value or category	Any scale; may be tied or nonunique

When the mean is noticeably higher than the median, that pattern can suggest right skew, especially when a histogram also shows a long right tail. The ordering of the two statistics alone does not prove the distribution's shape.

Variability

Four measures describe how much individuals differ:

Table 5: Measures of variability and their uses.

Measure	Formula	When to use
Range	$x_{\max} - x_{\min}$	Direct overview; sensitive to extremes
IQR	$Q_3 - Q_1$	Robust spread; paired with median
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$	Building block for SD and other statistics
SD	$s = \sqrt{s^2}$	Scale of deviations around the mean; paired with mean

The values $Q_1 - 1.5 \cdot \mathrm{IQR}$ and $Q_3 + 1.5 \cdot \mathrm{IQR}$ define the lower and upper fences. Whiskers extend to the most extreme observations inside those fences. Values beyond the fences are plotted separately as potential outliers.

Linear Transformations

A linear transformation $y_i = a + b \cdot x_i$ has predictable effects on summary statistics:

Table 6: Effects of a linear transformation on descriptive statistics.

Quantity	Transformation rule	Effect
Mean	$\bar{y} = a + b\bar{x}$	Shifted by a and scaled by b
Variance	$s_y^2 = b^2 s_x^2$	Scaled by b^2; unaffected by a
Standard deviation	$s_y = \lvert b \rvert s_x$	Scaled by $\lvert b \rvert$; unaffected by a
Rank and orientation	Preserved when $b > 0$	Reversed when $b < 0$

The transformed mean follows $\bar{y} = a + b\bar{x}$. The shift a moves the mean but does not affect variability, while $|b|$ scales distances and the standard deviation. A negative b also reflects the distribution and reverses ranks.

Z-Transformation

For a nonconstant sample with a defined, nonzero s, the z-transformation $z_i = (x_i - \bar{x})/s$ produces a sample mean of 0 and a sample SD of 1.

A z-score tells you how many standard deviations above or below the reference mean an observation lies. Comparing z-scores across instruments requires the same construct, suitable reference groups, and comparable measurement quality.

Diagram Selection by Scale Level

Table 7: Diagram selection by measurement level.

Scale level	Appropriate diagram
Nominal	Bar chart, pie chart
Ordinal	Ordered bar chart
Metric (interval, ratio, absolute)	Histogram, boxplot

In histograms, **area is proportional to frequency**. With equal bin widths, height and area both encode frequency correctly. With unequal widths, density height is f_j/w_j. Compare sensible bin widths because the apparent shape can change with the grouping.

Distribution Shapes

Table 8: Independent dimensions used to describe distribution shape.

Shape	Key feature	Interpretive implication
Symmetric	Equal tails on both sides	Mean and median are close
Right-skewed	Long tail toward high values	Mean often exceeds median; verify visually
Left-skewed	Long tail toward low values	Mean often falls below median; verify visually
Unimodal	Single peak	One main concentration, summarized according to its shape
Bimodal	Two peaks	Show and investigate both peaks; do not assume subgroups
High kurtosis	Greater weight on extreme deviations	More prone to distant observations
Low kurtosis	Less weight on extreme deviations	Less prone to distant observations

Symmetry, modality, and kurtosis can coexist in different combinations. Peak height alone does not define kurtosis.

Suggestive Graphics

Graphs can create misleading impressions through truncated axes, compressed or stretched scales, and inappropriate histogram bins. Percentages can mislead when their denominator, sample size, measurement method, or missing data are concealed. Both graphics and numbers require context and valid measurement.

What You Should Now Be Able to Do

You should now be able to identify cases, variables, values, and measurement levels; read the notation used in means, variances, and z-scores; build and interpret frequency summaries; choose compatible measures of center and spread; compare distributions with tables and graphs; predict the effects of linear transformations; standardize observations; and recognize presentation choices that can distort a statistical message. You should also be able to state the limits of a descriptive conclusion without turning an observed pattern into a population or causal claim.

Expanded Reference

Purpose and foundations

Descriptive statistics gives you a disciplined way to turn a collection of observations into an understandable account of what was observed. Begin by identifying the **cases**, the people or units represented by the rows of a dataset, and the **variables**, the characteristics recorded in its columns. A value is one recorded result for one variable and one case. Before calculating anything, ask what the variable means, how it was measured, and which values are possible. This prevents a mathematically correct calculation from becoming a misleading description.

The measurement scale guides what you may sensibly do with a variable. A nominal variable separates cases into categories without an order. An ordinal variable has ordered categories, but the distances between neighboring categories are not known to be equal. An interval variable has meaningful equal distances but no meaningful absolute zero. A ratio variable has equal distances and a meaningful zero, so ratios can be interpreted. The scale is a property of how a variable is defined and measured, not of how its values happen to look in one dataset.

Scale	What its values tell you	Suitable first summaries
Nominal	Whether cases belong to the same or different categories	Counts, proportions, mode
Ordinal	Category membership and order	Counts, proportions, median, quantiles
Interval/ratio	Order and meaningful numerical distance	Mean, median, variance, standard deviation, quantiles

Some teaching datasets in this learning sequence are **simulated data**, meaning computer-created values that follow stated rules rather than measurements collected from real people. A **simulation** is the process that creates those values. A computer uses a **random-number generator**, an algorithm designed to produce values that behave like chance outcomes. A **seed** is the starting value supplied to that algorithm. Reusing the same seed with the same instructions recreates the same dataset. This makes a teaching example reproducible: you and another learner can inspect the same observations and obtain the same results. Simulation supports learning, but it does not turn created values into evidence about a real population.

Core ideas

A distribution describes how the values of a variable are spread across their possible range. For a categorical variable, start with a frequency table. An absolute frequency is the number of cases in a category. A relative frequency is that count divided by the total number of valid cases. Relative frequencies can be reported as proportions or percentages. Always check whether missing values were excluded from the denominator, because a percentage is meaningful only when its reference total is known.

For a numerical variable, describe four features together: center, variability, shape, and unusual observations. The mean uses every value and represents the balance point of the distribution. The median is the middle ordered value and divides the data into two halves. The mode is the most frequent value or category. The range runs from the smallest to the largest value. The interquartile range covers the middle half of the ordered observations. Variance and standard deviation summarize how far values tend to lie from the mean.

Question	Useful evidence	Reading habit
Where is the distribution centered?	Mean, median, and sometimes mode	Compare mean and median instead of reporting one number alone
How much do observations differ?	Range, interquartile range, variance, standard deviation	State the units and watch for unusual values
What shape do the values form?	Histogram, boxplot, frequencies, skewness	Look for symmetry, skew, gaps, clusters, and more than one peak
Are any values surprising?	Raw values, plot, data checks, standardized scores	Investigate before deciding whether a value is an error

Shape affects interpretation. In a roughly symmetric distribution, the mean and median are often similar. A long right tail tends to pull the mean upward, while a long left tail tends to pull it downward. **Modality** describes the number and pattern of clear peaks or main concentrations. A distribution may be unimodal, with one main peak, or multimodal, with more than one concentration. **Kurtosis** describes how readily values occur far into the tails relative to a symmetric bell-shaped reference distribution with the same overall spread. Peak height alone does not define kurtosis. A number such as the mean cannot show these features, which is why a graph and numerical summaries should be read together.

An unusual observation is not automatically a mistake. It may be a valid but rare case, a coding error, a measurement problem, or a sign that different groups have been combined. Check the original definition and recording process before excluding anything. If an analytical decision changes after an observation is removed, report that sensitivity rather than hiding it.

Formula guide

For category or interval j, let n_j be its absolute frequency and let n be the number of valid observations. Its relative frequency is

$$f_j = \frac{n_j}{n}.$$

For ordered categories or numerical intervals, the cumulative relative frequency through category j is

$$F_j = \sum_{h=1}^{j} f_h.$$

Absolute frequencies should add to n, relative frequencies should add to 1 apart from rounding, and the final cumulative relative frequency should equal 1. Cumulative frequencies are meaningful only when the categories have a defensible order.

Let $x_1, x_2, ..., x_n$ be the observed values of a numerical variable. The sample mean adds all values and divides by the number of observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The notation $\sum$ means “add the indicated values.” The index i identifies one observation at a time, from the first observation to observation n. The deviation $x_i - \bar{x}$ tells you how far one value lies above or below the mean. Positive and negative deviations cancel when added, so variance first squares them. The sample variance uses $n - 1$ in the denominator:

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

The corresponding population variance uses the population mean μ and divides by the population size N:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Keep the two targets separate. The denominator $n-1$ belongs to the corrected sample variance used to estimate population variability, while division by N describes the complete population values themselves.

Because variance is expressed in squared units, take its square root to return to the original measurement unit. The sample standard deviation is therefore:

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

A standardized score expresses one value in standard-deviation units. It subtracts the sample mean and divides by the sample standard deviation:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A positive z_i places the value above the mean, a negative z_i places it below the mean, and the magnitude reports the distance in standard deviations. Standardization changes the unit and reference point, but it does not change the order or shape of the observations.

The median and quantiles begin with the ordered observations. The first quartile Q_1 marks the lower quarter, the median Q_2 marks the halfway point, and the third quartile Q_3 marks the lower three quarters. The range and interquartile range are

$$\text{range} = x_{\max} - x_{\min}, \qquad IQR = Q_3 - Q_1.$$

Sample-quantile conventions can interpolate differently, so software may report slightly different quartiles for a small dataset. A common boxplot diagnostic uses the inner fences

$$Q_1 - 1.5(IQR) \qquad \text{and} \qquad Q_3 + 1.5(IQR).$$

Values beyond a fence are potential outliers to investigate, not automatic errors to delete. The whiskers stop at the most extreme observed values still inside the fences; they do not necessarily end at the fence values themselves.

For a linear transformation $Y = a + bX$, the center and spread change according to

$$\bar{y} = a + b\bar{x}, \qquad s_y^2 = b^2 s_x^2, \qquad s_y = |b| s_x.$$

The shift a changes location but not spread. The multiplier b changes distances by $|b|$, so variance changes by b^2. Standardization is the special case that subtracts the mean and divides by the standard deviation.

Histogram height requires one final check. If bin j has relative frequency f_j and width w_j, its density height is

$$h_j = \frac{f_j}{w_j}.$$

The bar area is then $h_j w_j = f_j$. Equal-width bins allow height to track frequency directly. Unequal-width bins require density heights so that area, rather than height alone, continues to represent frequency.

Reading the explanatory figure

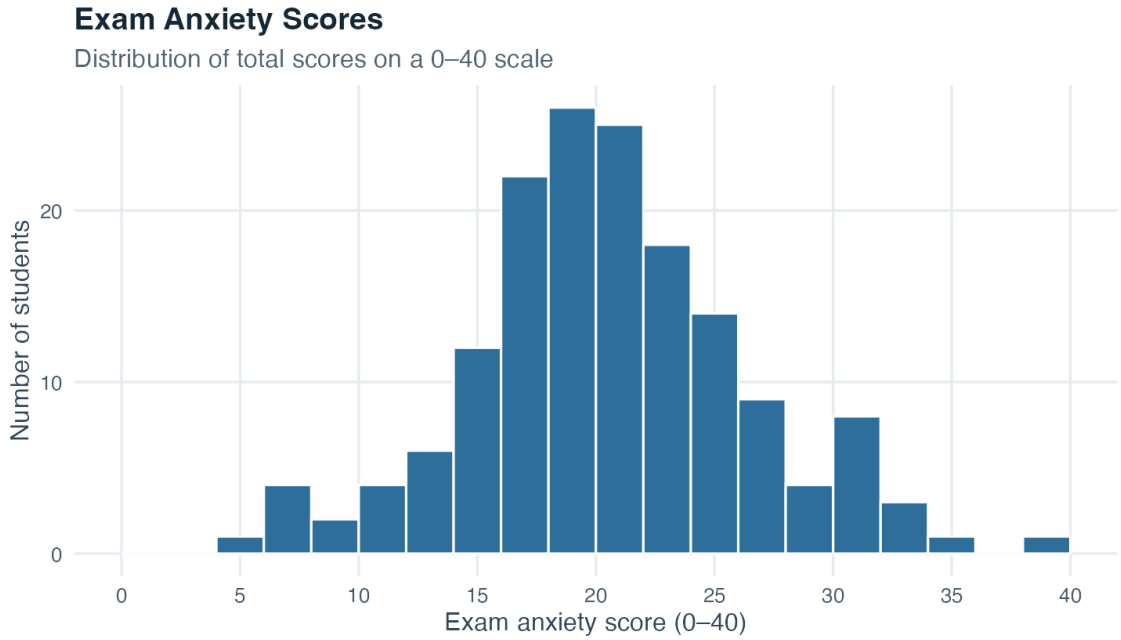

Figure 1: Histogram of simulated exam anxiety scores from zero to forty, with the tallest bars near the center and thinner tails toward both ends.

Read the horizontal axis first. It gives exam-anxiety scores on a scale from 0 to 40. The vertical axis gives the number of students, so the height of each bar is a frequency. Bars around scores 18 to 22 are the tallest, which places the main concentration near the center of the scale. Fewer observations appear at the lower and upper ends. One bar close to 40 shows that at least one high value is present, but the graph alone does not tell you whether that value is erroneous. You would return to the data definition and recording process before making that judgment.

The bars touch because a numerical scale is continuous across neighboring intervals. Their widths matter: changing the interval boundaries can make the same observations look more or less detailed. The figure should therefore be read as a view of one distribution, not as a set of separate categories. It supports statements about center, spread, shape, and unusual observations. It does not identify a cause of anxiety, compare a population with another population, or establish that the simulated pattern occurs among real students.

Interpretation checklist

Start every descriptive account by naming the cases, the variable, its scale, and the valid range. State the number of valid and missing observations. For a categorical variable, report counts together with their denominator and proportions. For a numerical variable, pair a plot with measures of center and variability. Use the mean and standard deviation when their interpretation fits the distribution, and also examine the median and interquartile range when skew or unusual observations matter.

Keep the unit visible. A standard deviation of five points means something different from five hours. Do not describe a group as homogeneous or variable without a reference that makes the comparison meaningful. Check whether a rounded table still adds to the expected total. Label simulated data as

simulated. Finally, separate description from explanation: a pattern in the observed values tells you what the dataset contains, while a causal explanation requires a research design and reasoning beyond descriptive statistics.

How this topic connects

Descriptive statistics supplies the language used throughout this learning sequence. Probability adds rules for reasoning about uncertain outcomes. Statistical inference then uses a sample and its variability to make careful statements about a population. Covariance and correlation ask how two variables vary together. Regression expresses an outcome as a function of one or more predictors, while partial correlation studies an association after linear adjustment. Analysis of variance compares group means by dividing total variability into meaningful components.

The central habit carries forward unchanged: understand the variables, inspect the distribution, calculate an appropriate summary, and interpret it in context. Later methods add uncertainty and models, but they never remove the need for a trustworthy description of the data that entered the analysis.

Document ID: `topic-01-descriptive-statistics-summary-en`

Notes 1 of 3

Descriptive Statistics | Topic 1

Notes 2 of 3

Descriptive Statistics | Topic 1

Notes 3 of 3