

Quantitative data analysis in financial literacy evaluation research – visualization and statistical inference

This article introduces the two main types of quantitative data analysis: descriptive statistics and statistical inference. First, we focus on data visualization, a key component of descriptive statistics that not only summarizes data but can also improve the clarity and communication of results. We then explain a number of key concepts underlying statistical inference which researchers use to draw conclusions about a larger population based on data from a sample.

Authors

Theresa Lorenz
OeNB, Financial Literacy and Culture Division
theresa.lorenz@oenb.at

Maximilian Zieser
OeNB, Financial Literacy and Culture Division
maximilian.zieser@oenb.at

Sofia Anyfantaki
Bank of Greece, Economic Analysis and Research Department
sanyfantaki@bankofgreece.gr



Effective data visualization

Effective data visuals use clear axes, intuitive colors, legible labels and minimal design. Researchers should avoid distortion, reduce clutter and match complexity to the audience. Annotations and highlights help guide attention and improve understanding.



Statistical inference for impact evaluations

Descriptive statistics alone are often insufficient for impact analysis, as they do not assess whether observed differences are specific to the sample or reflect broader patterns in the population. Using significance testing, inferences can be drawn from a random sample to arrive at conclusions about a larger population.

Opinions expressed by the authors of studies do not necessarily reflect the official viewpoint of the Oesterreichische Nationalbank, the Bank of Greece or the Eurosystem.

Abstract

To evaluate financial literacy programs, researchers often work with quantitative data. To derive meaningful insights from raw data requires data analysis. In this article, we introduce two types of quantitative data analysis: descriptive statistics and statistical inference. Data visualization is a powerful way of descriptive data analysis but is frequently overlooked in evaluation handbooks. This article therefore emphasizes effective methods of data visualization. We show different types of visualizations that are effective for plotting time series, proportions, relationships and distributions. Additionally, we provide general guidelines to ensure an accurate representation of data and facilitate the effective interpretation of the intended message. For impact evaluations, however, descriptive statistics alone may be insufficient. Methods of statistical inference enable us to gain insights into populations from sample data and quantify the uncertainty of these insights. We thus provide an introduction to statistical inference by focusing on the principles behind hypothesis and significance testing.

Introduction

Financial literacy evaluation research aims to assess the impact and underlying processes of financial literacy interventions. To this end, researchers must collect data, often in the form of structured quantitative information, e.g. from questionnaires or knowledge tests. However, raw data typically must be summarized to a high degree to enable general statements about the effectiveness of the intervention in question or of other relevant processes or associations. Data collected for this purpose can be qualitative or quantitative. Generally, each data type requires and allows for very different analysis techniques. Quantitative data (i.e. numbers in some sort of structured format) are analyzed using *statistical methods*. In contrast, qualitative data rely on textual information and a very different set of analysis techniques. In this paper, we provide techniques for quantitative data analysis, while qualitative and mixed-method approaches are discussed in other parts of the Financial Literacy Evaluation Series (see Felbermayr, 2024, and Lorenz, 2024).

The choice of quantitative analysis techniques depends on several factors: The first thing to consider is the goal of the analysis. Methods for quantitative data analysis can be assigned to either *descriptive statistics* or *statistical inference*. The goal of descriptive statistics is to summarize, visualize or identify patterns in the collected data. However, if the aim is to quantify the uncertainty of conclusions about a broader population (i.e. the target population of a financial literacy intervention) or test hypotheses, statistical inference is required. Because inferential statistics *infer* generalizations about the target population from a sample statistic, the related techniques must be based on a representative sample of that population to ensure validity. For impact evaluations, statistical inference is typically combined with descriptive statistics to reveal patterns in the data (e.g. outliers or missing data) and provide a basis for statistical inference analysis. Moreover, the same measures – like the mean – can be used in both descriptive statistics and statistical inference, but their purpose and interpretation differ. In descriptive statistics, the mean summarizes the sample data. In statistical inference, the sample mean is used to estimate or make conclusions about the population mean.

Another critical factor is understanding the nature of the collected data, particularly its scale of measurement (e.g. nominal, ordinal, interval or ratio). The scale of each variable collected defines what type of descriptive statistics and what sort of inferential statistics is applicable. For example, categorical data might require the visualization of proportions (descriptive statistics) or chi-square tests (statistical inference), while continuous data can be represented by distributions, measures of variability (descriptive statistics) or t-tests (statistical inference).

Finally, the target audience is another crucial factor when choosing the type of data analysis as it will significantly influence decisions regarding the choice of representation, such as tables or charts, the level of domain-specific language and how visually appealing an analysis needs to be. If the main purpose of an analysis is for the research team to better understand the data, the visual presentation and unambiguous interpretation will be less important than when the data are supposed to be communicated to a wider audience or used in publication.

This article is structured as follows: In section 1, we explore descriptive statistics, including fundamental summary statistics like mean and variance, which are essential for understanding data. The section also emphasizes the importance of data visualization, a powerful yet often underutilized tool in evaluation research. Moreover, it discusses guidelines and best practices for effectively visualizing relationships, time series, proportions and distributions. Section 2 focuses on statistical inference, covering key concepts such as hypothesis testing, statistical significance, p-values and regression analysis. As these methods are used in the majority of evaluation studies, a basic understanding of these inferential techniques is indispensable when it comes to interpreting findings and assessing evidence on the effectiveness of financial literacy interventions. Section 3 concludes.

Presenting statistical methods in a structured way and in the context of financial literacy research, this introductory article addresses evaluators, program designers and educators interested in descriptive statistics and statistical inference. It aims to refresh researchers' understanding of quantitative (inferential) analysis and advance their visualization skills. Moreover, it can be used by program designers or educators who wish to better understand the methods used in quantitative data analysis and derive meaningful insights from evaluations.

1 Descriptive statistics

1.1 Summary statistics

In quantitative data analysis, the concept of a *variable* is fundamental. A variable is an attribute that can vary over the unit of analysis. In the context of financial literacy interventions, the unit of analysis often covers *program participants* or *observations* across points in time. Variables can encompass all sorts of sociodemographic characteristics, typically age, gender, level of education or income. For instance, each participant (the unit of analysis) of an intervention has a value for their variable age. Sociodemographic variables are often used to give context or to provide results for certain subsamples.¹ Moreover, financial literacy evaluations often include variables that measure financial knowledge, behavior or attitudes. These are commonly referred to as *outcome variables* because they reflect the primary results of interest in an impact evaluation. In addition, for process evaluations or program monitoring, standard outcome variables might include the number of participants reached or participants' satisfaction with the program.

Summary statistics provide an overview of the variables and can be assigned to two different categories: statistics for *central tendency* or measures for *variability and spread* (Yoong, Mihaly, et al., 2013). Additionally, measures of association between variables – such as correlations – can also be considered summary statistics, as they capture relationships within the data. Table 1 shows the most important statistics for central tendency as well as variability and spread and their definitions.

¹ In regression analysis, as outlined in section 2, these types of variables are often called control variables.

Statistical measures

Central Tendency		Variability and Spread	
Measure	Definition	Measure	Definition
Mean	The sum of all data values for the variable divided by the number of values for the variable.	Range	The difference between the maximum and minimum data values of the variable.
Median	The middle value of the data values of the variable when the values are arranged in order.	Quartile	Divides the data values into four equal parts, each containing 25% of the values for the variable.
Mode	The data value that occurs most frequently for the variable.	Variance	Measures how spread out the values are by measuring how tightly the data are clustered around the mean.
Minimum	The smallest data value for the variable.	Standard Deviation	Is the square root of the variance and brings the measure back to the original units of the data, making it more interpretable.
Maximum	The largest data value for the variable.	Skewness	Measures asymmetry of the variable's distribution. positive values signal a longer right tail, negative values a longer left tail.

Source: Authors' compilation.

Another important differentiation is the scale level of a variable. Important distinctions are *continuous*, *discrete* and *categorical* variables.

Continuous variables can take on an infinite number of values within a given range, like height, weight, temperature or time. *Discrete* variables can take on a finite or countable number of values. They are often counts or categories with numerical labels, like the number of students in a class or the number of correctly answered financial knowledge questions. The difference between continuous and discrete variables is that the latter cannot be broken down meaningfully into smaller parts. For example, you can count 20 children in one class but not 20.5. Often, discrete variables with sufficient granularity are treated as continuous, e.g. age measured in full years.

Categorical data can be divided into distinct groups or categories, like educational levels (primary school, secondary school, bachelor's degree, master's degree, PhD), satisfaction ratings (unsatisfied, neutral, satisfied) or gender (male, female, diverse). Categorical variables such as educational levels and satisfaction ratings are so-called *ordinal* variables as they can be put in a meaningful order. This is not possible for *nominal* variables such as gender.

Understanding the scale of a variable is a prerequisite for choosing the right statistical methods and graphical representations for data analysis. Nominal data are usually reported as proportions (e.g. 50% of people observed in the dataset are male). Ordinal variables can be ranked and divided into intervals of equal probability; these are called quantiles. In this context, the median, quartiles, quintiles and percentiles are commonly used. For continuous and discrete variables, all central tendency and variability measures (table 1) as well as proportions and quantiles can be used to summarize the data.

While all these statistics can be reported in text form or in tables, in many cases it may be more valuable to visualize them in charts, which makes it considerably easier to grasp patterns in data quickly, especially for a nontechnical audience (Franconeri et al., 2021; Yoong, Mihaly, et al., 2013). Core data visualization techniques and tools are shown below.

1.2 Data visualization

Data visualization is the encoding of data through shapes, colors and animations, which must be chosen so that they can be easily decoded by the reader. Through decoding, you can examine variables from a different perspective and identify patterns that you might not have seen if they were presented in a table (Yau, 2011). Visualization can be applied at various levels of data aggregation, from plotting raw data points to presenting statistical information, including both summary statistics (e.g. means) and inferential statistics (e.g. confidence intervals). Because the core techniques for visualizing these different forms of data are largely similar, this subsection will focus on the visualization of raw data and summary statistics.

When approaching the task of visualizing a variable, the initial step is to consider the nature of the data – whether they are numeric or categorical, continuous or discrete.

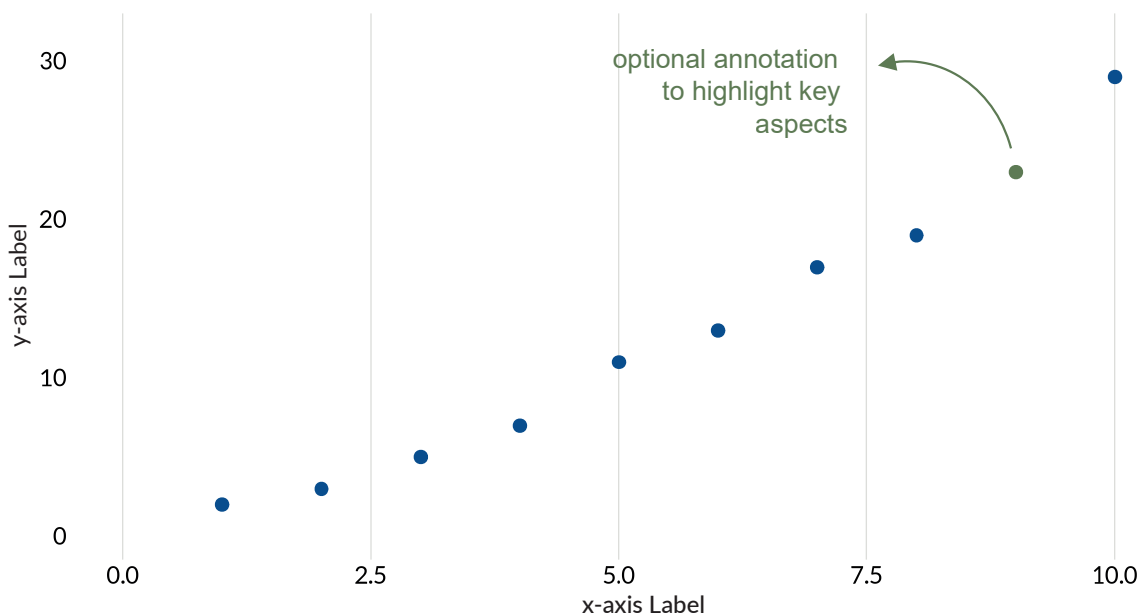
Next, reflecting on the goals you want your audience to achieve and the narrative you aim to convey, ask yourself questions like: Do you intend to illustrate temporal trends (*time series*)? Are you interested in conveying relative *proportions*? Do you seek to explore *relationships* between variables? Are you aiming to display the *distribution* of a variable or identify *outliers*? Depending on your answers to these questions, different types of charts and graphs are suitable for effective data representation.

Below, we will describe the core characteristics of some popular chart types and explain which of them can best be used to answer the above questions. For more details, see *Storytelling with data* (2023) or Yau (2011). Most charts, regardless of their type or purpose, typically contain some core elements that serve to convey information effectively. These elements include:

- (1) **Title:** A clear, concise title describes what a chart shows, providing context and helping viewers understand the main message.
- (2) **Axes:** Charts usually map data to a two-dimensional coordinate system, using a horizontal axis (x-axis) and a vertical axis (y-axis) and creating a structure for mapping data points. Without axes, it is not possible to understand the exact values or relative differences in the data. Some charts, such as pie charts, present values not along axes but according to circular segments based on angles or areas.
- (3) **Data points:** Data points are the visual representations of the data values of one or more variables and are placed along the axes according to their coordinates. The representation of data points can vary with the chart type.
- (4) **Labels:** Labels are used to identify specific data points or features of a chart. They can be data point labels, axis labels or annotations.
- (5) **Source and note:** Information about the data source and any important notes or explanations related to the chart should be provided.

Chart 1: Core elements of a chart

Title



Source: Simulated data.
 Note: Sample scatter plot.

These core elements serve to make a chart informative and accessible to the audience (see chart 1). Depending on the chart type and the complexity of the data, there may be additional elements like a legend, subtitle, error bars, trendlines or grid lines. The choice of chart elements should be based on the data being represented and the message researchers want to convey to their audience.

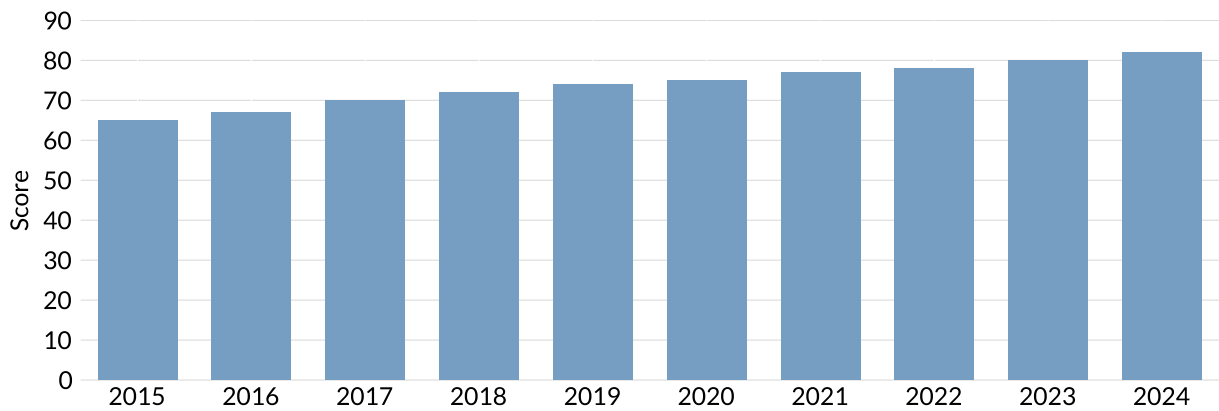
Below, we will show some typical chart types to represent 1) time series, 2) proportions, 3) relationships and 4) distributions. In the annex, we provide the R codes that reproduce the charts shown in this section.

1.2.1 Time series

A time series is a collection of data points or observations recorded at regular intervals over a specified period, with each data point being associated with a particular timestamp. These data points are ordered chronologically and are used to analyze and study how a variable or phenomenon changes and evolves over time. For instance, one might examine the evolution of financial literacy scores in a group of students over several years. Typically, the time dimension is plotted on the x-axis (time axis), while the y-axis represents the variable’s values. Time series are typically plotted using continuous or discrete variables. Common ways to represent time series data include bar plots and line charts. For a checklist on what to consider when designing a bar chart, see Rapp (2022). In chart 2 and chart 3, we show how a bar chart and a line chart can be used to represent a time series.

Chart 2: Bar plot to represent time series

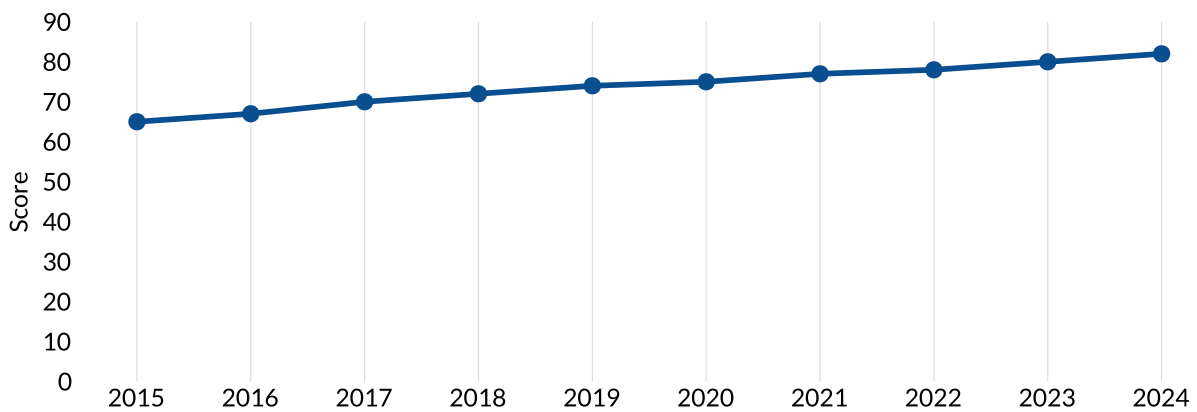
Financial literacy scores over the years



Source: Simulated data.
 Note: Financial literacy scores of students over the years.

Chart 3: Line chart to represent time series

Financial literacy scores over the years



Source: Simulated data.
 Note: Financial literacy scores of students over the years.

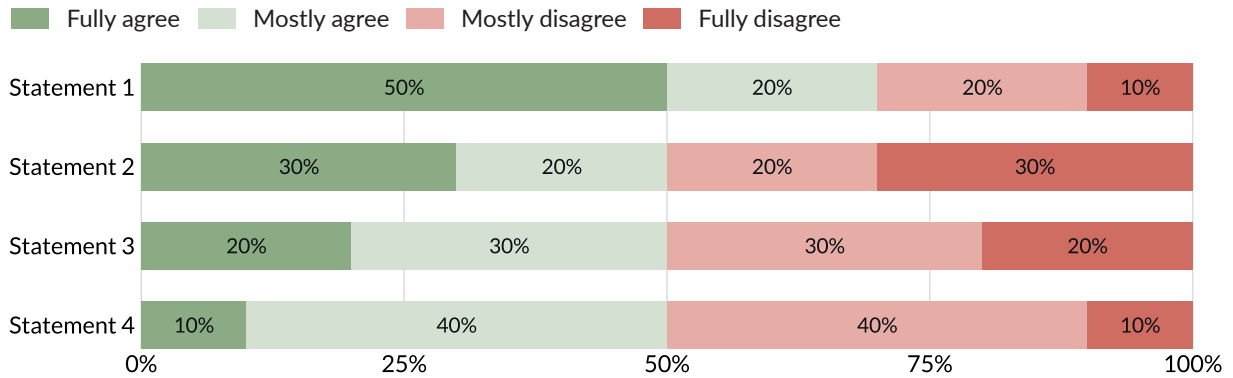
1.2.2 Proportions

Researchers frequently aim to illustrate the division of a group, entity or quantity into its individual components, with each part representing a portion of the whole to convey the relative significance of individual elements within the whole. For instance, you might analyze the distribution of a variable among men and women within a population or the percentage of respondents who agree or disagree on a particular topic. Common ways to represent proportions include pie charts, donut plots, (stacked) bar charts and (stacked) area charts. It is worth noting that bar charts and area charts are often preferred over pie charts, as they facilitate easier comparisons, especially with numerous categories or small differences in proportions (Muth, 2018a). Square area charts are an alternative to pie charts as they offer improved comparability. Area charts are useful for plotting the change of proportions over time. The defining feature of proportion-representing charts is their ability to showcase how each component contributes to the total, whether

through slices, areas or bars. Typically, categorical variables are used to represent proportions. In chart 4, we plot a stacked bar chart and in chart 5 an area chart to visualize proportions.

Chart 4: Stacked bar chart to represent proportions

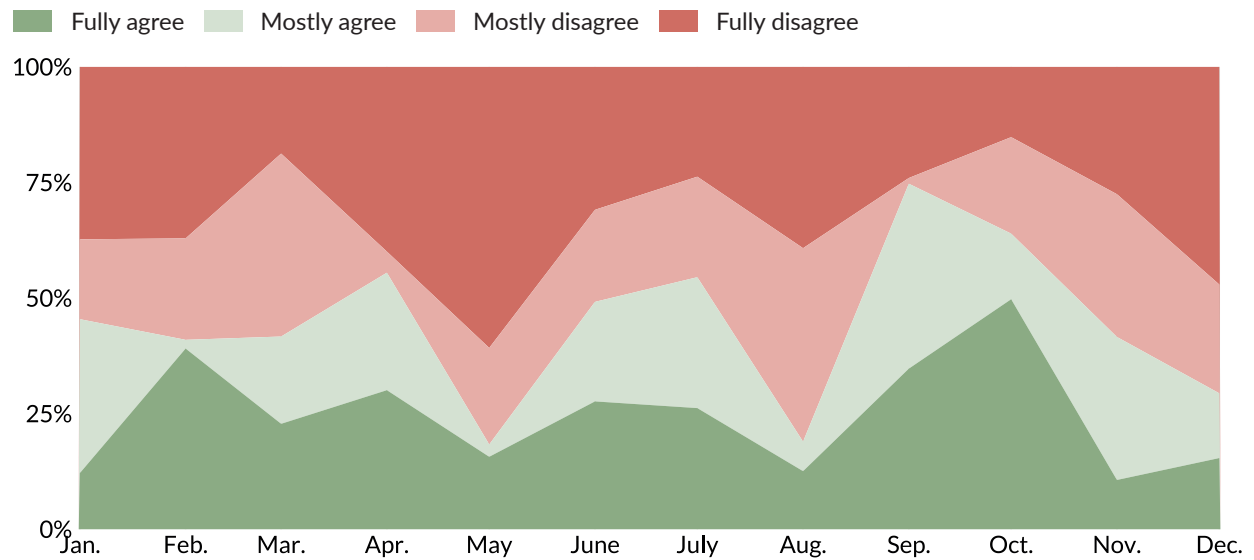
Agreement to statements



Source: Simulated data.

Chart 5: Area chart to represent proportions

Agreement to statement 4 | 2024



Source: Simulated data.

1.2.3 Relationships

When exploring how a variable’s values change with changes in another variable, we are interested in visualizing relationships or correlations. The go-to visualization for this purpose is a scatter plot (see e.g. chart 1). In a scatter plot, one variable is represented on the x-axis and the other variable is plotted on the y-axis. The choice of which variable goes on which axis depends on whether one variable influences the other. The dependent variable, which changes in response to the other (independent) variable, is typically

plotted on the y-axis, while the independent variable, which influences the dependent variable, is represented on the x-axis. While scatter plots share similarities with line charts in representing two discrete or continuous values, they are different in that they emphasize individual data points as dots. Line charts, on the other hand, typically connect multiple data points into a continuous line, as seen in time series data (see chart 3). Trend lines can be added to scatter plots to highlight the underlying relationship between the variables. However, caution is advised when using scatter plots for audiences who are not familiar with statistics, as scatter plots are exploratory in nature and can be misinterpreted as showing causal relationships.

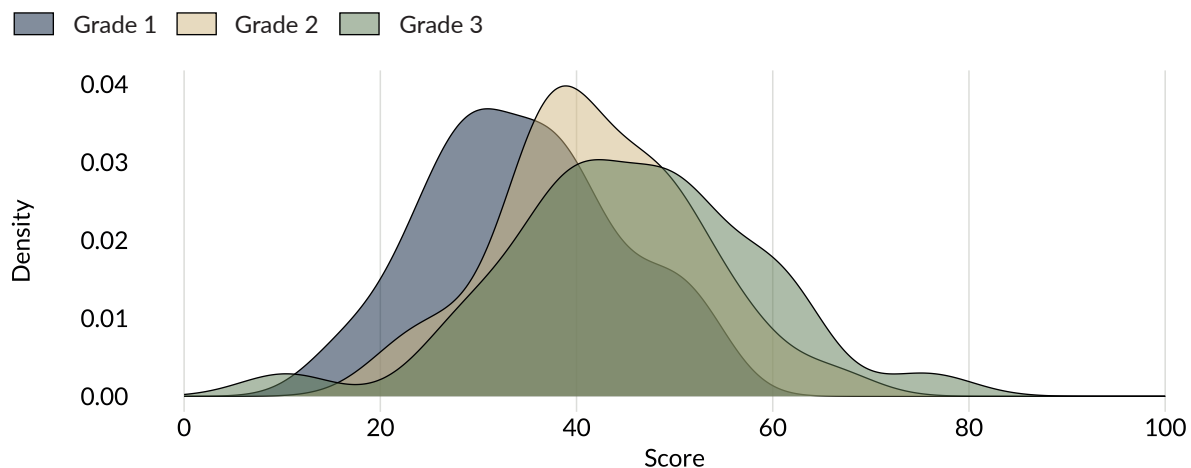
1.2.4 Distributions

When the goal is to comprehend the underlying structure of a variable, including the shape, central tendency and spread of its distribution, various visualization methods can be employed. Histograms, density plots and box plots are common choices. Histograms (not shown here) display the frequency or count of data points within different “bins” of a continuous or discrete variable. Density plots offer a continuous representation of the distribution and are suitable for continuous variables or fine-grained discrete variables. Box plots provide a summary of the distribution of continuous or discrete variables, displaying the median, quartiles and potential outliers, allowing for a quick overview of data distribution and skewness. To exemplify this, we show a density plot in chart 6 and a box plot in chart 7.

For more details on how to find the most appropriate chart for your data and for a variety of further types of visualization, see From Data to Viz (2018).

Chart 6: Density plots to represent distributions

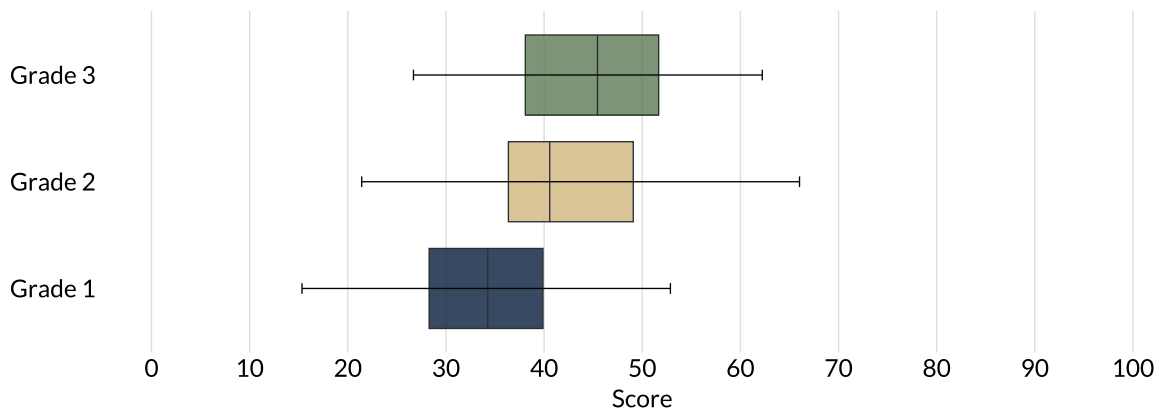
Financial literacy scores in three grades



Source: Simulated data.
 Note: Financial literacy scores of students.

Chart 7: Boxplots to represent distributions

Financial literacy scores in three grades



Source: Simulated data.

Note: Financial literacy scores of students.

1.2.5 Effective data visualization guidelines

Once the appropriate chart type has been selected, it is crucial to consider several fundamental visualization principles to ensure that the reader can easily understand the intended message. The visualization guidelines presented in this article are essential for creating effective and informative visual representations of data and support the accurate interpretation of charts. For further reading on visualization guidelines, we suggest Evergreen and Emery, (2016), Franconeri et al. (2021), Wilke (2019) and Yau (2011).

Visualizations can be misleading if chart elements, shapes and spacing are not chosen carefully. To mitigate illusion and prevent misinterpretation, axes should start at zero (Franconeri et al., 2021). Only in some cases can it be useful to start at values other than zero, especially if small differences are important for the interpretation of the data (Pearce, 2020). Also, axis intervals should maintain consistent spacing and have the same unit of measurement. For example, do not mix months and years in a time series presentation and ensure that all labels and annotations are readable, by potentially using horizontal annotation for easier reading also on the y-axis, and remove label redundancy. For instance, do not label every tick mark on your axis and do not use axis labeling and numeric labels together. Moreover, make the visualization as intuitive as possible by following a common understanding of directions (e.g., “up” and “right” indicating “more”). Differences in sizes of circles and squares are generally more difficult to capture than the area of bars. If you chose circles or squares, represent them based on their areas, not their radius, diameter or circumference as this creates distortions.

If used wisely, colors, annotations and highlighting provide a great way to improve storytelling and the readability of visualizations. You might wish to guide the audience in interpreting your visualization by providing visual hints through annotation and highlighting, especially when they may not be familiar with the data. For example, if one data point indicates a surprising result or conveys the main story of a chart, it can be useful to add a short explanation next to the relevant data point. Regarding the use of colors, you should take care to always choose colors according to their intuitive meaning, such as red for negative or lower values and green for positive or higher values. At the same time, it may be advisable to avoid an excessive use of colors as this can exaggerate minor differences and to use accessible color schemes, preferably combining shades of red and blue (for more details, see Muth, 2020). Moreover, sufficient contrast between colors prevents an overlap of intensities. Utilizing readable fonts and font sizes in your visualizations is also helpful (for further details, see Muth, 2022).

To facilitate the interpretation of your data, it is useful to reduce the cognitive load for the reader, which may involve replacing legends with direct labels and annotations where possible. If legends are necessary, care should be taken to order them logically (Wilke, 2019). Generally speaking, it is advisable to keep visualizations simple by eliminating noninformative elements like grid lines, background colors, excessive animation or decoration. Arranging data points, such as bars, in a meaningful order, e.g. by frequency or time, also facilitates readability and understanding. If a visualization contains too much information and becomes too complex, consider dividing it into multiple charts or more digestible segments to enhance comparison. For instance, use a split bar chart instead of a stacked bar chart if there are numerous categories to visualize. For more details on stacked bar charts, see Muth (2018b.)

Visualizations should be adapted to the target audience. The focus should be on what is important, meaningful and understandable for them. For example, complex statistical elements like error bars and confidence intervals should be communicated only when the intended audience will be able to interpret them correctly. If your audience is less experienced, you might want to provide guidance and rely on chart types they are familiar with.

For ethical data visualization, it is essential to consider several key aspects. First, be mindful of data privacy, particularly when working with sensitive or personal data. Avoid misleading or manipulative data representation and always aim for truthful and accurate representation. Ethical data visualization also includes the clear communication of all sources and processing of data. It is essential to recognize and transparently convey the limitations of data visualization. This involves ensuring that viewers are well informed about potential constraints or uncertainties in the data. This may include communicating issues such as missing values or a nonrepresentative sample. Additionally, it is crucial to remain mindful of the fact that relationships depicted through descriptive statistics are inherently correlational and that they should never be assumed to imply causation (Azzam et al., 2013). If data contain issues such as missing values or if they stem from unrepresentative samples, researchers should report these limitations clearly.

1.2.6 Tools for data visualization

Tools for data visualization can be categorized as out-of-the box tools or programming tools. Out-of-the box tools are user friendly, quick to get started with and suitable for common tasks but may lack customization and flexibility. Programming tools, while more complex to learn and use, offer more extensive customization and flexibility and make it possible to create tailored solutions for a wide range of data visualizations. The choice between the two types of tools depends on the specific needs and technical expertise of the researchers who will apply the tool (Yau, 2011).

Box 1

Data visualization tools

Out-of-the-box tools

- **Spreadsheet applications** such as Microsoft Excel or Google Sheets are perfect for users who want an easy-to-use solution with a standard set of graphics. They are particularly useful for small datasets, offering a straightforward way to gain rapid insights into your data.
- **Dedicated data visualization software** such as Tableau stands out for its interactivity and advanced visualization features. A potential drawback is data privacy. In the free version of Tableau, for example, data become publicly accessible once they are uploaded.

Programming software

- **Web-based tools** like HTML, JavaScript and CSS are excellent for creating interactive visualizations. They provide high levels of customization. Nevertheless, they require more prior knowledge compared to other programming tools.
- **Python** is highly effective for handling extensive volumes of data and features a straightforward syntax. While default charts may be unappealing, various libraries are available for creating attractive and customizable visualizations.
- **R** has a rich ecosystem of statistical tools and libraries, making it suitable to perform complex statistical analysis alongside data visualization (particularly in the ggplot universe). While it offers a relatively straightforward syntax, customization in R can be more time consuming than using web-based tools.
- **Interactive data visualizations** are possible in Python and R by integrating HTML, JavaScript and CSS libraries (e.g. R Shiny, highcharter, plotly) (Baruffa, 2023; Sievert, 2020) although dedicated interactive visualization programs can be superior in terms of performance and maintenance (e.g. updates).

For further reading and resources on data visualization, online references can be highly recommended, as there is an active and collaborative online open-source community dealing with the area of data visualization. This community comprises social media groups, blogs and websites that are often dedicated to specific data visualization tools and software. For R, see e.g. blogs (Rapp, 2025; Scherer, 2023), books (Baruffa, 2023; Nordmann et al. 2021) or tutorials (Scherer, 2019). Moreover, the community collaborates actively, contributing to open-source data visualization projects and sharing best practices. For instance, initiatives like the Tidy Tuesday challenges encourage participants to create visualizations and share their R codes with the broader community. Similar challenges exist for various other data visualization tools.

2 Statistical inference

2.1 Quantifying uncertainty

As outlined in the introduction of this paper, quantitative data analysis aims to present and summarize data in a way that allows for identifying patterns. Section 2 presents the fundamental principles behind such summarization, exploring variables, data types and their respective descriptive statistics. Descriptive statistics can also be regarded as the first step of statistical *inference*.

Statistical inference builds on the idea that the calculated statistics stem from a *random sample*, i.e. a randomly selected subset of the population. From the sample, researchers draw conclusions on the underlying population, e.g. on the target population of a financial education intervention.

In other words, statistical inference is the process of deriving insights about a population based on a sample from that same population. Importantly, statistical inference would be meaningless if we had reliable data on the whole (target) population; in such a case, purely descriptive statistics would be sufficient.

Researchers thus use sample data to calculate *sample statistics* that serve as estimates of the unknown true characteristic of a population, the *population parameters*. Population parameters and the respective sample statistics of interest may be any quantitative summary statistics (including the descriptive statistics presented in section 1), e.g. the average (mean) financial knowledge score, the variance of that score, the correlation between that score and financial well-being or the effect of a financial education intervention on the amount of monthly savings.

Methods of statistical inference allow us to *quantify uncertainty* from (small) samples. Suppose, for instance, that we want to test the effect of a financial education intervention on the financial knowledge of adults. We could sample only five persons from the (hypothetically very large) population of potential

participants to constitute the treatment group, and five nonparticipants as the control group. However, if we compared their average financial knowledge test scores, we would have little confidence that the difference is an accurate reflection of the true difference we should expect from the intervention.

Indeed, we must expect that financial knowledge varies considerably within the population and that choosing only five people at random could result in averages far above or below the groups' actual average financial knowledge. This randomness could influence our estimated averages much more than the intervention. We could conclude that we need larger samples, and rightfully so: Increasing the sample size reduces random variability in average scores, bringing the estimated averages closer to the true average level of financial knowledge in the population.

In essence, statistical inference is used to eliminate guesswork from judging a statistic as “probably correct” based on the underlying variability in the population and the sample size. Phrased more technically, statistical inference formalizes and quantifies the uncertainty with which we draw conclusions from sample data. It is used to derive the probability that a result – or a more extreme result, e.g. a difference of 10 or greater – would occur merely due to random sampling variability even if there is no actual difference in the population. This probability is the *p*-value. The *p*-value is often interpreted as the “significance” of a result in terms of the strength of evidence. In classical hypothesis testing, however, the *p*-value is compared against a pre-set criterion of just-acceptable uncertainty (often 5%).

In the context of financial literacy, statistical inference is typically applied to examine *differences* between subsamples (e.g. between men and women) and *relationships* between variables. In effect, it is used to evaluate the *effectiveness* of interventions (e.g. the effect of financial education on financial literacy) and identify other key factors associated with an outcome. In the context of quantitative evaluation research, statistical inference is almost universally accepted as the method to assess the impact of a program in terms of statistical significance.

However, statistical significance alone is not enough when it comes to evaluating an effect as practically relevant. A difference, for example, must also be interpreted in absolute terms (e.g. whether an average 50 cent increase in savings is relevant) or relative to the overall variability as a standardized effect size (e.g. a difference equivalent to 0.20 standard deviations).

The following subsections explore in more detail the basic principles behind inferential statistics. First, we describe the importance of clearly defining the population and the relevance of the random sample. We then present a short introduction to classical hypothesis testing and explore the basic principles behind common statistics and significance tests as they are commonly used in practice, building on simple calculations of the mean and the standard deviation. We conclude with a short overview of the many possible caveats of inferential statistics and the unavoidable omissions of this paper.

While we follow what is typically taught in introductory courses on statistical inference, we aim to provide an intuitive but nevertheless concrete understanding of the most important mathematical principles of statistical inference. More detail and many more statistical methods and their application using statistical software can be found in introductory textbooks such as Woodridge (2009), Anderson et al. (2020), Field (2013) or Janczyk and Pfister (2023).

2.2 Generalizing from a random sample to a population

A clear definition of the (*target*) *population* and a *random sample* are the basis of statistical inference. Indeed, depending on the definition of the population, the quality of the sample and the generalizability and usefulness of results may change considerably. Ultimately, however, it is up to the researchers to define the target population. For instance, 100 participants in a specific financial education program could be understood as 1) a sample of the general population, such as the adult residents of a country, 2) a sample of a population of potential program participants (e.g. due to their motivation and interests) or 3) the target population itself.

Additionally, common methods of statistical inference assume a *random* sample of a larger population. In a random sample (technically a *simple* random sample), every member of the target population has the same chance of being selected into the sample. This ensures that the selection of population members (e.g. survey participants) does not introduce any systematic patterns into the sample's properties that would lead to *bias*, i.e. systematic deviations of the estimated statistic from the true population parameter. Random sampling eliminates bias on average because differences in the distribution of characteristics between sample and population are canceled out with repeated sampling. With sufficient size, a random sample can thus be considered representative of the population.

In practice, large random (or equivalent probability) samples are difficult to achieve. Quantitative researchers are thus often concerned with *reducing bias* through (in some cases quite complex) sampling and weighting procedures.² Conceptually, however, random samples are the foundation of statistical inference and of estimating population parameters.

Any sample can be considered one of hypothetically infinite repetitions of the same procedure. Because a random sample is free of systematic bias (and if there are no other sources of bias), the estimated statistics vary around the true population parameter. The variability of these statistics is called the random sampling variance, and it determines the expected random sampling error, i.e. the deviation between a statistic and the population parameter that is to be expected. The assumptions and methods presented here are foundational to conventional frequentist statistics, which rely on the behavior of estimators across (hypothetical) repeated samples. Other approaches exist, such as Bayesian statistics, which have a fundamentally different understanding of probabilities, inference and "truth."

Luckily, it is possible to calculate both the statistic itself and its random sampling error (e.g. a mean and its standard error) from a single sample. This works because the variability of observations within a sample gives insight into how much the statistic would fluctuate across multiple (hypothetical) samples. From this, we can derive probabilities that indicate how likely the observed statistic (or a more extreme one) would be, assuming a certain population parameter. We can quantify, for example, how likely it is to find the observed difference between two groups if we assume that the true difference in the population is zero. This is the p-value and it is the basis for statistical hypothesis tests.

Box 2

Key terms of statistical inference

- **Statistical inference:** the method of drawing conclusions about a population from sample data
- **(Target) population:** the specific group a study aims to draw conclusions about, as defined by the researcher
- **Sample:** a smaller group selected from the population for the purpose of analysis
- **(Simple) random sample:** a sample where every member of the population has the same chance of being chosen; random samples are the basis of statistical inference
- **Population parameter:** a true value that describes a characteristic of a population, like the actual average financial knowledge of all individuals of the target population
- **Sample statistic:** a value calculated from the sample that estimates a population parameter
- **Generalizability:** the extent of how well the results from a sample apply to the larger population
- **Bias:** a systematic deviation of a sample statistic from the true population parameter, often due to flaws in sampling
- **Sampling variance:** the variability in a sample statistic that occurs because different random samples would yield different results

² Sampling methods are discussed in a different issue of the OeNB Financial Literacy Evaluation Series.

- **Sampling error:** the difference between a sample statistic and the true population parameter, caused by random chance in the sampling process
- **Standard error:** a measure of the expected variability of a sample statistic due to random sampling

2.3 Hypothesis testing

Hypotheses are an important part of inferential statistics. A hypothesis is a statement about a population that is tested using a sample. *Hypothesis testing* or *significance testing* is generally used as a process to assess whether differences (e.g. mean differences in financial literacy scores) or associations (e.g. between financial knowledge and financial well-being) are clear enough to be deemed statistically significant, i.e. not a product of a random sampling error. Performing hypothesis testing involves four key steps:

1. defining the hypotheses to be tested;
2. determining the criteria for deciding whether the claim being tested is true or not;
3. computing the sample statistic from the data sample and deriving its corresponding test statistic and probability value (p-value);
4. deciding to *reject* the claim as being true if there is a large discrepancy between the sample data and what we would expect to observe if the claim was true.

For step 1, we need to formulate two statements that form the two complementary and opposing hypotheses known as the *null hypothesis* (H_0) and the *alternative hypothesis* (H_1). The way these hypotheses are structured depends on whether the hypothesis test is two-tailed, examining effects in both directions (e.g. whether a mean is greater or smaller, i.e. more extreme), or one-tailed, focusing on a single direction. In most cases, two-tailed tests are chosen unless there is a clear justification for limiting the test to one direction. To simplify matters, we only provide examples of two-tailed tests in the following subsections.

The null hypothesis H_0 states that a given phenomenon does not exist in the population, e.g. that there is no association between variables or difference between groups. The alternative hypothesis H_1 , in contrast, claims the opposite. For example, a pair of hypotheses could be:

H_0 : *There is no correlation between financial knowledge and financial well-being.*

H_1 : *There is a correlation between financial knowledge and financial well-being.*

In step 2, we determine when to reject H_0 . To this end, a *level of significance* must be selected. This value, denoted by α , specifies the threshold for how small the p-value must be to justify rejecting H_0 . Common choices for α are 5% or 10%, though other values can also be applied depending on the context.

For step 3, we collect data and compute the sample statistic (e.g. the correlation coefficient) from the data sample, as well as its corresponding *test statistic*. For this test statistic, a *probability value* (p-value) is calculated. This value represents the likelihood of observing the test statistic under the assumption that H_0 is true. Specifically, a low p-value indicates that the observed result is unlikely to occur if H_0 is correct. Typically, these tests are conducted using statistical software or programming such as Stata, SPSS, R or Python. The choice of the test statistic depends on the specific context, though the general testing procedure remains the same.

In step 4, we decide whether to reject or fail to reject H_0 . Specifically, if the p-value $\leq \alpha$, we reject H_0 , meaning that the difference or association is statistically significant. Otherwise, we fail to reject H_0 .

As shown in table 2, there are four possible scenarios depending on the real effect in the population and the corresponding correct or false decision. As mentioned, the α level is selected by the researcher. Under ideal conditions, an α of 5% will lead to significant results and thus a false rejection of H_0 in exactly 5% out of every 100 (hypothetical and infinitely many) samples. Such a *false positive* result is called *type I error*. Its probability is set directly by the researchers themselves as the significance level α . Its complementary

is the confidence level $1 - \alpha$, which is the probability that H_0 is *not* rejected. If the null hypothesis is false (e.g. there is a true difference in the population) but it is not rejected, this is a false negative, whose probability is denoted by β . The complementary probability $1 - \beta$ is called *power*. It indicates how likely it is to find a significant result if there is a real phenomenon in the population. Power increases with larger effect sizes (e.g. the difference) and larger sample sizes. The power is greater the larger the effect (e.g. the difference) in the population or the larger the sample size. With it, researchers can derive the minimum required sample size to find an effect of prespecified size with the desired probability $1 - \beta$.

Table 2

Possible decisions and their (error) probabilities

	H_0 not rejected	H_0 rejected (significant result)
H_0 is true	Correct with probability $1 - \alpha$	Type I error: False positive with probability α
H_0 is false	Type II error: False negative with probability β	Correct with probability $1 - \beta$ (power)

Source: Authors' compilation.

Box 3

Key terms of hypothesis testing

- **Hypothesis:** a statement about a population to be tested; hypotheses are typically formulated as a null hypothesis (H_0) that assumes that no effect or difference exists in the population and an alternative hypothesis (H_1) that assumes that an effect or difference does exist
- **Hypothesis testing:** a method used to determine whether observed differences or associations in sample data are statistically significant or likely due to random chance
- **Statistical significance:** a result is statistically significant if the p-value is less than or equal to α , indicating the finding is unlikely to have occurred by chance
- **p-value:** the probability of observing the test statistic (or something more extreme) if the null hypothesis is true; a low p-value suggests the observed result is unlikely under H_0 .
- **Level of significance (α):** a threshold set by the researcher (commonly 0.05, i.e. 5%) that defines how small the p-value must be to reject the null hypothesis; it represents the risk of a type I error
- **Errors (type I and type II):** a false positive occurs when H_0 is wrongly rejected (type I error), with a probability equal to α ; a false negative happens when H_0 is wrongly not rejected (type II error), with probability β
- **Power ($1 - \beta$):** the probability of correctly rejecting the null hypothesis when it is false; higher power means greater ability to detect true effects, and this ability is influenced by effect size and sample size

2.4 Statistics and tests to estimate population parameters

This section illustrates in simple mathematical terms the principles behind estimating population parameters, using the example of the mean and standard deviation and the corresponding one-sample t-test, the difference between means and the independent two-sample t-test, the correlation coefficient and a simple regression.

The sample statistic (generally denoted with a Latin letter) is the estimate for the population parameter (generally denoted with a Greek letter). For instance, the mean \bar{x} (x-bar), e.g. the mean financial knowledge of a sample, is the estimate for the mean financial knowledge of the population μ (mu).

While the formulas presented in the following subsections may appear daunting at first glance, the principles behind these statistical tests follow a simple logic: For each statistic, such as an average (mean)

financial knowledge score, the *variability* of the underlying scores is calculated. This variability is generally based on the sum of all squared distances of values from the mean itself, the *sum of squares*.

Together with the sample size, the sum of squares is used to calculate variability metrics such as the variance and standard deviation of the values. Moreover, the sum of squares is used to estimate the variability of the statistic itself. With this information, the resulting statistic can be standardized and compared to a known distribution. From this comparison, we can derive the probability of a result that is equal or more extreme than the statistic, i.e. we can determine a *p-value* and whether a result is *significant*.

The presented statistics are a small part of statistical inference. There are many more different statistics to describe variables and their distributions and associations, such as the median, frequency proportions, skewness and rank correlations and countless increasingly complex methodological frameworks, all of which can be used for statistical inference. The presented statistics and test, however basic, are still prevalent in practice and are well-suited to illustrate the principles behind statistical inference.

2.4.1 The mean

The mean is likely the most important and most reported summary statistic. The mean and its properties also build the basis for the most common tests of statistical inference, such as the independent t-test.

The (arithmetic) mean of a sample (denoted as \bar{x}) is the sum of each observation (x_i) of a variable divided by the number of observations (n):

$$\bar{x} = \frac{1}{n} \sum x_i$$

For example, if a sample of five people score 5, 6, 6, 10 and 11 in a financial knowledge test, the mean is $38/5 = 7.6$.

The sample mean \bar{x} is an *estimate* of the population mean μ , i.e. it can be used to draw conclusions about the population. The sample mean is generally an *unbiased* estimator, which means that with increasing sizes of random samples, the sample mean will get closer to the population mean.

The mean has an important property: It is the value where the sum of the *squared* distances to each sample value, the *sum of squares*, is as small as possible. Using the scores from above, the mean $\bar{x} = 7.6$ minimizes the following sum of squares:

$$(5 - \bar{x})^2 + (6 - \bar{x})^2 + (6 - \bar{x})^2 + (10 - \bar{x})^2 + (11 - \bar{x})^2$$

The sum of squares can be written as:

$$SS = \sum (x_i - \bar{x})^2$$

The mean is thus the *least squares estimator* of central tendency. As the following sections will show, calculating or minimizing the sum of squares is the basis of many statistics and statistical tests. Importantly, the sum of squares is also used to calculate the sample variance and standard deviation.

Because squared differences are used, the sum of squares gives larger differences more weight: The mean gets “pulled” more strongly toward more distant values. The mean and all other estimates and significance tests based on the sum of squares is therefore sensitive to outliers (i.e. a few very large or small values) and can severely misrepresent the center of nonsymmetrical distributions. The mean stands in contrast to the *median*, which minimizes the absolute (and not the squared) distances. The median is largely robust to outliers and can often be a better measure of central tendency when a distribution is not symmetrical.

2.4.2 The standard deviation and variance

The mean alone gives us only a measure of the central tendency but no indication about the spread or variability of the underlying values. For an approximate description of the distribution, we thus need a measure that quantifies the spread of values around the mean – the standard deviation. Its calculation is closely related to the mean itself because it is based on the sum of squares.

The standard deviation s (or SD) is the square root of variance s^2 . Variance s^2 is the sum of squares divided by the sample size (although 1 is subtracted from the sample size to obtain unbiased estimates). The mean is therefore also the value that minimizes the variance. The variance is calculated as:

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

Using the same example as above, the variance of the five scores is $s^2 = 7.3$, the standard deviation is:

$$s = \sqrt{7.3} \approx 2.70$$

2.4.3 Standardization and the p-value of a single observation

Together, the mean and standard deviation can fully describe the most important theoretical distribution of statistical inference, the *normal distribution*. If we can assume that we are indeed working with a normal distribution of our values, we can use the mean and standard deviation to transform any value of a variable (e.g. any of the five scores from above) so that they can be evaluated and compared with a theoretical normal distribution or with observations from other samples. This is called *standardization* or *z-transformation*:

$$z = \frac{x - \bar{x}}{s}$$

For instance, for one of our five knowledge scores $x = 10$, the standardized z -value is approximately 0.89, which means the value is approximately 0.89 standard deviations above the mean. If all scores are standardized, the resulting distribution of standardized values has (by definition) a mean of 0 and a standard deviation of 1.

The theoretical normal distribution with a mean of 0 and a standard deviation of 1 is called standard normal distribution. Because its properties are known, we also know the probabilities to obtain certain ranges of values. The z -value of 0.89 can also be transformed into a *probability* that indicates how likely it is to get a value equal to or more extreme than 0.89 from the standard normal distribution. For example, the probability of sampling a single score that is more extreme than $z = 0.89$ from a standard normal distribution is $p = 0.37$.

Subsection 2.4.4 illustrates why the (standard) normal distribution can be used as the benchmark distribution for statistical tests and why our small hypothetical sample of five scores may not actually be suited for the tests presented here.

2.4.4 Why do statistics follow normal distributions?

Statistical tests rely on the fact that statistics from (large) samples often follow a normal distribution. This is because, to determine probabilities, statistical tests compare a standardized sample statistic (e.g. a mean) to the theoretical standard normal distribution. This section illustrates why statistics have this important attribute and how sample size affects the distribution of sample statistics.

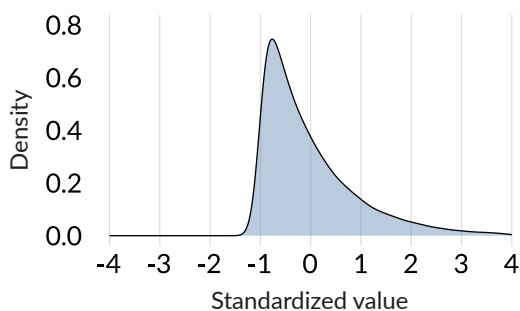
Measurements of real-world phenomena such as physical or psychological attributes (e.g. people's height or financial knowledge test scores) tend to cluster around their mean value, with fewer and fewer cases occurring with increasing differences from the mean. The measurements often tend to follow a bell-shaped distribution that is called "normal distribution."

The fact that many real-world measurements follow normal distributions is due to the central limit theorem (CLT). The CLT states that a summary score or a statistic (e.g. a sum or a mean) of many independently sampled values approach a normal distribution, regardless of the distribution of the underlying values. For example, even the sum of multiple coin toss results (where heads and tails are assigned any different values, such as 7 and 39) will approach a normal distribution if the samples are large enough. This means if we sum up the outcomes of 1,000 coin tosses and repeat this process to obtain many sums, the sums will be close to normally distributed.

Chart 8 shows the standardized distributions of means from samples of size $n=5$ or $n=50$ drawn from an asymmetric standardized distribution. As is apparent, the means from the larger samples are close to normally distributed. The means from the smaller samples, however, show considerable skewness.

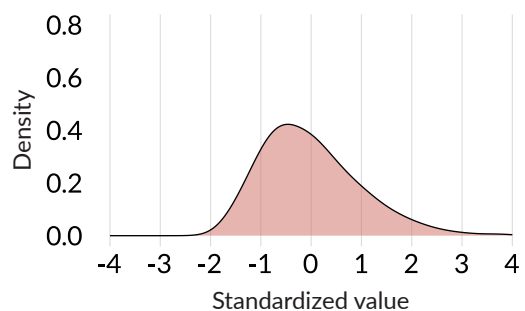
Chart 8: Illustration of the central limit theorem

Original skewed distribution



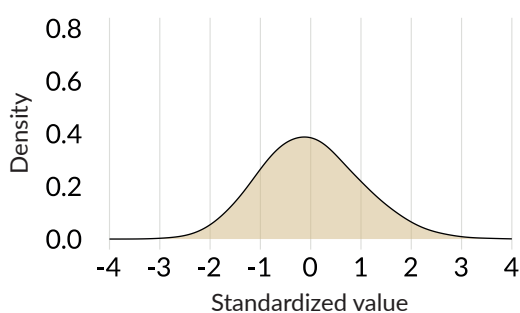
Source: Simulated data.

Distribution of means from $n=5$



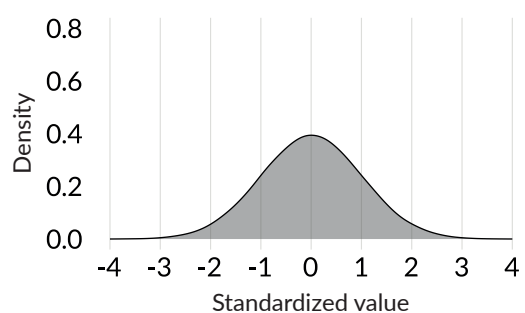
Source: Simulated data.

Distribution of means from $n=50$



Source: Simulated data.

Standard normal distribution



Source: Simulated data.

Importantly, the CLT is applicable on two levels, namely on the level of variables (such as a knowledge score) and on the level of individuals in the sample.

On the level of variables, real-world phenomena such as financial knowledge can be regarded as “summary” variables, i.e. as “statistics” on the level of an individual. Financial knowledge, for example, could be considered the result of many different mental processes. These processes may be independent enough that when we measure the financial knowledge of only one person, we effectively obtain a sample of many (unknown) *independent variables* summarized into one score. If we repeat this measurement process

with many people, the scores may thus follow a normal distribution. We could therefore assume (or have empirical evidence) that the knowledge score used in our examples is normally distributed in the population.

On the level of individuals, the CLT will also come into effect when we sample multiple *independent individuals* from the population. Even if the measured variable is not normally distributed (i.e. the conditions described above are not fulfilled), obtaining a large sample of independent individuals and thus knowledge scores will result in a summary statistic (such as the mean financial knowledge) that follows the normal distribution.

Statistical tests rely on this property. They compare a standardized sample statistics (e.g. a mean) to the theoretical standard normal distribution to determine probabilities, confidence intervals and significance levels. Statistics from large samples can thus be standardized and compared against a standardized normal distribution. The necessary sample size depends on the underlying distribution of the values: If the underlying distribution is very nonnormal, very large samples may be necessary.

The conditions for a normal distribution of summary statistics can be provided by the researcher by obtaining a large random sample. However, the CLT does *not* guarantee that values within a large sample follow a normal distribution (see Zhang et al., 2023, for an overview and typical misconceptions about the CLT).

2.4.5 The probability of a mean

In the previous section, we derived the probability to obtain a single value (or a more extreme value) from our distribution. Applying the same principle to the sample statistic, i.e. the estimate of the mean, finally allows us to draw conclusions about the population, namely about the probability of our mean estimate assuming some population mean. This is called the *one-sample t-test* and it is one of the most basic significance tests of statistical inference and the basis for confidence intervals.

To gauge the variability of the mean estimate across multiple (hypothetical) repetitions of our sample, the *standard error* of the mean is used. The standard error is the standard deviation of the mean itself and is derived from the standard deviation of the underlying values:

$$SE = \frac{s}{\sqrt{n}}$$

The standard error thus reflects the mean's variability across many (hypothetical) samples. It therefore relates to the mean *statistic* itself and not to its underlying values. For our five knowledge scores, the resulting standard error is approximately 1.21.

The mean can be standardized against its standard error, just like the single score can be standardized against the standard deviation in the example above. However, for the single score value of 10, there is a known mean to compare against. For the estimate of the mean itself, the population mean is not known. Consequently, we need to specify a population mean to compare against. The standardized value is then calculated as:

$$t = \frac{\bar{x} - \mu}{SE}$$

Because the standard error is itself only an estimate, the resulting value is not compared against a standard normal (z-)distribution, but against a *t-distribution*. t-distributions are based on standard normal distributions but compensate for the added uncertainty from estimating the standard error from the sample. There are different t-distributions for different sample sizes (technically expressed as *degrees of freedom*) for which probabilities can be looked up or calculated.

Specifying μ allows for obtaining the probability of how likely the sample mean $\bar{x} = 7.6$ is for a given mean of the population (μ). In other words, we can estimate the probability of the difference between the estimated mean (\bar{x}) and a specified population mean (μ). For example, we could test the null hypothesis that assumes the population mean $\mu = 0$. In this case, the t-statistic is approximately 6.29 for our estimated mean $\bar{x} = 7.6$, which corresponds to a probability of approximately 0.003 to obtain a sample mean of 7.6 (or more extreme) with $\mu = 0$.

0.003 is the p-value of this specific test. At the conventional significance level of $\alpha = 0.05$, the mean is indeed significantly larger than zero. It is thus very unlikely that the estimated mean stems from a population where the mean is zero, and we can reject the null hypothesis.

We could also compare the sample mean to a different population mean, e.g. a minimum mean score that characterizes a population as “financially literate.” Instead of testing against zero, we would formulate the null hypothesis that the population mean is not different from this specific value.

2.4.6 The confidence interval of a mean

Instead of testing our sample mean against a specific value of μ , we may be interested in a range of values of μ that would *not* be rejected as null hypotheses and thus reflects our level of confidence for our mean estimate. By a similar logic to that used in the one-sample t-test and by rearranging the previous formula for t , we can thus calculate the upper and lower limits of a *confidence interval* for the mean as:

$$CI = \bar{x} \pm t^* \cdot SE$$

Here, t^* indicates the *critical value* from the t-distribution based on our confidence level $1 - \alpha$ (e.g. 95%). The calculation gives us the range of μ for which the null hypothesis would *not* be rejected by the one-sample t-test. 95% confidence intervals thus represent the intervals that capture the true population mean in 95% of cases if we repeated the sampling process from the population many times. If the confidence interval does not include zero, the mean is significantly different from zero at $\alpha = 0.05$. For $\bar{x} = 7.6$ the 95% confidence interval ranges from 4.25 to 10.95.

2.4.7 Comparing two group means: the independent t-test

In practice, researchers may be more interested in testing the difference of the means of two samples instead of comparing a sample mean to a hypothetical population mean. The *independent two-sample t-test* (often just called *independent t-test*) can do just that. For instance, it can be used to test the effect of a financial education measure on the financial knowledge test score. One group could be the *treatment group* that received a financial education intervention, while the other is the *control group* that did not, with means reflecting each of the groups' knowledge test scores.

The independent t-test is thus an extension of the one-sample t-test. The t-statistic weights the difference between the two estimated means against its standard error, using both subsamples' variance and sample sizes:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Like with the one-sample t-test, the p-value can be determined for the now standardized difference. We can thus test the null hypothesis that there is no difference between the two means.

2.4.8 The covariance and correlation between two variables

Researchers are often interested in the association between two different variables (x and y , e.g. the financial knowledge score and the financial well-being score. For example, if a person scored highly on the financial knowledge test, is this knowledge also associated with higher financial well-being? To measure this association, the covariance can be used:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Unsurprisingly, the covariance is similar to the variance. Instead of the sum of squares, however, it uses the sum of each *product* of the value pairs' distances from the mean. If a person scores above average both in knowledge (x_i) and in financial well-being (y_i), the product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive. If a person scores below average in both, the product will also be positive. If a person scores above average in one and below in the other, the product will be negative. Results close to the average in one or both variables result in products close to zero. The mean of these products across all observations indicates the direction and size of the overall association and is represented by the formula above.

Although the direction of the association is clear from the covariance, the strength of the association is difficult to interpret. The result can thus be standardized as:

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

To get the correlation r , the covariance is weighted by the product of the two variables' standard deviations, so that it ranges from -1 (perfect negative association) to $+1$ (perfect positive association). Alternatively, we could also calculate the covariance of two standardized variables to get r .

The correlation coefficient r can be further transformed into its t-distributed test statistic to test for statistical significance, i.e. to obtain the probability of observing r or a larger (smaller) value if the association between the variables in the population is zero.

An important extension of the correlation is the *partial correlation*. For instance, researchers may suspect that the association between financial knowledge and financial well-being is not quite independent of other variables, e.g. income. They may thus want to calculate the correlation between financial knowledge and financial well-being net of their common influence from income (i.e. *controlling* for income). This can be achieved using the partial correlation. Calculating the partial correlation $r_{xy \cdot z}$ between x (e.g. financial knowledge) and y (e.g. financial well-being) net of the influence of z (e.g. income) requires the correlation coefficients between all variable pairs, where r_{xy} represents the correlation between x and y , r_{xz} represents the correlation between x and z and r_{yz} represents that between y and z :

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

2.4.9 Tests of correlation and difference in means are equivalent

Correlations and differences in means are conceptually and mathematically equivalent. The difference between two group means as tested with an independent t-test can also be thought of as the association between a binary (dichotomous) variable and a continuous variable.

If we assign a numeric value of 0 to the control group and a value of 1 to the treatment group – this process is called *dummy coding* – we can calculate the correlation between the numeric group variable and the knowledge test scores. This correlation has the exact same p-value as the independent t-test and it can even be interpreted as a measure of effect size. This type of correlation is called point-biserial correlation. The same principle also applies in regression analyses, as outlined in subsection 2.4.10, where both group variables (coded as 0 and 1) and continuous variables are used.

2.4.10 Regressions

Regression analyses (or simply *regressions*) can be considered the combination and generalization of all previous methods into a more comprehensive framework. This allows for comparing means, estimating associations between continuous variables and controlling for the effects of the other variables at the same time. In practice, impact evaluations often use regressions and their numerous extensions as the main statistical framework for statistical inference outcomes (Gertler et al., 2016).

In a regression, one *dependent variable* (also called *outcome variable*) and one or more *independent variables* (also called *predictor* or *control variables*) are analyzed simultaneously. As opposed to correlations and difference tests, a regression must be considered at least “directional” (i.e. one variable influences the other) in principle because the outcome is predicted or explained. In other words, a regression is a *model* of how the outcome came to be. Nevertheless, a regression model cannot per se give evidence on the causality of associations. The data and study design must be suitable for causal analysis and interpretation.

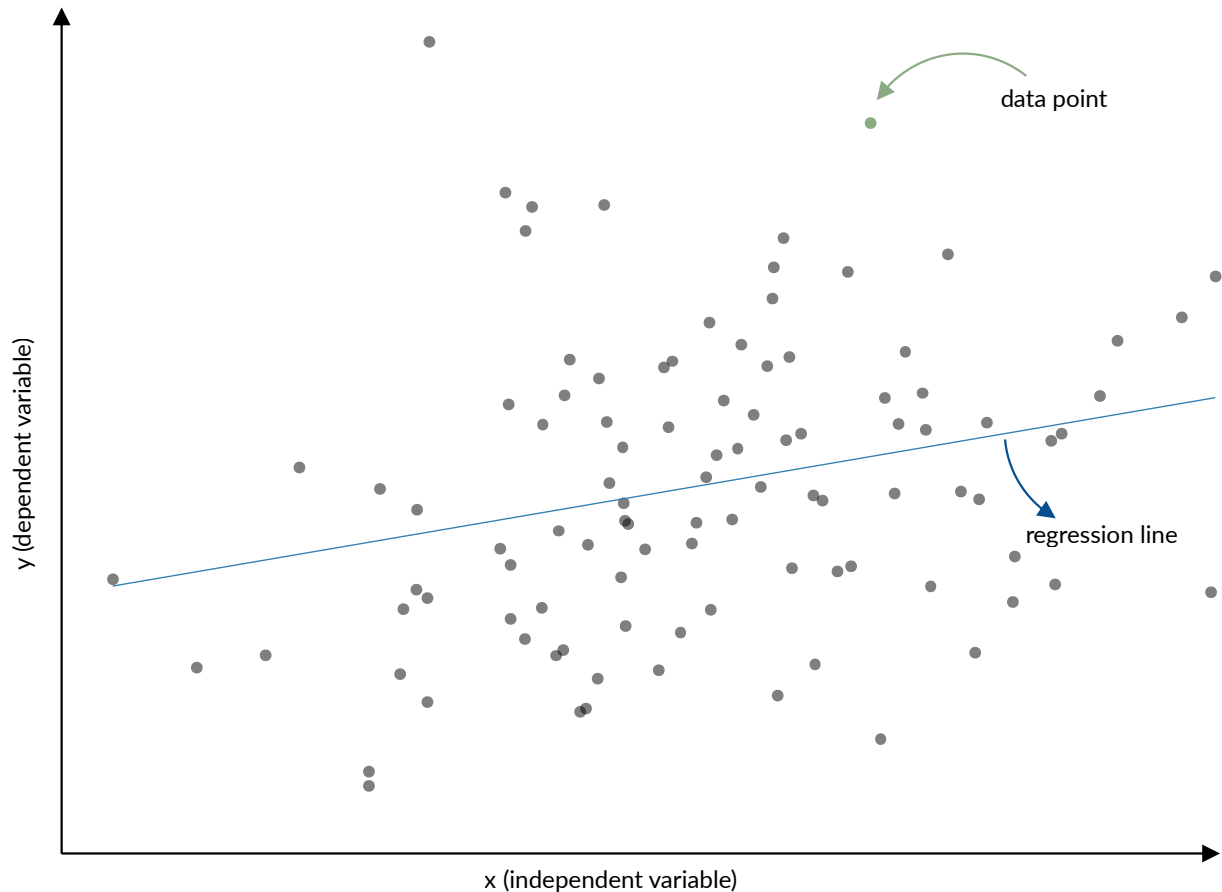
A simple linear regression with only one dependent variable and one independent variable can be represented as a geometric line in a two-dimensional coordinate system with a horizontal x-axis and a vertical y-axis (regression line). The values of the dependent variable are mapped on the vertical axis – the dependent variable is thus often denoted as y . The independent variable is mapped horizontally and denoted by x . The relationship between the two variables is modeled as a line:

$$y = b_0 + b_1x + e$$

b_1 represents the slope of the line, i.e. how much y increases with a one-unit increase of x . b_0 represents the intercept. It determines the position of the line in the coordinate system, i.e. at which level of y the line intersects the y -axis when x is zero. e is the error of the model. It is the variability of y that cannot be explained by the regression line. Chart 9 illustrates a simple regression model.

Chart 9: Illustration of a simple regression

A simple linear regression



Source: Simulated data.

Very similar to the calculation of a correlation, the regression slope can be calculated as the covariance of x and y divided by the variance of x :

$$b_1 = \frac{\text{Cov}(x, y)}{s_x^2}$$

Unlike the correlation, however, the covariance is scaled by the variance only of x so that an increase of x by 1 leads to a b_1 increase in y . However, using two standardized (z-transformed) variables in a regression gives the regression coefficient as b_1 . The intercept ensures that the line goes through the mean of y (\bar{y}). It is calculated as:

$$b_0 = \bar{y} - b_1 \bar{x}$$

The regression line has an important property: It minimizes the sum of squared *vertical* distances of y to the regression line. The distances are also called *residuals* in regression analysis and are denoted by e in the regression model, representing the unexplained variability of y .

The regression line can thus also be thought of as the straight line that best reflects the *mean* of y for all values of x . A regression where the only independent variable is a group variable that has been coded as 0 (e.g. for the control group) and 1 (e.g. for the treatment group) will result in a b_1 that reflects the difference in means between the two groups and is exactly equivalent to an independent t-test. Just as the mean, the independent t-test and the correlation, regressions are susceptible to outliers because they are, in effect, based on the sum of squares.

As with the statistical tests before, the b_1 and b_0 in a regression can be standardized to a t-statistic, and a p-value can be obtained for significance tests. Moreover, similar to the partial correlation, a regression can include two or more independent variables where each slope is controlled for all other variables that were entered into the model. Many more extensions to the simple regression model exist that exceed the scope of this paper. The following models are important to mention: 1) nonlinear regression models, where the slope of the line changes with x , 2) models with interaction terms, where the effect of one variable depends on the level of another variable, and 3) hierarchical or mixed models, where clustered data structures can be modeled explicitly.

2.5 Caveats and alternatives

This short introduction to statistical inference omits many important assumptions and caveats about the presented tests, leaves out countless alternative methods and approaches to hypothesis testing and statistical inference and has generally simplified a continuously growing, complex field of science.

First, it is important to recognize that *statistical significance* does not always mean *practical significance*, which refers to the actual, real-world impact of an outcome. For example, a result could be highly significant but economically or educationally irrelevant. Practical significance can, for example, be assessed using effect sizes.

Second, inferential statistical analysis is based on specific *assumptions* about the sample and population. This short guide strategically ignores many aspects. For instance, the methods presented above generally assume that variables are on an interval scale, which may not necessarily be true for many variables collected through questionnaires. Nevertheless, Likert-scale data typically collected in questionnaires are often (and only sometimes rightfully) treated as being continuous. All tests presented above are also parametric, meaning they make strict assumptions about the distributions of the variables in the population (e.g. that they are normally distributed). The presented tests are often robust to the assumption of normality with larger sample sizes due to the central limit theorem (see section 2.4.4) and other violations may not bias results severely under many circumstances. However, alternative nonparametric tests may be better suited in many situations. An example for a nonparametric test that can be used to compare the central tendency of two groups is the Mann-Whitney U test.

Third, the presented methods cover only a small subset of the vast landscape of statistical methodology. While many statistical frameworks, methods and tests are inherently related, they often use different terminology and may be used across different fields and for distinct purposes. Differences in means, for instance, can be tested using analysis of variance (ANOVA), measurement instruments can be validated using factor analysis or item response theory (IRT, see de Ayala, 2013) and complex relationships between different types of variables can be analyzed and modeled e.g. by using structural equation modeling (Kline, 2016).

Fourth, the presented approaches of classical frequentist inference are typically taught in a similar way in social sciences. However, other approaches exist that have different general assumptions about how to quantify reality. Most importantly, in Bayesian statistics, prior information about the world is integrated into the analysis and beliefs about parameters are updated as new data become available. Unlike frequentist statistics, which assumes that population parameters exist as fixed, unknown values, Bayesian statistics

treats parameters as random variables with distributions that reflect this uncertainty (see Greenberg, 2012).

Fifth, statistical inference is sometimes criticized for becoming an end in itself. It can be argued that the obsession with statistically significant results has created a culture in many disciplines that puts undue emphasis on yielding a p-value smaller than 0.05. Worrisome phenomena in this context include the replication crisis (see Shrout and Rodgers, 2018), p-hacking (see Moniz et al. 2025, and Head et al., 2015), hypothesizing after the results are known (HARKing, see Kerr, 1998) and publication bias (see Maier et al., 2022), although efforts such as preregistration are attempts to counteract these issues (see Brodeur et al., 2024).

3 Summary and concluding remarks

This article underscores the crucial role of quantitative data analysis in evaluating financial literacy interventions, offering an accessible introduction to quantitative methods for evaluators, program designers and educators alike.

Section 1 focuses on *descriptive statistics*, emphasizing the importance of effective data visualization. The foundation of impactful visualization lies in selecting the right chart type. To support this, we present commonly used chart types for time series, proportions, relationships and distributions, alongside their prerequisites and associated variables. Additionally, we summarize guidelines for creating clear and meaningful visualizations that balance the need for clarity without generating an information overload. Key principles include adhering to common visualization rules, such as consistent axes and labeling, using intuitive visual cues (e.g. “up” to signify growth) and leveraging thoughtful highlights and colors (e.g. green for agreement, red for disagreement). Ethical data visualization is also addressed, with recommendations for truthful representation, consideration of privacy and transparency about data limitations to prevent misinterpretations, such as confusing correlation for causation. Section 1 concludes by introducing two types of tools for data visualization: out-of-the-box tools and programming-based tools. We summarize their characteristics to help researchers choose the best fit for their needs.

Section 2 provides an introduction to *statistical inference*. While descriptive statistics summarize data, inferential statistics allow for generalizations and the exploration of relationships. We begin by covering key concepts such as quantifying uncertainty and making inferences from a random sample to a larger population. This is followed by an introduction to hypothesis testing, a method used to determine whether differences between groups or associations among variables are statistically significant. Building on these fundamentals, we then show how statistical inference is applied in practice, by exploring how sample statistics are calculated, standardized and compared to theoretical distributions.

In evaluation research, descriptive statistics and statistical inference are essential for describing patterns in data and estimating their strength and generalizability with respect to the population. They are thus the main tools for describing and testing program impacts and identifying the key drivers of outcomes of financial education interventions.

References

- Anderson, D. R., T. A. Williams and J. J. Cochran. 2020.** Statistics for business & economics. Cengage Learning.
- Azzam, T., S. Evergreen, A. A. Germuth and S. J. Kistler. 2013.** Data visualization and evaluation. *New Directions for Evaluation* 2013(139). 7–32.
- Baruffa, O. 2023.** Big book of R. <https://www.bigbookofr.com/chapters/data%20visualization>
- Brodeur, A., N. M. Cook, J. S. Hartley and A. Heyes. 2024.** Do preregistration and preanalysis plans reduce p-hacking and publication bias? evidence from 15,992 test statistics and suggestions for improvement. In: *Journal of Political Economy Microeconomics* 2(3). 527–561.
- de Ayala, R. J. 2013.** The Theory and Practice of Item Response Theory. Guilford Press.
- Evergreen, S. and A. K. Emery. 2016.** Data visualization checklist. https://stephanieevergreen.com/wp-content/uploads/2016/10/DataVizChecklist_May2016.pdf
- Felbermayr, K. 2024.** Qualitative research evaluation – how to get from first ideas to a final paper. OeNB Financial Literacy Evaluation Series.
- Field, A. 2013.** Discovering Statistics Using IBM SPSS statistics (4th edition). Sage Publications.
- Franconeri, S. L., L. M Padilla, P. Shah, J. M. Zacks and J. Hullman. 2021.** The science of visual data communication: What works. *Psychological Science in the Public Interest* 22(3). 110–161.
- From Data to Viz. 2018.** From Data to Viz. <https://www.data-to-viz.com/#explore>
- Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings and C. M. Vermeersch. 2016.** Impact evaluation in practice. World Bank Publications.
- Greenberg, E. 2012.** Introduction to Bayesian econometrics. Cambridge University Press.
- Head, M. L., L. Holman, R. Lanfear, A. T. Kahn and M. D. Jennions. 2015.** The extent and consequences of p-hacking in science. In: *PLOS Biology* 13(3). e1002106.
- Janczyk, M. and R. Pfister. 2023.** Understanding Inferential Statistics. From A for Significance Test to Z for Confidence Interval. Springer.
- Kline, R. B. 2016.** Principles and Practice of Structural Equation Modeling (4th edition) 2016. Guilford Press.
- Kerr, N. L. 1998.** HARKing: Hypothesizing after the results are known. In: *Personality and Social Psychology Review* 2(3). 196–217.
- Lorenz, T. 2024.** Mixed methods – a practical guide for the gold standard of evaluation research. OeNB Financial Literacy Evaluation Series.
- Maier, M., F. Bartoš, T. D. Stanley, D. R. Shanks, A. J. Harris and E. J. Wagenmakers. 2022.** No evidence for nudging after adjusting for publication bias. In: *Proceedings of the National Academy of Sciences* 119(31). e2200300119.
- Moniz, P., J. N. Druckman and J. Freese. 2025.** The file drawer problem in social science survey experiments. In: *Proceedings of the National Academy of Sciences* 122(12). e2426937122
- Muth, L. C. 2018a.** What to consider when creating pie charts. <https://blog.datawrapper.de/pie-charts/>
- Muth, L. C. 2018b.** What to consider when creating stacked column charts. <https://blog.datawrapper.de/stacked-column-charts/>
- Muth, L. C. 2020.** What to consider when visualizing data for colorblind readers: Part 2 of a three-part series on colorblindness. <https://blog.datawrapper.de/colorblindness-part2/>
- Muth, L. C. 2022.** Which fonts to use for your charts and tables and how to customize them. <https://blog.datawrapper.de/fonts-for-data-visualization/>
- Nordmann, E., P. AcAleur, W. Toivo, H. Paterson and L. M. De Bruine. 2022.** Data visualization using R, for researchers who don't use R. https://osf.io/bj83f/?view_only=

- Pearce, R. 2020.** Why you sometimes need to break the rules in data viz. <https://medium.economist.com/why-you-sometimes-need-to-break-the-rules-in-data-viz-4d8ece284919>
- Rapp, A. 2022.** Bar plot checklist. https://albert-rapp.de/posts/ggplot2-tips/16_bars_checklist/16_bars_checklist
- Rapp, A. 2025.** Blog. <https://albert-rapp.de/blog>
- Scherer, C. 2019.** A ggplot2 tutorial for beautiful plotting in R. <https://www.cedricscherer.com/2019/08/05/a-ggplot2-tutorial-for-beautiful-plotting-in-r/>
- Scherer, C. 2023.** Data Visualization & Information Design. <https://www.cedricscherer.com/>
- Shrout, P. E. and J. L. Rodgers. 2018.** Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. In: *Annual Review of Psychology* 69(1). 487–510.
- Sievert, C. 2020.** Interactive web-based data visualization with r, plotly, and shiny. Chapman and Hall/CRC.
- Storytelling with data. 2023.** Choose an effective visual with the SWD Chart Guide. <https://www.storytellingwithdata.com/chart-guide>
- Wilke, C. O. 2019.** Fundamentals of data visualization: A primer on making informative and compelling figures. O'Reilly Media.
- Wooldridge, J. M. 2009.** Introductory econometrics: A modern approach (4th edition). South-Western.
- Yau, N. 2011.** Visualize this: The FlowingData guide to design, visualization, and statistics. John Wiley & Sons.
- Yoong, J., K. Mihaly, S. Bauhoff, L. Rabinovich and A. Hung. 2013.** A Toolkit for the Evaluation of Financial Capability Programs in Low-, and Middle-Income Countries. Washington, DC: The World Bank.
- Zhang, X., O. L. O. Astivia, E. Kroc and B. D. Zumbo. 2023.** How to think clearly about the central limit theorem. In: *Psychological Methods* 28(6). 1427.

Annex: R code to reproduce the charts shown in this article

```

options(scipen = 999)
options(pillar.sigfig = 4)

## Library
setwd("/path")

library(readxl)
library(haven)
library(srvyr)
library(tidyverse)
library(camcorder)
library(scales)
library(ggplot2)
library(reshape2)
library(ggtext)
library(patchwork)
library(ggrepel)
library(glue)
library(unhcrthemes)
library(lubridate)
library(showtext)

#### load font ####
font_add_google("Lato", "lato")      # Regular Lato
font_add_google("Lato", "lato_bold") # Optional: bold version

# Turn on showtext
showtext_auto()

th <- theme(
  legend.title = element_blank(),
  axis.title.y = element_text(size = 8, family = "lato", color = "black"),
  axis.title.x = element_text(size = 8, family = "lato", color = "black"),
  axis.text.x = element_text(size = 8, family = "lato", colour = "black"),
  axis.text.y = element_text(size = 8, family = "lato", colour = "black"),
  plot.margin = margin(25, 25, 10, 25),
  plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
margin=margin(0,0,10,0)),
  plot.subtitle = element_text(size = 8, family = "lato_bold", face="bold", color = "black",
margin=margin(0,0,10,0)),
  plot.tag = element_text(size = 11, family = "lato"),
  plot.tag.position = "topright",

```

```

plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,0,0)),
plot.caption.position = "plot",
panel.background = element_rect(fill = "#FFFFFF"),
axis.ticks = element_blank(),
axis.ticks.length = unit(0, "cm"),
panel.grid.major.x = element_line(color = "#E5E4E2", size = 0.2),
panel.grid.minor.x = element_blank(),
legend.position = "top",
legend.direction = 'horizontal',
legend.text = element_text(
  size = 8,
  family = "lato",
  color = "black"),
legend.margin = margin(0, 0, 0, 0),
legend.justification.top = "left",
legend.justification.left = "top",
legend.justification.bottom = "right",
legend.justification.inside = c(1, 1),
legend.location = "plot",
plot.title.position = "plot",
legend.key.height = unit(0.3,"cm"),
legend.key.width = unit(0.5,"cm")
) +
theme(
  plot.background = element_rect(fill = "transparent", colour = NA)
)

```

```
##### set Corporate Design #####
```

```

colors <- c(
  `OeNB-Blau` = '#004289', #rgb(0,66,137,maxColorValue = 255),
  `OeNB-Sand` = '#f2e5c0', #rgb(242,229,192,maxColorValue = 255),
  `OeNB-Salbei` = '#95af83', #rgb(149,175,131,maxColorValue = 255),
  `OeNB-Türkis` = '#46aaaf', #rgb(70,170,175,maxColorValue = 255),
  `OeNB-Violett` = '#6a5988', #rgb(106,89,136,maxColorValue = 255),
  `OeNB-Koralle` = '#d36a5f', #rgb(211,106,95,maxColorValue = 255),
  `OeNB-Bronze` = '#bc8034', #rgb(188,128,52,maxColorValue = 255),
  `OeNB-Grau` = '#d1d9dd', #rgb(209,217,221,maxColorValue = 255),
  `Österreich (rot)` = '#e30613', #rgb(220,0,0,maxColorValue = 255),
  `EU (blau)` = '#164194', #rgb(22,65,148,maxColorValue = 255),

  `OeNB-Blau hell` = '#80A2C8',
  `OeNB-Sand hell` = '#faf7ef',
  `OeNB-Salbei hell` = '#dfe7da',
  `OeNB-Türkis hell` = '#9aced5',

```

```
`OeNB-Violett hell` = '#9e90b6',
`OeNB-Koralle hell` = '#e9b5af',
`OeNB-Bronze hell` = '#dcb47f',
`OeNB-Grau hell` = '#f3f5f6',

`OeNB-Blau dunkel` = '#001E3D',
`OeNB-Sand dunkel` = '#d9be81',
`OeNB-Salbei dunkel` = '#627c50',
`OeNB-Türkis dunkel` = '#0c453f',
`OeNB-Violett dunkel` = '#30283e',
`OeNB-Koralle dunkel` = '#903227',
`OeNB-Bronze dunkel` = '#704d1f',
`OeNB-Grau dunkel` = '#95a8b1')
```

scatter plot

Sample data

```
data1 <- data.frame(x = 1:10,
                    y = c(2, 3, 5, 7, 11, 13, 17, 19, 23, 29))
```

Create the plot

```
scatter <- ggplot(data1, aes(x = x, y = y)) +
  geom_point(color = "#004289") +
  labs(
    title = "Title",
    x = "x-axis Label",
    y = "y-axis Label",
    caption = "Source: Simulated data. \nNote: Sample scatter plot."
  ) +
  theme(
    scale_x_continuous(limits = c(0, max(data1$x)), expand = expansion(mult = c(0.05, 0.1))) + # X-axis
    with more spacing
    scale_y_continuous(limits = c(0, max(data1$y) + 1), expand = expansion(mult = c(0.05, 0.1))) # Y-axis
    with more spacing
    theme(plot.caption = element_text(lineheight = 1.3)) +
    theme(plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,-20,0))) +
    theme(plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
    margin=margin(-20,0,10,0)))
```

save plot as svg, use device = svg to save it correctly as svg

```
ggsave(scatter, file = "img/scatter.svg", height = 16 * 20 / 32, width = 16, units = "cm", device = svg)
```

time series - bar and line chart

bar chart #####*# Sample time series data*

```
data2 <- data.frame(Year = 2015:2024,
                    Score = c(65, 67, 70, 72, 74, 75, 77, 78, 80, 82))
```

Create the bar chart

```
bar <- ggplot(data2, aes(x = Year, y = Score)) +
  geom_bar(stat = "identity",
           fill = "#80A2C8",
           width = 0.7) +
  labs(
    title = "Financial literacy scores over the years",
    x = "",
    y = "Score",
    caption = "Source: Simulated data. \nNote: Financial literacy scores of students over the years."
  ) +
  theme(
    scale_x_continuous(breaks = 2015:2024, limits = c(2014.5, 2024.5)) +
    scale_y_continuous(breaks = seq(0,90, by = 10), limits = c(0, 90)) +
    coord_cartesian(expand = FALSE, clip = "on") +
    panel.grid.major.x = element_line(color = "white", size = 0.2), # add theme options
    panel.grid.major.y = element_line(color = "#E5E4E2", size = 0.2))
```

line chart

```
line <- ggplot(data2, aes(x = Year, y = Score)) +
  geom_line(color = "#004289", size = 1) +
  geom_point(color = "#004289", size = 2) +
  labs(
    title = "Financial Literacy Scores Over Years",
    x = "",
    y = "Score",
    caption = "Source: Simulated data. \nNote: Financial literacy scores of students over the years."
  ) +
  theme(
    scale_x_continuous(breaks = 2015:2024, limits = c(2014.5, 2024.5)) +
    scale_y_continuous(breaks = seq(0,90, by = 10), limits = c(0, 90)) +
    coord_cartesian(expand = FALSE, clip = "on") +
    theme(panel.grid.major.x = element_line(color = "#E5E4E2", size = 0.2),
          panel.grid.major.y = element_line(color = "white", size = 0.2))
```

combine bar and line plot and add lines

```
timeseries <- patchwork::wrap_plots(bar, line,
                                    ncol = 1, nrow = 2)
```

```
ggsave(timeseries, file = "img/timeseries.svg", height = 16 * 32 / 32, width = 16, units = "cm", device =
svg)
```

proportions - stacked bar chart and area chart ####**##### stacked bar chart #####***# sample data*

```
data3 <- data.frame(
  Category = c("Statement 4", "Statement 3", "Statement 2", "Statement 1"),
  fully_Agree = c(10, 20, 30, 50),
  mostly_Agree = c(40, 30, 20, 20),
  mostly_Disagree = c(40, 30, 20, 20),
  fully_Disagree = c(10, 20, 30, 10)
)
```

Convert data to long format

```
data3_long <-
tidyr::pivot_longer(
  data3,
  cols = -Category,
  names_to = "Response",
  values_to = "Count"
)
```

Calculate percentages

```
data3_long <- data3_long %>%
  group_by(Category) %>%
  mutate(Percentage = round(Count / sum(Count), 2)) %>%
  mutate(
    Response = case_when(
      Response == "fully_Agree" ~ "Fully agree",
      Response == "mostly_Agree" ~ "Mostly agree",
      Response == "mostly_Disagree" ~ "Mostly disagree",
      Response == "fully_Disagree" ~ "Fully disagree"
    )
  ) %>%
  ungroup()
```

Create the stacked bar chart

```
stackedbar <-
ggplot(data3_long, aes(
  x = factor(
    Category,
    levels = c("Statement 4", "Statement 3", "Statement 2", "Statement 1")
  ),
  y = Percentage,
  fill = factor(
    Response,
    levels = c(
      "Fully disagree",
```

```

    "Mostly disagree",
    "Mostly agree",
    "Fully agree"
  )
)
)) +
geom_bar(stat = "identity", width = 0.6) +
scale_fill_manual(
  values = c(
    "Fully agree" = "#95af83",
    "Mostly agree" = "#dfe7da",
    "Mostly disagree" = "#e9b5af",
    "Fully disagree" = "#d36a5f"
  )
) +
coord_flip(expand = FALSE, clip = "off") +
th +
labs(title = "Agreement to statements",
      fill = "Response",
      caption = "Source: Simulated data.") +
theme(axis.title.x = element_blank()) +
theme(axis.title.y = element_blank()) +
guides(fill = guide_legend(reverse = TRUE)) +
scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
scale_x_discrete(labels = \(Category) str_wrap(Category, 40)) + # required for wrap
geom_text(
  aes(label = paste0(round(Percentage * 100, 0), "%")),
  size = 2.5,
  position = position_stack(vjust = 0.5),
  family = "lato_bold"
) +
theme(plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,20,0))) +
theme(plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
margin=margin(-20,0,10,0)))

```

area chart

Create the 'month' column

```

months <-
rep(seq.Date(
  from = as.Date("2022-01-01"),
  by = "month",
  length.out = 12
), each = 4)

```

```
# Create the 'values' column
```

```
values <-
  rep(c(
    "Fully agree",
    "Mostly agree",
    "Mostly disagree",
    "Fully disagree"
  ),
  times = 12)
```

```
# Create a random 'proportion' column such that each group of months sums up to 1
```

```
set.seed(123) # For reproducibility
proportions <- unlist(lapply(1:12, function(x) {
  sample_values <- runif(4)
  sample_values / sum(sample_values)
}))
```

```
# Create the final data frame
```

```
data5 <-
  data.frame(month = months,
            values = values,
            proportion = proportions)
```

```
# Plot
```

```
area <- ggplot(data5) +
  geom_area(aes(
    x = months,
    y = proportions,
    fill = factor(
      values,
      levels = c(
        "Fully disagree",
        "Mostly disagree",
        "Mostly agree",
        "Fully agree"
      )
    )
  )) +
  labs(title = "Agreement to statement 4 | 2024",
       caption = "Source: Simulated data.") +
  scale_x_date(
    breaks = seq(as.Date("2022-01-01"), by = "month", length.out = 12),
    labels = c("Jan.", "Feb.", "Mar.", "Apr.", "May", "June", "July", "Aug.", "Sep.", "Oct.", "Nov.", "Dec.")
  ) +
```

```

scale_fill_manual(
  values = c(
    "Fully agree" = "#95af83",
    "Mostly agree" = "#dfe7da",
    "Mostly disagree" = "#e9b5af",
    "Fully disagree" = "#d36a5f"
  )
) +
th +
theme(axis.title.y = element_blank()) +
theme(axis.title.x = element_blank()) +
scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
guides(fill = guide_legend(reverse = TRUE)) +
coord_cartesian(expand = FALSE, clip = "on") +
theme(plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,20,0))) +
theme(plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
margin=margin(-20,0,10,0)))

```

```

ggsave(stackedbar, file = "img/stackedbar.svg", height = 16 * 14 / 32, width = 16, units = "cm", device =
svg)
ggsave(area, file = "img/area.svg", height = 16 * 18 / 32, width = 16, units = "cm", device = svg)

```

distributions - density plot and boxplot

boxplot

```

set.seed(123) # For reproducibility
grades <- c("Grade 1", "Grade 2", "Grade 3")
data6 <- data.frame(Grade = factor(rep(grades, each = 30)),
  Score = c(
    rnorm(30, mean = 35, sd = 10),
    # Scores for Grade 1
    rnorm(30, mean = 40, sd = 12),
    # Scores for Grade 2
    rnorm(30, mean = 45, sd = 15) # Scores for Grade 3
  ))

```

Create the boxplot

```

box <- ggplot(data6, aes(
  x = Grade,
  y = Score,
  fill = factor(Grade, levels = c("Grade 1", "Grade 2", "Grade 3")))
) +

```

```

geom_boxplot(varwidth = TRUE,
             alpha = 0.8,
             outlier.shape = NA) +
stat_boxplot(geom = 'errorbar', width = 0.1) +
th +
scale_color_manual(values = c(
  "Grade 1" = "#001E3D",
  "Grade 2" = "#d9be81",
  "Grade 3" = "#627c50"
)) +
scale_fill_manual(values = c(
  "Grade 1" = "#001E3D",
  "Grade 2" = "#d9be81",
  "Grade 3" = "#627c50"
)) +
theme(legend.position = "none",
      panel.grid.major.y = element_blank()) +
labs(title = "Financial literacy scores in three grades",
      caption = 'Source: Simulated data. \nNote: Financial literacy scores of students.') +
coord_flip(expand = TRUE, clip = "off") +
scale_y_continuous(limits = c(0, 100),
                  breaks = seq(0, 100, by = 10)) +
theme(axis.title.y = element_blank()) +
#geom_jitter(aes(color = Grade), size = 1) +
theme(plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,20,0))) +
theme(plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
margin=margin(-20,0,10,0)))

```

density plot

```

density <-
ggplot(data6, aes(x = Score, fill = factor(
  Grade, levels = c("Grade 1", "Grade 2", "Grade 3")
))) +

geom_density(alpha = 0.5) +
labs(
  x = "Score",
  y = "Density",
  title = "Financial literacy scores in three grades",
  caption = 'Source: Simulated data. \nNote: Financial literacy scores of students.'
) +
th +
scale_x_continuous(limits = c(0, 100),
                  breaks = seq(0, 100, by = 20)) +
scale_color_manual(values = c(

```

```

"Grade 1" = "#001E3D",
"Grade 2" = "#d9be81",
"Grade 3" = "#627c50"
)) +
scale_fill_manual(values = c(
  "Grade 1" = "#001E3D",
  "Grade 2" = "#d9be81",
  "Grade 3" = "#627c50"
)) +
theme(plot.caption = element_text(size = 8, family = "lato", hjust = 0, margin=margin(10,0,20,0))) +
theme(plot.title = element_text(size = 11, family = "lato_bold", face="bold", color = "#004289",
margin=margin(-20,0,10,0)))

# combine bar and line plot and add lines
distribution <- patchwork::wrap_plots(density, box,
                                     ncol = 1, nrow = 2)

ggsave(distribution, file = "img/distribution.svg", height = 16 * 32 / 32, width = 16, units = "cm",
device = svg)
#### print all charts ####
scatter
timeseries
stackedbar
area
distribution

```