MATHEMATICS FRAMEWORK

for California Public Schools

Kindergarten Through Grade Twelve

Adopted by the California State Board
of Education November 2023

Published by the California Department
of Education Sacramento, 2024

CHAPTER 5

Mathematical Foundations for Data Science

Introduction

The ability to work with and understand data is an essential life skill in a world continually inundated with data. As data become omnipresent in all sectors of life—personal, business, academic, and education—community members and citizens need the attitudes, skills, and practices to use data to make informed decisions in professional and personal matters. Data drive students' lives, whether they see it or not. Making sense of data, being able to identify data that are misleading, and using data to make decisions are all important skills for students in their roles as global citizens. Almost all occupations now require that employees collect feedback data and use this information to adjust their practice. Stories about the world are illuminated by massive quantities of data, and community members telling and listening to those stories need to be able to make sense of data to understand their health, their finances, and news.

What Are Data Literacy and Data Science?

Many groups have used different terms to describe the ability to work with and derive meaning from data. These terms include statistical literacy (Bargagliotti et al. 2020), data literacy (Education Development Center 2016), data fluency, and data acumen (National Academies of Sciences, Engineering, and Medicine 2018). Because a full discussion of the academic and industry differences between these terms is outside the scope of this framework, the rest of this chapter uses the terms "data literacy" and "data science."

Annika Wolff and colleagues describe "data literacy" as

> the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use of data. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to select, clean, analyze, visualize, critique, and interpret data, as well as to communicate stories from data and use data as part of a design process. (2016, 10)

Within most mathematics courses, students at the prekindergarten through grade twelve level will be building strong mathematical foundations and be engaged in work that supports data literacy; all California students should graduate from high school with data literacy. However, students should also have access to experiences that extend beyond what many currently experience in their mathematics classrooms and that prepare them for future work in an emerging field called data science.

"Content expectations across multiple school subjects in US primary and secondary education already incorporate at least some learning about data collection, utilization, and analysis. Data-related concepts consistently appear in mathematics, science, computer science, and social studies across states. These existing standards may provide the building blocks or even partially comprise a data science education." (Drozda, Johnstone, and Van Horne 2022, 1)

At the broadest level, data science is "the science of learning from data" (International Data Science in Schools Project (IDSSP) Curriculum Team, 2019, 9) or "the processes and systems that enable the extraction of knowledge or insights from data in various forms, either structured or unstructured" (Berman et al. 2016, 2). There has been an expansion in computing and visualization tools that have made many more techniques available for finding meaning in data—often relying on innovative visualizations of complex data that enable major features to be identified and explored further.

Data science is an emerging discipline, and there is yet to be a consensus on exactly what content constitutes "data science." Data science education organizations have conceptualized it as a cyclical process that includes problem formulation, data collection, data analysis, and interpretation and communication of findings (Bargagliotti et al. 2020; IDSSP Curriculum Team 2019). While the data science field continues to be shaped by emerging technologies and techniques deeply influenced by academia and business sectors, most definitions of data science recognize an intersection of mathematics, statistics, and computer science. Data science may also include domain knowledge, ethics, and communication skills as well as specific approaches such as data mining (especially for data collected through the internet and electronic devices) and machine learning (Bargagliotti et al. 2020; Berman et al. 2016; IDSSP Curriculum Team 2019; Rawlings-Goss et al. 2018). Although students may not encounter topics such as machine learning until after high school, kindergarten through grade twelve (K–12) mathematics provides an essential foundation in statistical concepts necessary for future learning in data science.

Data literacy and data science should be thought of more as a continuum than as distinct concepts. The focus of data literacy is the ability to use an inquiry process to extract answers to real-world questions from data sets. All K–12 graduates should have developed data literacy through rich data experiences in each grade level. Data literacy is part of data science, but data science also includes advanced mathematics, statistics, and computational skills that build upon—and go far beyond—the content contained in the K–12 California mathematics standards.

Why a Chapter on the Foundations of Data Science in the Math Framework?

This dedicated chapter on the foundations of data science is included in the framework for two primary reasons. The first has to do with relevance: Data literacy is increasingly central to being able to understand and fully participate in modern life. Discussions around health care, finance, electoral politics, and other major challenges of the current era all revolve around data. Since students frequently encounter claims made from such large data sets, it is crucial that all students have experiences in which they explore the ways such claims are made. The second reason has to do with preparing students for the future. Data science is and will continue to be a fast-growing and lucrative field (Bureau of Labor Statistics 2022). Both of these reasons lead to the same conclusion: Students should have equitable access to data literacy and introductory data science at the K–12 level to facilitate equitable participation in a data-driven world as adults.

Studies by Gregory M. Walton and colleagues show that many students, particularly girls and students of color, do not feel that they belong in certain disciplines (2015). These feelings are often due to a history of negative and off-putting messages (Chestnut et al. 2018). Other studies have shown that different topics and teaching approaches can lead to feelings of belonging or not belonging (Boaler, Cordero, and Dieckmann 2019). Data investigation can support teachers as they seek to create climates of belonging for students, inviting them to investigate real data that is likely relevant to their lives. This meaningful engagement can create opportunities for students to develop self-confidence and self-efficacy with mathematics.

Unequal access to data education in K–12 leads to unequal participation in occupations, systems, and outcomes that create or are heavily impacted by data-related products and endeavors. Consequently, historically marginalized communities have fewer opportunities to reap the economic benefits of data-driven industry. Unequal participation in data science can also lead to bias, not only in the solutions that are developed (e.g., résumé-screening tools that overlook historically underrepresented groups in the industry) but also in the selection of problems that get attention and resources.

Particular aspects of the California Common Core State Standards for Mathematics (CA CCSSM), especially the standards related to statistics instruction, help build the data understanding and skills that high school graduates require. However, the progression of these ideas—from counting, categorizing, and simple picture graphs studied at younger grade levels, to the complex skills and understanding that older students may develop—requires careful thought and considerably more focus throughout the K–12 curriculum than most students have historically experienced. This chapter is a first step in identifying how the current standards can support data literacy across the grade levels and help K–12 students develop foundational knowledge and skills for data science. Subsequent chapters (6–8) will provide additional grade band–specific examples of how data can be integrated into mathematics. Learning about the mathematical and statistical concepts and practices associated with data science may help teachers energize their mathematics instruction and extend their students' mathematical experiences in new ways.

California is not alone in giving attention to this growing field. State departments of education in Georgia, Ohio, Oregon, Utah, and Virginia are also exploring how to increase access to data science concepts for their students, whether by revising mathematics standards, creating new mathematics course pathways or frameworks, or providing micro-credentials in data science for teachers (National Center for Education Research, Institute of Education Sciences 2021). Other countries, including China, New Zealand, and South Korea, are also investing in more data science instruction for all K–12 students (Data Science 4 Everyone 2022).

In the past, statistics instruction focused on just a few key ideas and procedures (mean, median, standard deviation, interquartile range, correlation, and linear regression, along with a few data visualizations such as line plots and scatter plots) or was overlooked altogether. As students progress through school, they should learn different approaches to statistical analysis, culminating in the investigation of large data sets using appropriate technological tools. Elevating statistics as a way to understand the world and solve problems at the K–12 level is an important step in supporting data literacy for all students and building a pathway for an introduction to data science in the third or fourth year of high school.

This emphasis is not meant to suggest that statistics should replace other math content. Statistics and other math domains are mutually reinforcing. For example, an understanding of linear regression is closely related to an understanding of functions and polynomials. A comprehensive understanding of all domains in K–12 mathematics is necessary for successful postsecondary work in data science. If students are intending to pursue STEM majors in college (including data science), they should take courses that, at a minimum, allow them to enter college having completed the prerequisites for calculus. As of this writing, undergraduate data science programs typically require a core math sequence that includes calculus and linear algebra.

The points made above are not meant to suggest that learning statistics is equivalent to learning data science. The types of data being collected are vast and the types of techniques used to extract insights from the data depend on a strong understanding of multiple areas. Students will need to learn how to use computational tools to store, transform, and analyze the quantity of data being generated. Students will also need to have domain knowledge to identify questions that can be investigated through data science and to interpret the results of their analyses in a thoughtful way.

Statistics has become increasingly relevant in applications of mathematics and provides many contemporary illustrations of the significance of the CA CCSSM. Accordingly, this chapter focuses on ways in which teachers can create rich statistics and data experiences across Pre-K–12 that can help (a) modernize the teaching of Statistics and Probability standards, (b) engage students in the kind of authentic problem solving that broadens participation in STEM fields generally and data science specifically, and (c) prepare students for life and work in the data age.

To avoid confusion about terminology, the remainder of this chapter will use the term "data science" to encompass K–12 work that serves these ends.

Using Statistics and Data Science in a Problem-Solving Process to Support the Standards for Mathematical Practice and to Make Connections to Other Domains

With data serving as the basis of large-scale decisions and predictions, all California high school graduates need skills in interpreting and visualizing data, making and critiquing data-based arguments, and gaining some facility with spreadsheets and other tools used to store and analyze data. It is crucial for students to develop the ability to identify types of questions that are subject to exploration through data. Just as crucial is their understanding of some misuses of data. Students must ultimately approach data science and statistics as a problem-solving process that consists of formulating statistical investigative questions, collecting and interrogating existing data, analyzing data, and interpreting and communicating findings (Bargagliotti et al. 2020). Across the K–12 grade levels, students should have opportunities to experience data in different ways, such as the following:

- **Encountering and understanding the role of data in the world:** How do we explore and interpret data and make ethical decisions about how it is used? Students should experience working with data from a context that is meaningful to them personally. They should have opportunities to solve problems of value to them and to their schools and communities.
- **Collecting and exploring data:** How can we collect data? Data explorations should be investigative and collaborative, with students working together to ask investigative questions or engage in statistics as a problem-solving process. Students should have multiple opportunities to become familiar with a variety of technology and modern tools to access, collect, explore, and make sense of data.
- **Considering variability and engaging in multivariate thinking:** How can we describe, display, and compare data effectively? How can we determine the relationship between different variables or quantities? Students should learn to engage with real data that include multiple variables. At first, students can learn to understand the relationship between two variables with bivariate data; as they progress through the grades, they can learn to handle multivariable data and multivariate thinking. Multivariable data often include three or more variables. For example, students could categorize their favorite stuffed animals by considering their size, fluffiness, and type of animal (three variables).
- **Considering data sampling and probability:** How can we use random sampling to help understand a population? How can we determine the chances that an event or events will occur? Technology and tool use should become

more complex as students progress through the grades and can help them explore the role of sampling and probability.

- **Interpreting and communicating findings:** What do our data mean? Does our analysis address any of our questions? What are the best ways to communicate our findings? What impacts might the findings have? As students learn to interpret data in increasingly sophisticated ways, they should also have opportunities to make statements about the data and to practice using data visualizations to communicate results. Especially in middle and high school, students' encounters with data should revisit the context from which the data originated, interpreting results in that context.

Throughout the CA CCSSM, there are multiple opportunities to support such data-rich experiences and integrate the five components of equitable and engaging teaching described in chapter two, even if the standards domains do not appear explicitly within a grade or grade band. When approaching the grade band chapters in this framework from a data science lens, educators can find additional moments to integrate data into students' mathematical experiences. For example, chapter six includes a vignette describing Mrs. Verners' fourth-grade lessons supporting Number and Operations in Base Ten and Operations and Algebraic Thinking ("Comparing Numbers and Place Value Relationships in Grade Four, with Integrated English Language Development"). In her class, students explore population data by making estimates based on prior knowledge, exploring data both in written and standard form, considering place value, and making multiplicative comparisons. The lessons help students focus on changing mathematical quantities while also connecting to social studies content and integrating English language arts/English language development standards in a meaningful way. Similarly, as described in chapter eight, alternative third- and fourth-year high school courses can also provide valuable opportunities to explore important data science topics beyond statistics, such as ethics, data modeling and simulations, and data cleaning. Two important sources for contexts in which to explore statistics and data science are:

- The California Next Generation Science Standards (CA NGSS) (California Department of Education 2013a)
- The California Environmental Principles and Concepts (EP&Cs) (California Department of Education 2013b)

In addition to connecting data-rich experiences to other content, data investigations can support students to draw on the Standards for Mathematical Practices (SMPs) as they engage in this problem-solving process across every grade band. For example, students should reason abstractly and quantitatively by engaging in statistical thinking while considering where data come from (SMP.2); apply statistical models to include descriptions of the variability present in data (SMP.4); and consider available tools such as calculators, spreadsheets, applets, statistical packages, and graphical displays to help facilitate the statistical problem-solving process (SMP.5). When students participate in the analysis of large data sets, they should be able to decide which questions matter and identify which ones can be answered with a given data set (SMP.4). Figure 5.1 illustrates an example of how the SMPs can be highlighted within an elementary data investigation.

Figure 5.1: Investigating Ladybugs to Support the Standards for Mathematical Practice

Brief Description of the Learning Activity (Level A Ladybugs Example)	Connections to the Standards for Mathematical Practice
Students formulate statistical investigative questions: "How many spots do ladybugs typically have?" Or "Do red-bodied ladybugs tend to have more spots than black-bodied ladybugs?"	**SMP.1: Make sense of problems** Consider which of our questions can be answered with data. Interesting investigations anticipate that data collected will vary or are not the same for every observation.
Students collect data: Students recognize that they can use photographs to help answer data collection questions—e.g., "What is the body color?" or "How many spots are on each ladybug?" These questions generate data that are both numeric and categorical.	**SMP.5: Use appropriate tools strategically** Students use a nontraditional data source—photographs of ladybugs—and develop a data collection plan for the class to use. Tables are helpful tools for organizing individual data or for collecting and organizing data from multiple students. **SMP.6: Attend to precision** Humans make decisions that impact data collection and resulting analyses or interpretations.
Students analyze data by making plots and describing them: Students make dot plots for their ladybug data and describe the number of spots that were most common, visually estimate the median, and identify how these values compared for ladybugs of different colors. Students can use a probability chart (0 = not probable, 1/2 = equal chance, 1 = very likely) to express whether they think something will happen.	**SMP.2: Reason abstractly and quantitatively** Students can compare the distribution of the number of spots for ladybugs of different colors. Students make informal associations between ladybug color and their numbers of spots, for example black-bodied ladybugs have fewer spots than red-bodied and orange-bodied ladybugs. Students use data to consider the likelihood that something will happen, such as finding a black-bodied ladybug with 14 spots.
Students interpret and use the plots to answer their initial questions: Students describe the distribution of spot numbers and range of spot numbers for ladybugs.	**SMP.3: Construct viable arguments** Students consider the limitations of their data—for example, that their information probably is not sufficient to describe all the ladybugs in the world (make population inferences).

Source: Adapted from Bargagliotti et al. (2020)

Thematic Topics within the CA CCSSM That Directly Support Data Science

Mathematics and statistics make up a significant portion of a data science education. Topics related to data primarily are found within two domains of the standards: (a) Measurement and Data and (b) Statistics and Probability. This section describes three thematic topics developed from the CA CCSSM that support data science and demonstrate how the topics progress across the grade bands. Thematic topics were created by applying the "Essential Understandings" of statistics described by the National Council of Teachers of Mathematics (NCTM) (Kader et al. 2013; Peck et al. 2013) and aligned to the CCSSM and the statistics developmental levels from GAISE II (Bargagliotti et al. 2020).

The thematic topics are:

- Understanding and describing variability in data and data distributions
- Data collection, sampling, and random processes
- Comparing distributions and identifying associations between variables

The subsequent sections of this chapter show how the thematic topics connect across grade bands to create the foundational knowledge for data science, beginning with simple counting and categorizing activities in kindergarten and culminating in high school where students integrate all these concepts during sophisticated investigations involving linear models and inferential statistics (see figure 5.2). In figure 5.2, the bullet points represent the CA CCSSM clusters from the Measurement and Data and Statistics and Probability domains. Additional details on how the individual standards can be implemented appear in the sections that follow, which are specific to each grade band.

Figure 5.2: Data-Focused CA CCSSM Content Clusters, Organized into Thematic Topics That Span K–12

Grade Levels	Understanding and describing variability in data and data distributions	Data collection, sampling, and random processes	Comparing distributions and identifying associations between variables
K–5 Measurement and data	• Describe and compare measurable attributes • Represent and interpret data	• Classify objects and count the number of objects in categories	n/a
6–8 Statistics and probability	• Develop an understanding of statistical variability • Summarize and describe distributions	• Use random sampling to draw inferences about a population • Investigate chance processes and develop use and evaluate probability models	• Draw informal comparative inferences about two populations • Investigate patterns of associations in bivariate data

Figure 5.2: Data-Focused CA CCSSM Content Clusters, Organized into Thematic Topics That Span K–12 (cont.)

Grade Levels	Understanding and describing variability in data and data distributions	Data collection, sampling, and random processes	Comparing distributions and identifying associations between variables
High School Statistics and probability	• Summarize, represent, and interpret data on a single count of measurement variable • Summarize, represent, and interpret data on two categorical and quantitative variables • Understand and evaluate random processes underlying statistical investigation • Interpret linear models • Make inferences and justify conclusions from sample surveys, experiments, and observational studies • Understand independence and conditional probability and use them to interpret data • Use the rules of probability to compute probabilities of compound events in a uniform probability model • Use probability to evaluate outcomes of decisions	• Summarize, represent, and interpret data on a single count of measurement variable • Summarize, represent, and interpret data on two categorical and quantitative variables • Understand and evaluate random processes underlying statistical investigation • Interpret linear models • Make inferences and justify conclusions from sample surveys, experiments, and observational studies • Understand independence and conditional probability and use them to interpret data • Use the rules of probability to compute probabilities of compound events in a uniform probability model • Use probability to evaluate outcomes of decisions	• Summarize, represent, and interpret data on a single count of measurement variable • Summarize, represent, and interpret data on two categorical and quantitative variables • Understand and evaluate random processes underlying statistical investigation • Interpret linear models • Make inferences and justify conclusions from sample surveys, experiments, and observational studies • Understand independence and conditional probability and use them to interpret data • Use the rules of probability to compute probabilities of compound events in a uniform probability model • Use probability to evaluate outcomes of decisions

Source: Adapted from the CA CCSSM (California Department of Education 2013c)

Understanding and Describing Variability in Data and Data Distributions

Many important outcomes (e.g., health, wealth, education) vary in the world. Gathering data provides a way to capture how these outcomes vary in order to understand the causes of the variation. The patterns of variation that are seen in the data are called distributions. Across the curriculum, students consider how their observed, counted, or measured values and data characteristics might not be the same—that is, they vary. The statistical work of understanding and describing variability provides a strong footing for students to engage in the work of data science. In kindergarten through grade five, it is essential that students encounter variation in a variety of ways, including by counting, measuring, and observing quantities and characteristics that vary in order to be prepared for more sophisticated work with statistics later. Elementary students develop visualizations to show variability in data. Early elementary students begin with creating picture graphs, showing one or more categories of data in whole units. By the end of elementary school, students should have had experience with line plots and bar graphs for data with three or four categories and experience with plots with scales in fractions of units.

From sixth grade, students begin learning more formal methods to understand data and create models of variation. Students continue to produce data visualizations. They learn to describe distributions by their overall shape (e.g., symmetric versus skewed) as well as measures of center (mean, median, mode) and spread. This foundational work is important for being able to compare distributions and identify associations between variables beginning in seventh grade.

Data Collection, Sampling, and Random Processes

Data collection can underpin data science activities. As students look at the world through a data lens, they might notice that data can take different forms. Data collected and represented fall into two categories: categorical (non-numerical, or qualitative) data and numerical or quantitative data. For instance, consider a set of colored blocks in the classroom. Color is a categorical or qualitative variable that students could observe about each block, while length is quantitative, a data point generated through measuring. The standards focus on students developing an understanding of categorical data in kindergarten through grade three; in grade two, students begin to also learn about measurement data. Figure 5.3 illustrates several examples of categorical and quantitative data.

Figure 5.3: Examples of Categorical and Quantitative Data

Categorical Data	Quantitative (or Measurement) Data
• Temperature (hot, room temperature, cold) • Color (red, green, blue, yellow) of blocks in the classroom • Species of trees at the school • Identification of schools in the district as "elementary school," "middle school," or "high school"	• Temperature (80°F) • Pixel or RGB color values • Height (or circumference of trunk, or biomass) of trees at the school • Number of pages (or weight, or height) of books in the classroom • Annual income for households in a census tract

As students pose statistical investigative questions, they should also encounter opportunities to help determine how data might be produced to answer those questions, and what forms of data would be best to use for producing the answers. In addition to producing data directly through their own observations, students should gain exposure to designing and using surveys and simple experiments. By producing their own data from their classroom or community, students recognize data as having context and deriving from observation and measurement, and they come to see data (and mathematics more broadly) as a tool to help think about their worlds.

In seventh grade, students are introduced to the idea of random sampling and the idea that data collected from a subset of a population can help them understand the whole population. Students are also introduced to probability and chance processes in seventh grade, building theoretical probability models and conducting experiments to calculate long-run probabilities of chance events. Students should continue to develop their understanding of sampling, random sampling, and probability models through eighth grade to prepare for work in high school.

Comparing Distributions and Identifying Associations Between Variables

In elementary school and early in middle school, students are primarily working with data sets that include a single variable measured in a single population in mathematics and one to three variables in science. In seventh grade, students continue working with univariate data but begin informally comparing a single variable measured across two populations or at two points in time.

In eighth grade, students begin working with bivariate data: two variables measured in the same population. Students are introduced to the use of scatter plots to visualize bivariate data and depict how the two variables are associated. Students also begin working with informally fitting linear models to scatter plots that suggest a linear association and using those linear models to solve problems and make predictions.

Although students’ statistical explorations of linear models are informal in middle school, middle school work with expressions, equations, and functions is critical to preparing students for more formal use of linear models in high school.

In high school, students use technological tools to create a line of best fit and compute a correlation coefficient. At the high school level, students should be able to interpret the slope and intercept of a linear model and distinguish between correlation and causation.

In high school, students integrate their knowledge of random sampling, comparing distributions, and identifying associations into more complex statistical investigations in which they make inferences from data. Students begin asking whether observed differences between two samples could happen through random chance.

Data Science in Each Grade Band

Transitioning from Prekindergarten

Before kindergarten, children begin to describe their world in language, identifying characteristics of objects, places, people, and events: *The ball is red. My classroom is warm. My teacher is old or young. Our trip to the park was too short.* Identifying characteristics is the beginning of using data and wondering about characteristics—including countable characteristics—is the beginning of asking questions that data can help to answer. In the California Preschool Learning Foundations, this content is located under the heading, “Algebra and Functions (Classification and Patterning),” in which children “sort and classify objects in their everyday environment” (by one attribute at around 48 months and by more than one attribute at around 60 months of age); and in “Measurement,” in which students compare and order objects directly at around 48 months of age and may use an intermediate object for comparison at around 60 months of age (California Department of Education 2008). These preschool activities directly enable the types of kindergarten through grade-five learning trajectories described below.

Kindergarten Through Grade Five

Within the kindergarten through fifth grade band, a sense of curiosity about the world is a crucial first step in building an understanding of what data are and how they can be generated. All work with data should begin with noticing and wondering: “I notice that …” or “I wonder what …” or “I wonder how many ….” To prompt students to wonder, teachers can ask: “What do you notice or wonder about here [in this context] that we could [count/measure/keep track of] to figure out or explore further?” To establish effective routines, and to support language development in “I wonder” activities, it can be effective to provide these examples as sentence starters. Early data explorations might begin with students asking questions that can be answered with a single value, “How many students are there in our class?” or “How long is recess?” (SMP.1). With support from their teachers, students can also start to pose or explore statistical investigative questions that involve multiple variables, such as “I wonder if plants grow more with more sunlight?” or “I wonder if age affects which color people like?” Questions guide much of students’ work with mathematics and include “those used to frame an investigation, those used to collect data, and those used to guide analysis and interpretation” (Bargagliotti et al. 2020, 44).

At the lower elementary level, students encounter data through exposure to small data sets or numbers that were collected manually through counting, classifying, comparing, or possibly measuring objects. Simple peer-questioning activities such as gathering answers to questions like “How many siblings do you have?” or “Was the sky clear or cloudy today?” engage students with basic data collection. Activities focused on a single attribute (e.g., number of siblings) also provide an opportunity for students to engage in the SMPs while the students represent data graphically, and these activities support looking for patterns in data, which is crucial work for any statistical investigation. Upper elementary students explore various types of data representations and use those data representations, generated by themselves or others, to describe the world around them (SMP.2). These experiences lay a critical foundation for mathematical and statistical thinking within middle school and are crucial steps in supporting data literacy.

As students gain confidence in their ability to communicate the mathematical ideas, the teacher should encourage students to generate questions themselves to build their agency in using mathematics to make sense of their worlds or to use their data to develop claims in response to their questions (SMP.3). For example, a weekly whole-class "I wonder" routine—in which students propose questions to investigate by collecting data—contributes to students' development of modeling with mathematics (SMP.4). Exploring data is an opportunity to help students see how mathematics can be used to make sense of problems or answer questions that are relevant to them (SMP.1).

Because the mathematical experiences that support data science increase in sophistication substantially between the early and later elementary grades, thematic topics are discussed separately for the K–2 and 3–5 grade bands in the following sections.

Kindergarten Through Second Grade

Understanding and Describing Variability in Data and Data Distributions

Students naturally encounter simple variation in their everyday lives through tasks that invite qualitative descriptions or comparisons, such as "The same kinds of plant are different sizes" or "The contents of our lunch boxes differ."

Once students are invited to notice things in a context and wonder about a question, they begin to describe measurable, countable, and observable attributes of objects or situations (K.MD.1, K.G.1, K.G.4) and classify objects and count the number in each category (K.MD.3), such as categorizing a set of cubes by color. Sorting objects into two or three categories and representing these categories by their count (K.MD.3, 1.MD.4) are early examples of students representing data to help make sense of their worlds (SMP.4). Basic summary statements of objects pulled from a bag—"The shape is square" and "This cube is red" (categorical data) or "There are 13 red cubes in the set" (numeric data)—represent early work that builds toward an understanding of variability. Notably, most of the focus on number in kindergarten and first grade should be with numbers representing quantities of counted objects (SMP.2). Sorting and categorizing activities help students recognize that objects can naturally vary and help students develop language to express this variability, such as "We have three different shapes: squares, triangles and circles" or "Our red cubes come in two different sizes." Students also begin to use total counts to describe their observation—e.g., "When we pulled shapes from a bag one at a time, 12 were square and 6 were circles."

Many opportunities to explore variation arise as students measure time to the nearest five minutes (2.MD.7) and measure length to the nearest whole unit (2.MD.9), using different standard units (centimeters, meters, inches, feet) (2.MD.3) and several tools (2.MD.1). They might recognize that their measurements are not always the same. Working with data collected in tables and in visualizations provides students an opportunity to explore questions such as, "For the objects measured, what was the most common length in inches? What was the smallest (minimum) or largest (maximum) object? What is the difference between our largest and smallest object (range)?" By second grade, students begin to expand their focus on data representation, being introduced to line plots (whole number units only; 2.MD.9), picture graphs, and bar graphs. These graphs can be used to answer put-together, take-apart, and compare questions (2.MD.10).

Data Collection, Sampling, and Random Processes

Data investigations should be investigative and collaborative, with students working together to learn about and describe the world around them. Collecting data through measurements, surveys, and experiments is an important part of the statistical investigative process and supports young learners in building their awareness of what data are and where data come from. Simple classroom polls provide opportunities for early elementary students to work with simple addition and subtraction equations to express relationships between the collected student responses. For example, the question "How many people took the bus today compared to yesterday?" requires students to collect data and consider how, and perhaps why, quantities might change from day to day. Simple surveys in the form of interviews help students practice expressing counts verbally and symbolically and provide opportunities for students to communicate with each other about their data.

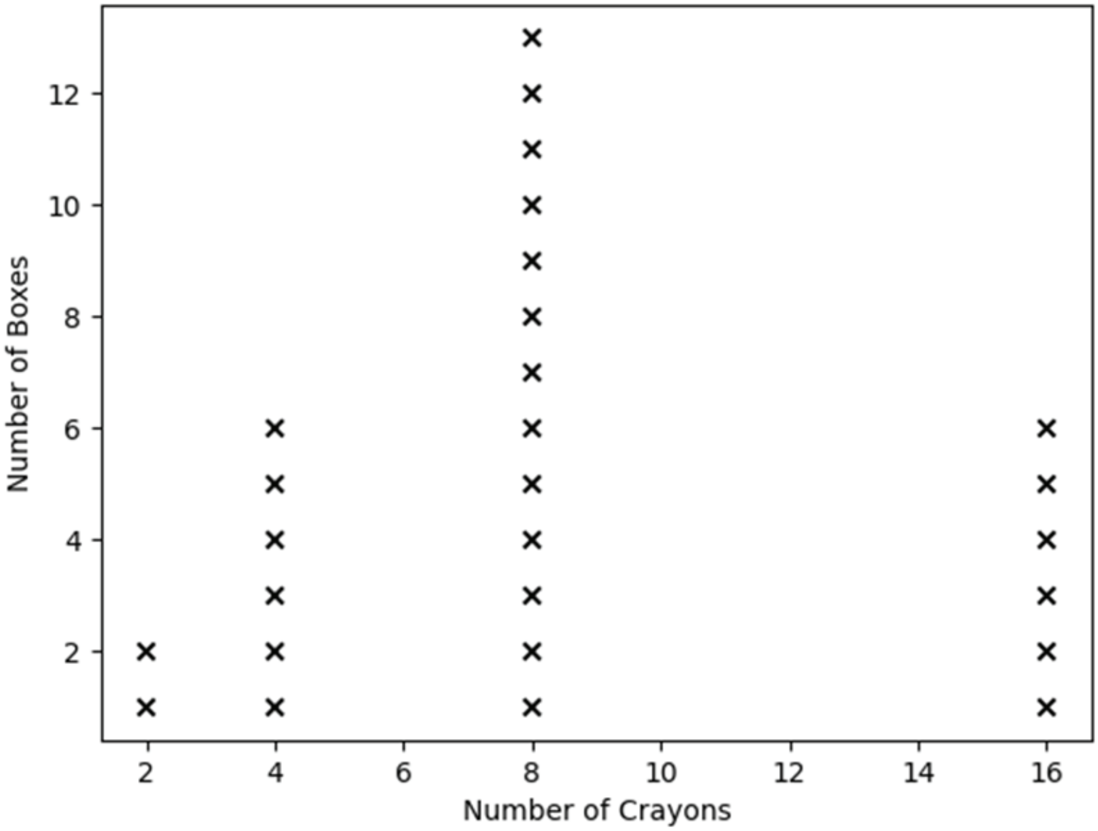
Data collection tasks in the early elementary grades are usually constrained to the context of the classroom. When choosing data tasks, it is important to consider the grade-level expectations for counting (up to 10 objects scattered or up to 20 if arranged in a line, array, or circle in kindergarten [K.CC.5]; 120 by the end of first grade [1.NBT.1]; and up to 1,000 by the end of second grade [2.NBT.2]). Counting tasks can also be structured to build understanding of place value.

Comparing Distributions and Identifying Associations Between Variables

Within kindergarten through second grade, students use language, counts, and measures to describe and compare objects. Students compare the numbers of objects in different categories (K.CC.6) to answer "Which has more?" questions (e.g., "I wonder whether there are more square blocks or more triangular blocks on the desk?"). At first, the teacher suggests or specifies categories; eventually students generate ideas for classification. They also directly compare objects (rather than measuring each with a unit or an intermediate object) with common measurable or countable attributes to see which has more (K.MD.2, K.G.4) ("I wonder which shape has more sides? Which kind of block is heaviest?" [answering this question by using a balance or an informal one-in-each-hand comparison rather than a scale]). "I wonder ..." questions should explore two-category "Which is more?" questions as well as comparisons of objects according to length, height, weight, and countable attributes like number of sides. Student-generated questions provide opportunities to work on precision of language as well; for example, by asking students to clarify what they mean by “bigger.” Mathematics discussions that are rooted in academic language can help students understand mathematical concepts more deeply and discover new ones.

To explore foundational ideas of distribution, students might explore the variety of colors offered in crayons or markers in their classroom. To do so, students collect the counts of crayons or markers in each of the boxes, choosing from different methods of counting, such as using dots, arrays, or tallies. The counts are reported to the teacher, who constructs a basic dot plot, graphically representing the class counts, as shown in figure 5.4. Students can use the counts to first describe the shape of their data by describing where they see “hills” or where the data form a group or crowd (cluster). Students can use their counts to find the smallest count and largest count (range) and to describe the most common crayon box size. Further extensions might include asking which box sizes occur more often or asking students to visually estimate where the middle of the plot would be. These early activities help expose students to key ideas (median and frequency). The teacher might ask the students to use the data to predict. “Imagine if we put all the crayon boxes into a large black bag. If we closed our eyes and reached in, which crayon box size (2, 4, 8, or 16) do we think we would grab first?”

Figure 5.4: A Teacher’s Dot Plot of the Data to Determine the Most Common Crayon Box Size



Long description of figure 5.4

The following elementary school snapshot illustrates many of the ideas discussed in this section. Experiences related to thematic topics (*understanding and describing variability in data; data collection, sampling, and random processes; and comparing distributions and identifying associations between variables*) are included in parentheses where relevant.

Snapshot: Logan's Early Elementary School Explorations with Data

In first grade, student teams were asked to think of two similar things at school for which they were not sure which was taller and then to find a way to compare the objects' heights. A variety of materials was available to use in the comparison (*data collection*). Logan's team was able to compare the height of the slide in front of the school with the height of the slide behind school, measuring the height of both using towers of large DUPLO® bricks. The whole class used their data to discuss how much taller the slide in front of the school was compared to the one in back. Afterward, Logan wanted to build DUPLO® towers to measure height and length of lots of things and was disappointed that the class did not have enough bricks to measure the height of the school (and that their teacher would not let them climb the school).

In another class activity, students recorded the length of each day (sunrise to sunset) by looking at the weather station in the main office. The class maintained a visible running tally of the number of school days with less than 11 hours of daylight, 11 to 13 hours, and more than 13 hours for the entire year (*understanding and describing variability in data*). The class then discussed what the students thought might happen to the number of hours of daylight in the future and checked the data a month later to see whether their predictions were correct.

The students in Logan's second grade class made their own yardsticks by marking a blank wooden rod in inches, using only a three-inch by five-inch card to measure the marks. The class then used the yardsticks extensively to measure objects of interest to the nearest inch. Later, they added centimeter markings to the other side of the yardsticks and discovered that measuring the same things with smaller units led to larger number measurements and improved the quality of data (*data collection*).

When choosing an activity for measuring time, Logan's group decided to time and record the amount of time in a week that team members spent reading in school and then to compare those measurements over several weeks. (This activity had the benefit that team members read much more during those weeks!) Other teams measured time spent playing outside, listening to announcements, and working at math stations (*data collection*). Teams made line plots of their data and compared the line plots of different activities to discuss how students typically spend their school time (*understanding and describing variability in data, and comparing distributions*).

Grades Three Through Five

Understanding and Describing Variability in Data and Data Distributions

Data collected from observations and measurements often vary; that is to say, the values reported are not identical. Variability is a term used to describe how much the values differ from each other or are consistent. If the lengths of 10 NBA basketball courts are measured, the values would be expected to be very nearly identical, but if the numbers of pages in third-grade math textbooks from different publishers were counted, these values would be expected to differ. Foundational work in variability begins in elementary mathematics when students count, measure, observe, and describe their data. The use of simple plots to identify patterns is an integral part of preparing students for the statistical concepts in variability that are covered in grades six through eight.

When working with visualizations of data, students not only should consider the most popular value in a data set (the mode) but also should describe the shape and spread of data distributions. Identifying the maximum and minimum values of quantitative data sets can help students appreciate the concept of range as a measure of spread. Looking for clusters and gaps in a distribution can begin to help them attend to the shapes of data sets. As students engage in experiences in which they produce their own data through measurement, teachers should highlight for students the variation that results. Measuring the same variable on multiple individuals or objects, for example, results in data that vary, and students should consider the causes or sources that might have given rise to the variation they have observed, working as they do so to differentiate between variation and error. For example, if students plant a particular variety of flower seed at multiple locations around the school, then measure the plants' height and the amount of sunlight each month, they can conduct investigations into the ways that plant growth and sunlight relate to each other. They should discuss and describe any patterns in their data and discuss reasons for the variability. Upper elementary students should have the opportunity to represent their data through plots that they themselves create. This process helps students notice variation within the data as well as communicate their thinking in multiple ways.

Students in grades three through five refine their measurements of length and time and expand the set of units they use, adding area and volume measurement to their repertoires. By the fifth grade, students should understand that data sets can include both categorical and numerical data. They should recognize that an individual instance or object can possess attributes that exemplify these different types of variables, and they should have gained experience measuring, characterizing, and analyzing such diverse types of data and associating them together. Mathematical and scientific work that can reveal variability of measured dimensions, mass, and volume present natural opportunities for students to explore variation in their different measurements graphically—such as in a common classroom activity in which students compare the mass or volume of objects to other dimensions.

The following snapshot on Logan's fifth grade exploration with data illustrates many of the ideas discussed in this section. Experiences related to thematic topics (*understanding and describing variability in data; data collection, sampling, and random processes; and comparing distributions and identifying associations between variables*) are included in parentheses where relevant.

Snapshot: An Example of Logan's Fifth-Grade Exploration with Data

By fifth grade, Logan and classmates had constructed many line plots and thus often wondered about quantities that vary on repeated measurement, such as the following:

- The cartons of milk from lunch say they each contain 8 fluid ounces, but yours feels heavier than mine. Does my container have less milk? Are you getting more milk than me?
- The weather site says the average high temperature here is 57°F (degrees Fahrenheit) in November, but today it got up to 65°F. How can we check whether this month is near average?

To explore the first question, the school donated 20 cartons of milk to the experiment so students could measure the volume (*data collection*). When they examined the line plot of their measured volumes, they saw that it had a tightly clustered shape, with a minimum measurement of 7.8 fluid ounces and a maximum of 8.2 fluid ounces, and that the most frequent value was 7.9 ounces (*understanding and describing variability in data*). One student in the group thought that some milk probably remained in the containers, so the group spent a while trying to figure out how they might identify how much had been left inside. The teams came up with several methods, laying the groundwork to talk about random measurement.

For the second question, the class recorded the daily high temperature for each day of the month, recorded these temperatures on a line plot that also had marked the "average" high temperature from the weather site, and used the line plot at the end of the month to discuss whether their measurement was consistent with the stated average (without computing an average of the data).

Fifth grade does not extend the expected set of data representations, but students do use line plots in a sophisticated way that sets the stage for understanding the most common measure of center for a data set—the mean (commonly called the average)—in sixth grade. Namely, fifth grade students use a line plot to decide how a repeatedly measured quantity could be redistributed equally (5.MD.2): "Given different measurements of liquid in identical beakers, find the amount of liquid each beaker would contain if the total amount in all the beakers were redistributed equally."

Although the data visualizations mastered by fifth grade include only picture graphs, bar graphs, and line plots, students do not need to be restricted to these. Each of these represents repeated measurements of a single varying quantity; science curricula, and many questions of interest in general, require the consideration of relationships between *two or more different* changing quantities, such as erosion and time (NGSS 4-ESS2-1 Earth's Systems) or length or direction of shadows and time (NGSS 5-ESS1-2 Earth's Place in the Universe). Reasoning that involves such multiple variables is an important aspect of modern encounters with data, and students should experience this kind of reasoning at all levels (SMP.2). These science investigations represent an excellent opportunity to compare distributions between variables by posing questions such as "How does the shadow length change between fall and spring?"

In recent years, new technological tools and developments have prompted an explosion in interesting data visualizations, many of which are quite comprehensible to young students with some exploration. Technology and the power of computing play an important role in data science and can be incorporated into the kindergarten-through-grade-five experience. The ability to use technology to collect, organize, represent, and share data is fundamental to the development of data literacy. California's 2018 Computer Science Standards include computer-based data sorting, categorizing, and visualizing for students in kindergarten through grade two and for grades three through five (CS K–2.DA.8, K–2.DA.9, 3–5.DA.8). Working toward these standards is important preparation for using data software in middle and high school to visualize and interpret large data sets. Experiences with different types of visualizations will further expand students' sense-making opportunities and encourage them to think about what they can understand by looking at data sets in different ways (SMP.4). Newspapers and online news sources offer specific examples; student-gathered examples help build buy-in for a "Can we figure out what this visualization is trying to help us understand?" routine.

Data Collection, Sampling, and Random Processes

Remaining alert to students wondering about their everyday experiences—perhaps in attendance, weather, or lunch-count data—may generate opportunities for the class to explore how the collection of data can help answer questions asked by the class or the teacher. In addition to their own observations, students should gain exposure to designing and using surveys and simple experiments as ways to collect data. By producing their own data from their classroom or community ("How does the age of students relate to their enjoyment of school? Does time on social media apps increase with age? How much waste is generated by different companies or our school?") students recognize data as having context and deriving from observation and measurement, and they come to see how mathematics and data are tools to help think about their worlds (SMP.4).

As students seek data to address authentic questions similar to those described above, they should also encounter opportunities to help determine how data might be produced and to consider how their choices might impact the data. When conducting classroom surveys, students can begin to grapple with very basic ideas of fairness, laying a foundation for sampling. For example, "Is it fair to interview only my friends?" or "Should I measure only my tallest seedling?" Also, they should consider their own measurement techniques and how confident they are that all of the students measured the same way (so that if someone else measured, such as for height or amount of sunlight, they would get the same results) (SMP.6). Students should also encounter data collected by other people for a similar purpose.

Students at the elementary level often express informal wonderings about probability, randomness, and uncertainty. For example, “How likely is it that it rains when we have recess?” or “Can we predict who will come through the door next or what color cube we will draw out of a bag?” Randomness is a complex idea encompassing uncertainty and a level of predictability. When (blindly) drawing a cube out of a bag containing three blue cubes, two red cubes, and one yellow cube, nobody can predict with certainty what will happen on a single draw. But, over many draws, the person who always predicts a blue cube will be right about half the time. Activities that demonstrate this concept can be used to generate data for many of the explorations of the thematic topics above, which will leave students well-prepared for a more formal treatment of randomness and probability in middle school. At this point, students should begin to conceive of probability as a general measure—e.g., not likely, likely, very likely chance that something will happen—and should see it as a basic measure of certainty or uncertainty.

Interpreting data is a matter of making inferences from the data available. Although students will encounter quantitative and nuanced techniques for making inferences in later grades, they should nevertheless encounter opportunities to make claims and infer conclusions across their kindergarten through grade five years (SMP.3). When they do, students should learn to wonder whether patterns or trends they notice extend to larger populations (including considering ways in which a group might not be representative of the larger population). Additionally, students should learn that good claims draw upon data as evidence and that they always come hand in hand with a degree of uncertainty. Modeling the use of appropriate terminology such as “tends to,” “typical,” “usually,” and “similar” can help lay important groundwork for this concept (Rubin 2019).

Upper elementary students begin to reason more abstractly and work toward using all four operations to solve problems. The classroom, home, and community present meaningful opportunities for students to apply tools to measure and describe the world around them, including collecting data for one or more attributes. For example, students may be challenged to discover which location (inside or outside the classroom) has the “best” types of tables for collaborative group work. As a part of this task, students explore the idea of “best” and discuss features such as size of the tabletop, height of the table, and shape. The teacher guides the students to consider which attributes of the table could be measured and then provides a template for student pairs to collect their observations and measurements—e.g., of width, height, and shape of the table. After collecting data, students notice that some of the table shapes were hard to measure because of their unexpected shape—e.g., trapezoids, kidney beans, and circles.

The teacher guides a discussion by asking students what was fun, easy, or hard about the data collection process. While students share, the teacher tracks some of the challenges and prompts students to brainstorm reasons they encountered the challenge, as shown in figure 5.5.

Figure 5.5: Tracking Challenges and Reasons for Challenges

Challenge	Reasons for the Challenge
We measured the same type of table but found different widths or heights.	• "Maybe we measured wrong." • "We measured different units." • "We didn't round up or down in the same way." • "Some of us used 1/2s, like 5.5."
We only measured certain tables.	• "We didn't know what to do for circle or trapezoid tables." • "The shape was weird (irregular)." • "We ignored certain tables." • "Tables are packed away in the cafeteria or were being used in the library so we couldn't measure them."

The process of collecting and working with data helps students recognize that data collection is a human endeavor and includes decisions and sometimes errors made by people. Activities such as the tabletop area exercise help students begin to develop a list of features important when designing surveys or experiments:

- Is it fair to choose the biggest tables in one room and then only the smallest tables to measure in the next (randomness)?
- What happens when someone designs a data collection method but it does not work in the real world—for example, the large tables in the cafeteria were stored away?
- How many tables (number of cases or sample size) should be measured?

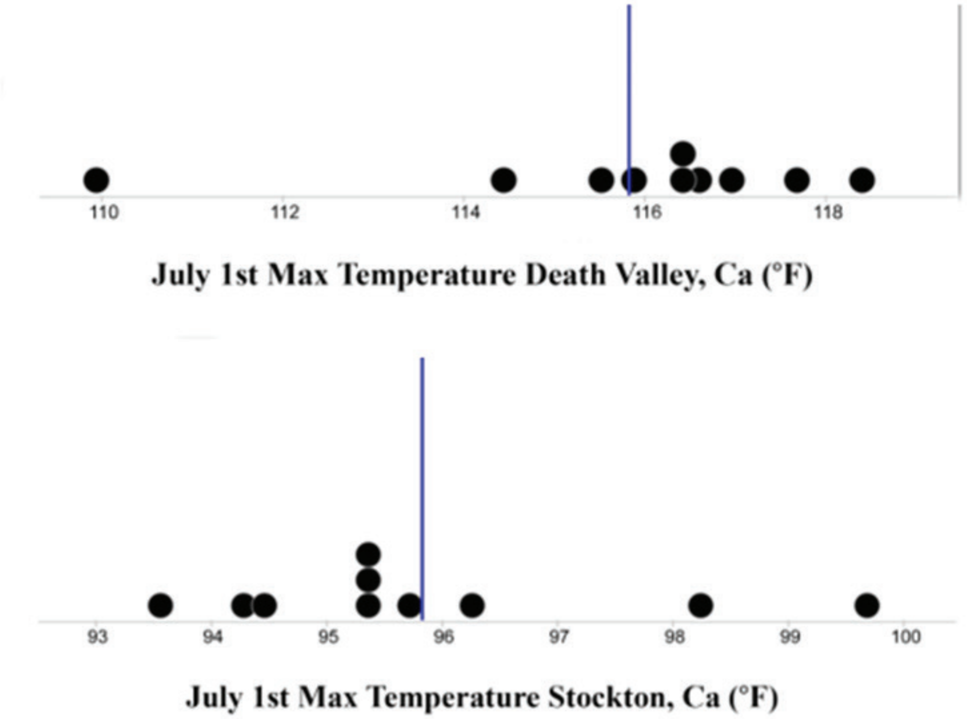
Statistical investigations that expose students to data that are sometimes messy or involve a process that can be ambiguous (SMP.1) are crucial to data science. Students can also come to recognize that they can ask questions about the data that are given to them—e.g., questions about how or when the data were collected and for what purposes. As students ask questions, the data required to answer them may not be accessible or possible to collect firsthand. Data gathered by others (such as other students in the discussion) can help answer questions students generate about their own communities and can open discussion about randomness and probability.

Comparing Distributions and Identifying Associations Between Variables

In kindergarten through fifth grade, students are invited to ask questions about the world around them, especially about objects inside the classroom. When students compare measurements and frequently reported values between groups, they are engaged in tasks that lay the groundwork for more sophisticated comparisons. Invitations for students to describe their data and make predictions about why the values might vary are important early opportunities. Sentence starters, tally tables, dot plots, line plots, pie charts, and bar graphs are all important tools that help students describe patterns within data, especially when making comparisons between groups.

Notably, many questions that students might wonder about in science and other fields will not be fully answerable using the tools and mathematical understanding available to them in kindergarten through grade five. It is important that teachers have resources for helping students figure out which aspects of questions can be investigated with currently available tools and that teachers have some understanding of technological tools that students will encounter later. For example, many students will wonder about relationships between two different variables: “If I get up earlier, do I feel tired earlier in the afternoon at school? Do students who skip lunch eat more candy in the afternoon?” When one of the variables is categorical (like for the skipping lunch question), separate line plots can be made for each category and the line plots compared. When both variables are quantitative, students can input data into an online graphing and visualization tool such as the Common Online Data Analysis Platform (CODAP), Desmos, or TinkerPlots, and then investigate the relationships by plotting their data on graphs, observing their distributions, and adding line plots. Another option is that one of the variables can be made into a categorical variable by defining categories in terms of the quantitative variable. For instance, waking-up times could be classified into “early” and “late” (ideally with a student-generated cut-point between early and late) and then dot plots of “time in the evening when I felt tired” created for each category. Science investigations can provide opportunities for students to compare mean values between two locations or experimental conditions, as shown in figure 5.6. As in the figure, the use of stickers to create line plots with different symbols or graphing programs can help with making plots quickly and easily. Note that dot plots are not formally introduced until sixth grade.

Figure 5.6: Temperature Plots to Compare Mean Values for Two Cities in California



Source: Generated with CODAP NOAA Plugin

Long description of figure 5.6

As students are invited to ask questions through data and measurement, teachers should be mindful of the types of comparisons being generated. Questions such as "What time will it be when the next person walks into the classroom?" or "Which book in the classroom is the most read?" compare events or objects within a shared space and are generally preferred. Questions and data collection tied to personal characteristics ("Who is the shortest in our class?") or that serve as potential markers for economic or social status ("What brand of shoes is the most popular in our class?") usually should be avoided.

Grades three through five provide opportunities to investigate questions using data that should include volume and mass measurement (grams, kilograms, and liters, but not compound units such as cm^3) in addition to the length, time, and money contexts from earlier grades (3.MD.2). Time measurements are refined to the nearest minute (3.MD.1) and length now includes half- and quarter-inches (3.MD.4). An increased ability to report lengths more precisely helps students begin to notice that some data can fall into specific counts (discrete) while other types of data (measuring length of objects in millimeters) are continuous. A significant context for data-investigation questions is classification and analysis of two-dimensional shapes (4.G.2). Incorporating this geometry standard to help build data understanding can foster the important practice of analyzing by attributes—one instance of SMP.7 (Look for and make use of structure).

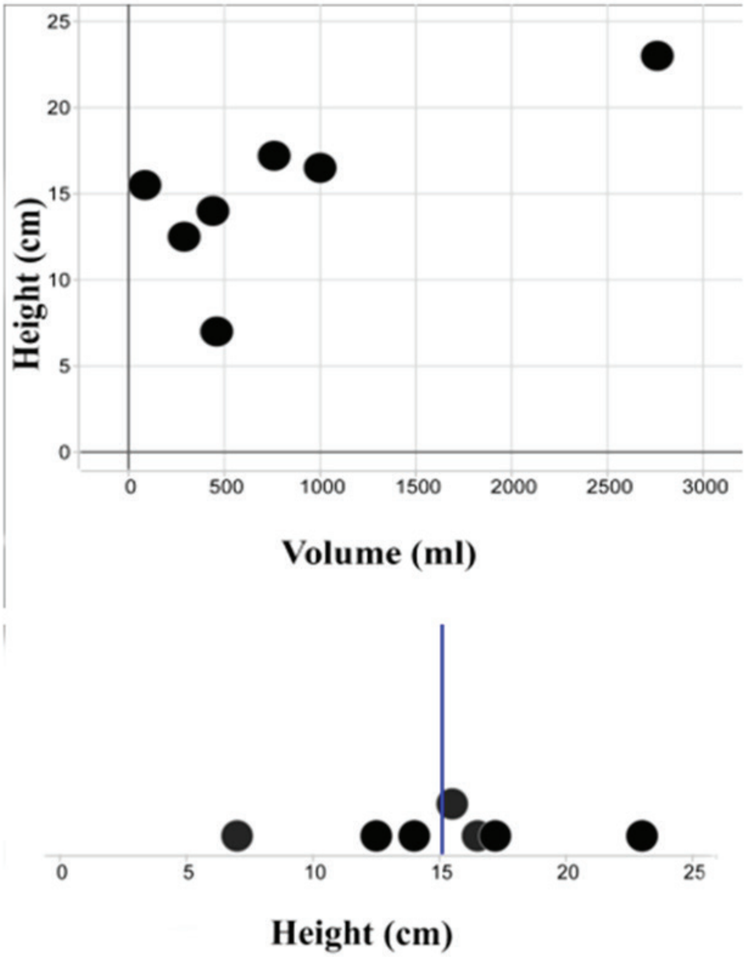
In this grade band, students extend the set of units they work with (4.MD.1) and can generate data about area for more complex shapes. Fifth graders deepen their understanding of volume to include unit cubes, making this an important context for data-inquiry questions. For example, a teacher could invite students to build a structure out of multilink cubes and then collect data from the class by inquiring how many cubes they used, the height and width of their structures, or which colors they used. Invitations to collect data on multiple variables produce data sets that allow students to compare measures across plots, such as comparing the average amount of time it takes for students to walk to school versus drive to school.

The snapshot of Logan's third- and fourth-grade explorations with data describes two possible encounters a student in this grade band might have with the thematic topics (*understanding and describing variability in data; data collection, sampling, and random processes; and comparing distributions and identifying associations between variables*) and are highlighted in parentheses where relevant.

Snapshot: Logan's Third- and Fourth- Grade Explorations with Data

In third grade, as mass and volume became characteristics to measure, Logan's class used length, height, mass, and volume measurements they had collected to examine sets of objects (*data collection*). In the science corner, the line plots of the masses looked quite different from the line plots of the lengths/heights of the objects, as did the line plots of volume, height, and mass of all objects in the room that hold water (vases, cups, etc.) (*understanding and describing variability in data and data distributions*; see figure 5.7). Logan's team had a great disagreement about whether a taller vase should hold more water than a shorter vase (*comparing distributions and identifying associations between variables*); the class eventually decided that it was usually but not always true that taller vases hold more water.

Figure 5.7: Logan's Vase Measurement Data Visualized in CODAP

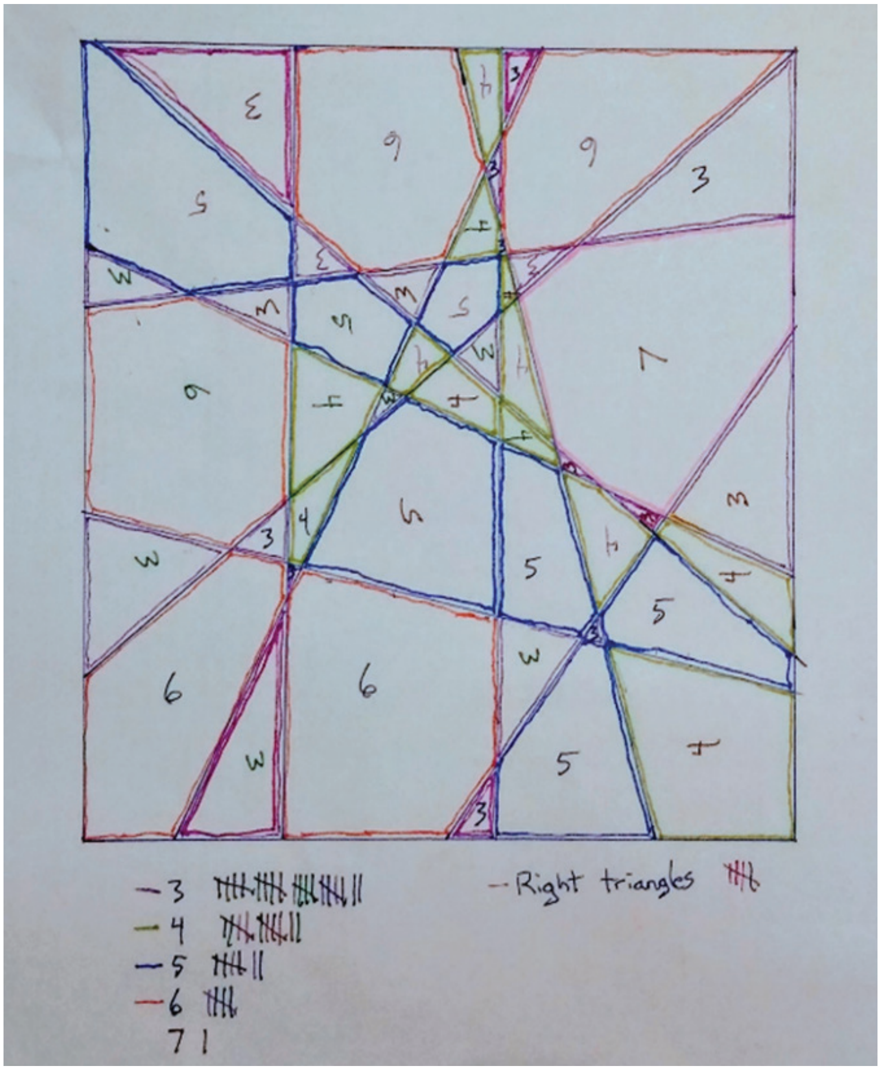


Source: Collected student data in CODAP

Long description of figure 5.7

One of Logan's favorite activities in fourth grade was one that combined data work with classifying shapes by attributes: creating collaborative art pieces. For this activity, each team had a 1/2-meter by 1/2-meter square on the board, and each student in the team drew in two edge-to-edge straight lines of their choice, using their meter sticks. Then one student in class chose a shape to try to find in the drawings, and each team outlined each new instance of that shape they found and described how they knew it was a triangle, rectangle, right triangle, quadrilateral, etc.; this process was repeated for several other shapes. Team members collected data by making an individual card to represent each piece of artwork (as shown in figure 5.8), using the card to represent the different variables they measured for each piece (how many triangles, how many instances of each color, how clean or messy each line was, etc.). When they had made a full set of cards, they sorted them in various ways, then made a table to compare the tallies for the different pieces (*understanding and describing variability in data and data distributions*), discussing the different features of the art and the process of creating it that might help explain the variations in their data.

Figure 5.8: Using Data to Classify Shapes



Long description of figure 5.8

Grades Six Through Eight

As in earlier grades, students in grades six through eight can understand their world via a process that begins with wondering questions. This grade span is also the beginning of when students experience the mathematical modeling cycle (Pelesko 2015) and investigations in science (NGSS Lead States 2013). In middle school, students develop a formal understanding of several key ideas in statistics, including describing distributions and variability in data and random processes. Students begin informally comparing and identifying associations in eighth grade in preparation for work in high school that develops a more formal understanding of linear models and statistical tests.

At the middle school level, students should encounter data sets that are small (a few to a dozen to a few hundred data points) and, when possible, can encounter larger data sets that contain thousands of data points. Many statistical concepts are more intuitive and accessible when illustrated with large data sets.

The following sections provide examples of how the same concepts can be illustrated with "little data" versus "big data." Working with bigger data sets requires the use of computational tools, some of which may require programming skills. California has adopted K–12 computer science standards which can be consulted to determine what level programming is appropriate for middle school and high school (California State Board of Education 2022). Working with complex data sets provides students with opportunities to engage in multivariate explorations, conduct simulations, and quickly create plots to reveal patterns—all nearly impossible to do by hand. Knowing when and how to leverage the power of computational tools is a crucial skill in data science. Students and teachers will need additional support with selecting and using these tools.

Understanding and Describing Variability in Data and Data Distributions

Sixth-grade students build on earlier experiences by distinguishing between statistical questions that can be investigated using data that varies (e.g., analysis of social media usage by age of students) versus questions for which there are no variations in (correct) responses (How many days are there in January?) (6.SP.1). When considering a statistical question, they understand that the variation in numerical data has a distribution which can be described by its center (first the median, then the mean); by its variability (also called spread, which is described both qualitatively and via a numerical measure—either interquartile range [IQR], range, or mean absolute deviation); and by an overall shape (including descriptors such as symmetric, skewed left or right, peak, gap, and outlier) (6.SP.2, 6.SP.3). As students explore data sets, they can produce visual representations of the distributions of their data; they can look at the shape of distributions that have different measures of center and spread and can develop visual understandings of the shape of distributions. In sixth grade, visual representations of distributions include box plots and histograms, adding to the line plots (called dot plots from grade six onward) from earlier grades (6.SP.4). In addition, students learn to report and interpret measures of center and variability, and descriptions of distributions, in the context in which the data arose (6.SP.5.d).

Beginning in sixth grade, students should have experiences deciding which measure of center is a more useful descriptor of a typical value for data sets with different shapes. Because the mean is sensitive to extreme values, the median is often a more useful measure for skewed distributions; in this case, the interquartile range is a useful measure of variability. For some distributions—with multiple clusters, for example—students may decide that neither median nor mean is a useful measure and might decide that a single number cannot reasonably represent a typical value (6.SP.5).

The following snapshot illustrates several themes discussed in this section, with thematic topics included in parentheses where relevant.

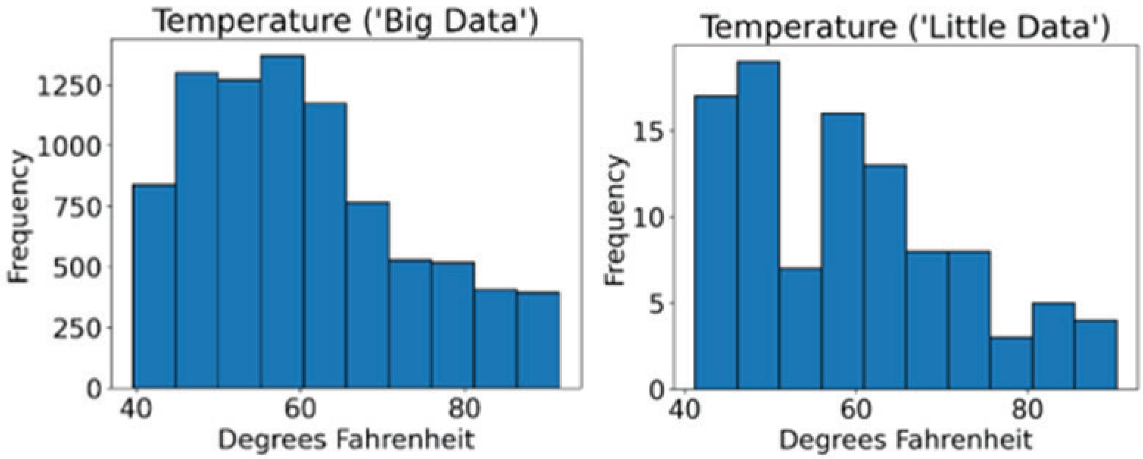
Snapshot: Óscar's Visual Proof for Finding a Mean

Óscar did not enjoy learning about mean, median, and mode. He often confused the different measures and felt they had little meaning. His parent contacted Maria, his teacher, to let her know that Óscar had been expressing frustration about the meaning of the terms since his last assessment. Óscar was not alone; Maria knew that many of the students were still struggling with the meanings of these measures of average. Based on results from an electronic, anonymous "exit ticket" survey used as formative assessment, Maria approached the students with the idea to build physical models so they could experience the averages in visual and physical ways, encouraging important brain connections.

Maria gave her students cubes and asked them to make six different towers of cubes that represented the numbers 1, 6, 3, 2, 4, and 2. She asked them how they might construct a physical proof to show the mean of the numbers. Some of the students were able to calculate the answer; however, she kept pushing them to build a visual proof while remaining open to multiple means of representation. This strategy, based on specific UDL guidelines, allowed Maria to ensure scaffolds and supports would exist to help highlight the patterns of language and draw on background knowledge to express what students know in ways that are authentic and meaningful. Óscar and his group members came up with the idea of moving the cubes from tower to tower to show that they could make six towers that were all the same height. They just needed to average out all the blocks (*understanding and describing variability in data and data distributions*). Óscar and his group excitedly explained to the class how they had made a physical proof of finding the mean of the blocks (*understanding and describing variability in data and data distributions*). They shared the calculation with the class and compared it to the method they used of moving the blocks. After her students had discussed finding mean, Maria asked them to make a visual proof for the median and the mode.

A key characteristic of data science is asking questions of “big data”—a data set that has many cases and variables and needs to be manipulated and analyzed computationally. One benefit of working with bigger data sets is that discerning the shape of a distribution is often easier. Both histograms in figure 5.9 show distributions of air temperature measurements taken in Sacramento, CA, in 2013. The table on the left shows 8,700 measurements taken hourly from January 1 to December 31, 2013, while the table on the right contains 100 measurements randomly sampled from the larger data set on the left. The “big data” distribution on the left shows a fairly smooth distribution with a rightward skew and modal temperatures of approximately 55 degrees. The “little data” distribution on the right also shows a rightward skew but contains peaks and valleys, with modal temperatures of approximately 45 degrees.

Figure 5.9: Comparing Distributions for Large and Small Data Sets



Source: National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2023)

Long description of figure 5.9

The following snapshot describes a classroom scenario in which students investigate hurricane data from multiple years and use a range of data displays to understand the science of hurricanes and generate additional questions. Thematic topics are shown in parentheses within the body of the snapshot where relevant.

Snapshot: Quincey’s Investigation of Hurricane Data

The sixth-grade math teacher, Leonora, decided to have her students explore the “shape” of some weather data. The context is hurricanes in the Atlantic Ocean and uses real data collected from 5 years of hurricanes at successive 10-year intervals. One student, Quincey, showed real interest and engaged in the lesson’s opening discussion of 2017 hurricane data displayed on a line plot (*understanding and describing variability in data and data distributions*). Quincey and the class were really interested in the number of hurricanes that were in the tropical storm category.

Next, students worked in groups to study hurricane category data for the years 1977, 1987, 1997, and 2007 (*data collection, sampling, and random processes*). Each decade’s data were presented in different ways: bar graphs, line plots, tables, and sentences. Quincey enjoyed the analysis and was taken with the different ways of displaying data as well as the changes in the spread of data from decade to decade.

Quincey asked important questions about the science of hurricanes. “How do they develop? What makes them get larger? What is the difference between a Category 3 storm and a Category 5 storm?” At the close of the lesson, Leonora was convinced that students understood that different visual displays of data can make it easier to recognize how a situation might be changing over time (*comparing distributions and identifying associations between variables*). The class reflected that the changes were easier to see in line plots and histograms than through the data being shared in writing or in a table of values. Quincey decided to further investigate the number of Category 4 and 5 hurricanes over the past 100 years and how these storms become stronger, and Quincey set out to gather more data and ask questions of the data. Others in the class decided to investigate why the number of Category 4 and 5 storms is increasing.

Data Collection, Sampling, and Random Processes

Sampling

Prior to seventh grade, students’ work with data focused exclusively on using data to understand, describe, and compare the particular collection of objects or situations that have been collected by observations, experiments, or measurements.

Seventh grade includes the first introduction to sampling, the process of collecting data from a subset of a population in an attempt to understand or describe the whole population. This focus represents a big jump in sophistication from earlier work. As an example, suppose all students who come in to play basketball before school are asked to track their screen usage for the week. The class analyzes the data and determines that those sampled spent an average of 862 minutes on the screen. Small- and whole-group discussions invite students to consider whether this sample can extend to (i.e., is representative of) the entire student population at the school or to all students who are the same age. Although the 862 minutes may be the typical screen time for the defined group of students where the data were collected, it may not extend to everyone. Explorations and discussions, including considering some obviously nonrepresentative samples, can help students understand the idea of a random sample.

It is important for students to have multiple experiences selecting samples from known populations in ways that are random (for instance, drawing numbered ping-pong balls from an opaque bag or drawing student names on identical slips of paper from a hat) and in ways that are not random (for instance, asking survey questions only of the students who sit near you in class). The goal is for students to develop an understanding that random sampling tends to produce samples that are representative of the population—that is, their distribution of the quantities under consideration are close to the distribution for the population as a whole (7.SP.1)—and for students to have a sense of the variability when using samples to make inferences and estimates for a population (7.SP.2). Many computational tools enable students to quickly draw samples from data sets using a variety of methods (e.g., randomly, selecting every fifth record or the first 100 records).

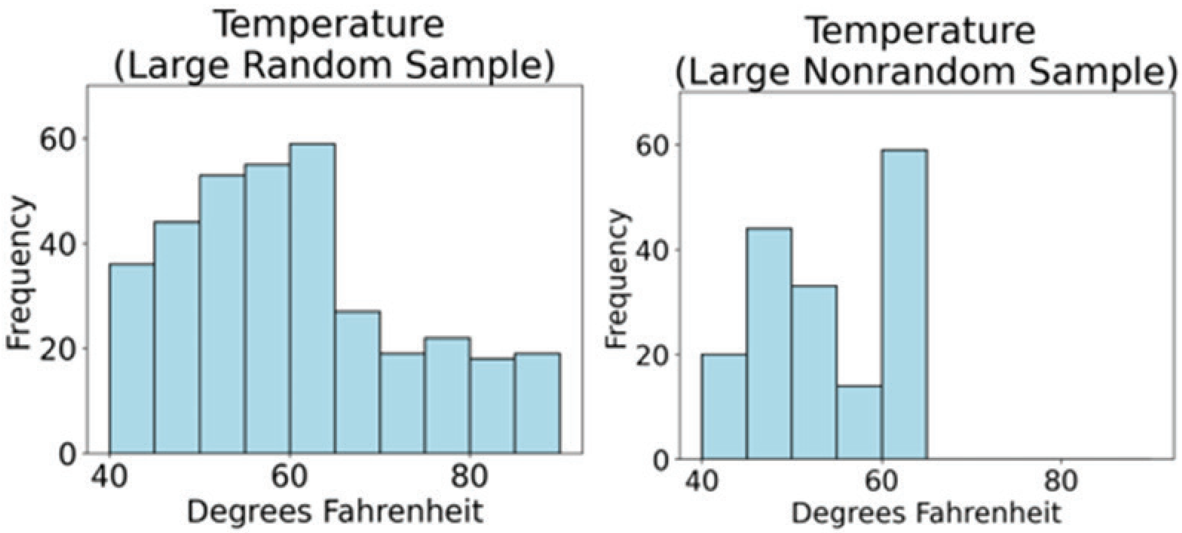
Although sampling is not explicitly named again in the standards until high school, eighth grade students may benefit from additional opportunities to deepen their understanding of sampling as they work with bivariate data. Random sampling can become a tool to engage in data explorations of interest to students—e.g., "I wonder how long on average it takes students from different grades to get from home to school?" or "How much food is wasted in the lunchroom every month?"

Nonrandom sampling (such as attempting to understand the school as a whole by collecting data only from one's friends, or by asking about eating habits at the gym after school) produces biased conclusions, even when the bias in the sample selection might not be obviously linked to the quantity being measured in the measurement or observation. Bias in the statistical setting does not refer to temperament or outlook (prejudice), which is one meaning of the word; instead, it means a systematic error.

Students often believe that arbitrary sampling schemes ("The first 10 students I meet" or "Every tenth student alphabetically") are random; they need to understand the difference between these schemes and choosing by chance so that every possible sample has an equal likelihood of being selected.

Figure 5.10 has two samples drawn from the Sacramento temperature data first shown in figure 5.9 above. The table on the left shows a random sample of 365 measurements from the original data set, while the table on the right shows a nonrandom sample of 365 measurements: the first measurement of the day for every day in 2013. The shape of the distribution on the left is much closer to the shape of the original distribution that contained 8,700 measurements (figure 5.9).

Figure 5.10: Comparing Random and Nonrandom Samples



Source: National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2023)

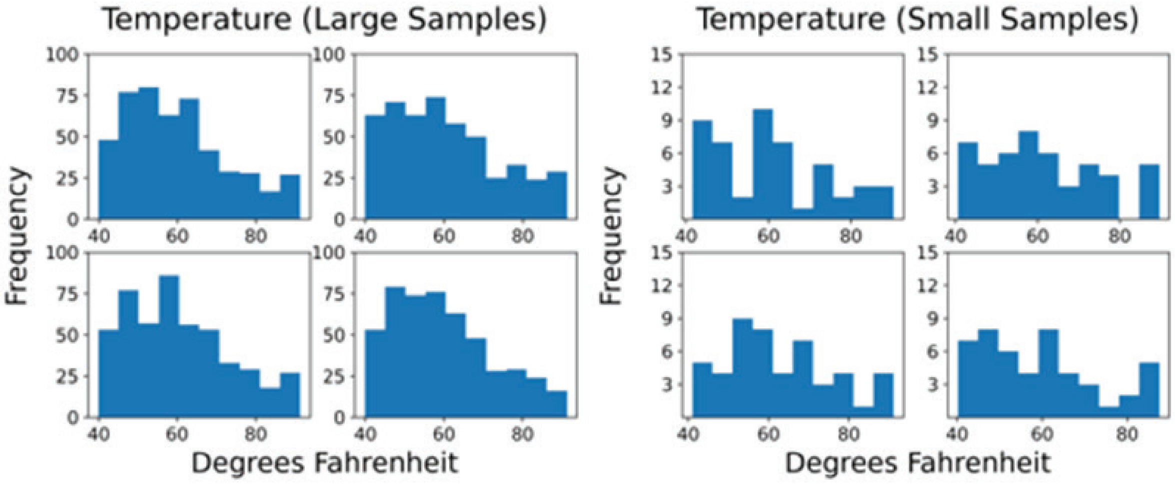
Long description of figure 5.10

Probability and Random Processes

Randomly selecting from a population and measuring a characteristic (in which variation is expected across the population) is a chance process: It may result in different results and its outcomes follow some distribution.

Although students can generate samples through non-computational means, computational tools can enable students to quickly and easily draw samples from a data set and visualize or summarize each sample in order to compare and contrast the results of different sampling methods. The tables on the left in Figure 5.11 show four different random samples of 500 data points from the same Sacramento temperature data shown in figure 5.9, and tables on the right show four different random but much smaller samples of 50 data points from this data set. Students should notice that the shapes of the small samples are much more variable compared to the shapes of the large samples, even though all samples were generated randomly. A small sample, even if random, is less likely to be representative of the population than is a large random sample.

Figure 5.11: Comparing Distributions for Large and Small Random Samples



Source: National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2023)

Long description of figure 5.11

Probability expresses the chance of an outcome as a number between 0 and 1 (7.SP.5). Probability is combined with statistics in the grade-seven standards.

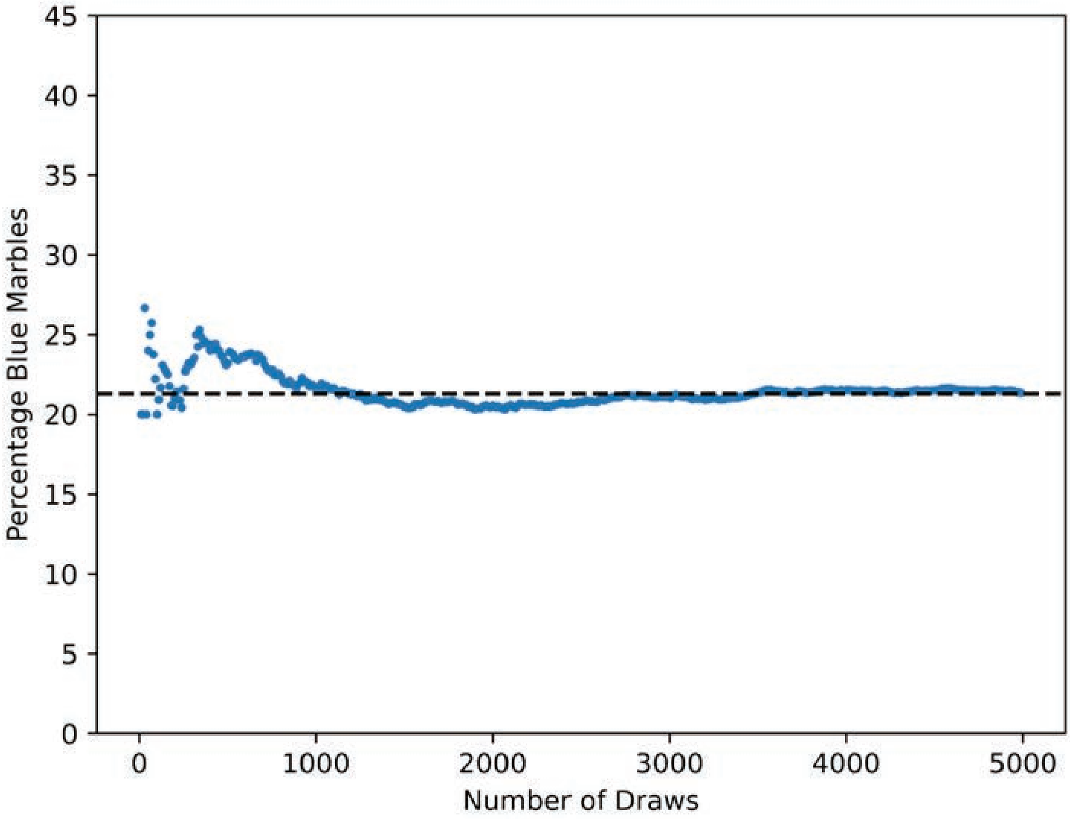
Statistics and probability are historically linked because statistical claims and estimates are based on the mathematical field of probability. Using sampled data to predict events, such as elections outcomes, is based on probabilistic reasoning.

Students sometimes struggle to see clear connections between probability and statistics, especially when their experiences focus on procedures and calculations rather than exploration, context, and interpretation. Statistics produces estimates for parameters in probabilistic models. There is much work with probability that does not support statistical reasoning and may not be applicable to a setting of interest (e.g., calculating theoretical probabilities for the sum of two dice without using those theoretical probabilities to decide whether a given pair of dice are likely fair), and middle school probability experiences should be carefully designed to support reasoning with interesting and meaningful data.

In seventh grade, students gather data to estimate the probability of outcomes by observing their long-run relative frequency; that is, they compute experimental probability. Consider repeating this experiment 150 times: Draw a marble from a bag with marbles in it, record its color, then put the marble back in the bag. If you get a blue marble 32 times, your estimate for the probability of getting a blue marble on any particular draw is 32/150 (7.SP.6, 7.SP.7.b). This is really an estimate of the fraction of blue marbles in the bag.

Compare the marble experiment just described to another, placing the following marbles in a bag (all identical except for color): 16 blue marbles, 31 red marbles, 16 green marbles, and 12 white marbles (75 total marbles). If you blindly pull a marble from the bag, what is the probability that you will get a blue marble? If you repeat this 150 times (putting the marble back each time), about how many times do you expect to get a blue marble? Students can create a theoretical probability model and calculate the expected frequency by multiplying the probability of drawing a blue marble in one draw by the number of draws (16/75 × 150 = 32). After calculating this expectation that students will draw blue marbles 32 out of 150 times, or 21.3 percent of the time, students might try to verify their theoretical probability model and create a simulation that can pull a marble from the bag 150 or 1500 or 15,000 times. Students can then compare the simulated frequency of drawing a blue marble with their theoretical expectations (CSS 6–8.AP.10). Figure 5.12 shows the results from such a simulation. The long-run proportion of blue marbles reaches an asymptote at the theoretical probability of 21.3 percent.

Figure 5.12: Results from a Simulation Containing 5,000 Trials of the Marble Experiment



Note the difference between the two different marble experiments described in the previous two paragraphs. In the first, students repeat an experiment many times and use the long-run frequency of drawing a blue marble to estimate the theoretical probability of drawing a blue marble from the bag. In the second, students build a (theoretical) probability model and use it to estimate the long-run frequency of drawing a blue marble from the bag (7.SP.7). If the relative frequencies of experimental outcomes do not seem close to predictions from the probability model, then students need to be able to discuss possible sources of discrepancy

(7.SP.7): Perhaps the green marbles have a different texture and tend to be drawn more frequently than predicted. Maybe somebody changed the mix of marbles in the bag. Or perhaps not enough draws were performed to see the relative frequencies approach the probability model.

Finally, seventh grade students find probabilities of compound events (events which are made up of several simple events)—for example, drawing two marbles from the bag of 75 described above and getting one white (W) and one blue (B) marble (7.SP.8).

The recognition that some events (repeat the draw five times, get all blue; or repeat the draw five times, obtain W-B-W-W-B in that order) are much less likely than others (repeat the draw five times, get three white and two blue) is key to understanding claims made from statistics.

In fact, most statistical claims depend on a comparison of a (theoretical and hypothetical) probability model with observed data, as in 7.SP.7. To prepare middle school students for future statistical work, teachers should offer experiences that develop an awareness that more data tend to produce relative frequencies closer to actual probabilities. Computational tools can support activities that use larger data sets and the creation of simulations that enable students to compare experimental and theoretical probabilities.

The following snapshot describes how a student in this grade band might encounter ideas from this section. Major themes are highlighted in parentheses where relevant.

Snapshot: Rosa's Students Experience Random Sampling

Understanding the ways Rosa's seventh-grade students have responded to the probability activities offered through her instruction has influenced the next steps in her planning. Overall, Rosa has not been satisfied with the students' understanding of random sampling. She decides to give students a more visual and physical experience of the concept. Her plan calls for six paper bags filled with differently colored cubes. The sum of cubes and the color distribution of the cubes in the bags are as follows:

Bag One, 15 total: 15 blue

Bag Two, 12 total: 11 blue, 1 red

Bag Three, 20 total: 15 blue, 4 yellow, 1 red

Bag Four, 10 total: 5 red, 5 yellow

Bag Five, 12 total: 5 blue, 4 red, 3 yellow

Bag Six, 20 total: 8 blue, 8 red, 4 yellow

Rosa explains the task by telling students they will determine the contents of each bag through sampling. She chooses not to tell them how many times to sample but she does tell them to sample from the bags by selecting one cube at a time and then putting it back into the bag. Rosa also asks students to determine the chance of drawing a blue cube from each bag (*data collection, sampling, and random processes*).

Students engage in the activity, brainstorming methods for collecting and recording their information. When each group of students feels satisfied with their determinations of the number of cubes and color distributions of the contents of each bag, she asks them to choose which bag belongs to which card showing the contents of each bag (*understanding and describing variability in data and data distributions*). In setting up the lesson, Rosa filled the bags differently and made sure to have a bag for which the probability of drawing a blue cube would be 1 and another for which it would be 0. After the activity and class discussion, Rosa is happy to hear her students later talking about situations in which the probability is 1 or 0 and other situations representing everything in between. Her students recognized the number of times they sampled usually led to better predictions about the contents of the bags. They also realized that sampling without replacement would have shown them the exact contents of the bag. The class engaged in a rich conversation about sampling with and without replacement, recognizing that it would be unproductive to draw all the cubes if there were a million.

Comparing Distributions and Identifying Associations Between Variables

Prior to grade seven, students typically work with a single collection of data that measures a single variable. In grade seven, they compare the same variable measured across two populations, either by actually measuring the whole populations or obtaining estimates for the population distributions via sampling. They can plot data and draw from different statistical methods such as creating box plots and dot plots to informally assess the degree of overlap of two populations.

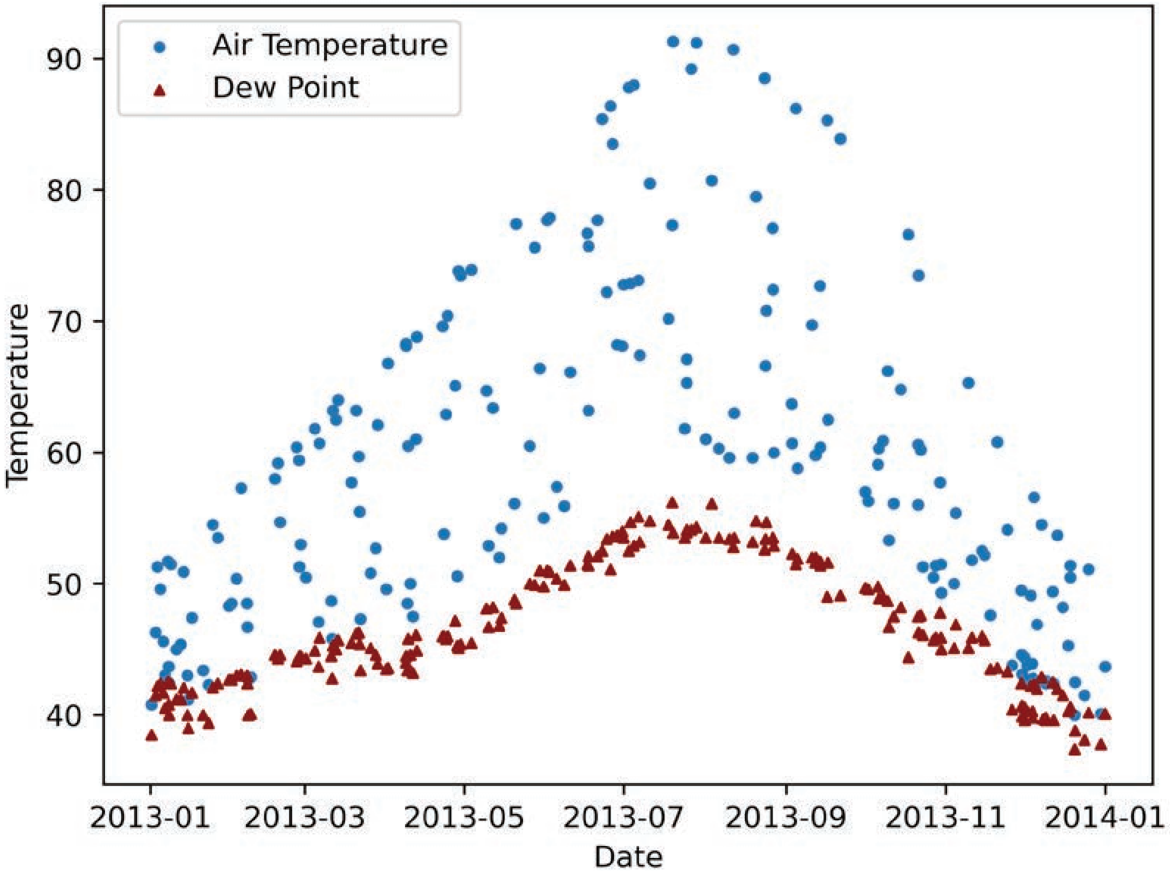
In eighth grade, the focus is on formally describing bivariate data: two quantities or categorical variables measured or observed across a population or across a sample drawn from a population (8.SP.1). Eighth-grade students use two-way frequency tables as tools to see associations in bivariate categorical data (8.SP.4). This work has important connections to linear equations and modeling.

The scatter plot as a visual representation of quantitative bivariate data is one of the most important ideas introduced in eighth grade. A survey of students collecting data on both time and distance for traveling from home to school might reveal clusters, outliers, and any of various types of association (positive, negative, linear, nonlinear). Students should describe such patterns in a scatter plot and interpret the patterns in the context of the data (8.SP.1). Once a scatter plot is created, an association between the two variables may become visually identifiable. Fitting a function to the data is the creation of a mathematical model for the association.

In eighth grade, students choose a line to fit the data by visual approximation on the scatter plot, and they compare and argue for whose line fits "best" (8.SP.2). They then interpret the meaning of the slope and intercept of their chosen model line and use the line to make predictions for one variable when the other variable is specified (8.SP.3). Finally, eighth-grade students use two-way frequency tables as tools to see associations in bivariate categorical data (8.SP.4).

Although the type of function that is used most frequently is a line (a linear function), students also need experiences plotting associations that are clearly nonlinear, as in figure 5.13, and fitting other types of functions (quadratic, exponential) to the plot.

Figure 5.13: Scatterplot Showing a Nonlinear Association Between Air Temperature and Dew Point in Sacramento in 2013



Source: National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2023)

Any standard data software (including spreadsheets, Desmos, Geogebra, CODAP) will fit lines, quadratic functions, and exponential functions to given data. Students are not expected to know the specific standard technique for identifying a line (or quadratic or exponential function) of best fit (least squares regression), but students should have experiences fitting lines and some other functions visually (by adjusting parameters on appropriate function types in graphing software) and using appropriate software tools which perform the regression calculations.

High School

Students' prior work learning about describing and comparing distributions and random sampling comes together in high school. High school students continue to visualize and represent univariate data with dot plots, histograms, and box plots; use measures of center and spread to describe such distributions (S-ID.4); and compare distributions from different populations or samples using these representations and statistics (S-ID.1–3). A major difference between students' data experiences in kindergarten through grade eight and what is explored in high school is the richness and complexity of available data sets, even more so than their sheer size.

As high school students work with these data sets, they can draw upon the statistical understandings they have developed in their mathematics lessons from kindergarten through grade eight. Instruction should emphasize opportunities for questioning and interpreting alongside technical procedures.

Data exploration begins with a search for available data about a context of interest. The data set is then examined for hidden patterns and associations. At the high school level, visualization of data can illustrate unexpected structure. Any patterns or associations discovered can lead to new hypotheses or questions to investigate further. Students began this process in eighth grade and continue in high school with experiences in which they examine data sets with multiple variables that are measured for each member of the sample. They plot pairs of variables to decide which ones might show associations. Important discussions for students to engage in when working with existing data sets include the following:

- Prior to exploring: Do you expect any of these variables to be associated? Why?
- Might the association you see just be a result of the way in which the data were collected rather than truly reflective of the population? What features of the data collection might make conclusions suspect, and what features might give confidence? Note that a large sample size is not enough to have confidence in conclusions.
- Can you think of possible explanations for the association(s) you see? Can you think of ways you could decide which explanations might be accurate?

After data exploration identifies some association(s) of interest, the stage of model building follows. Technical methods are reserved for the specialized statistics or data science course, which is described below, but all students need to explore questions such as the following:

- Could you use some variables to predict others? Doing so is a hugely important use of data because some factors are easier to measure or observe than others. Medicine and many other fields often require using presently available information to try to predict future outcomes.

Most importantly, high school students (like kindergarten through grade-eight students) must experience statistics as a set of tools for making sense of their worlds in ways that matter to them.

Bringing It All Together: Introduction to Inferential Statistics

High school students begin learning about inferential statistics, which aim to generalize from a sample and draw conclusions about a population (S-IC.1). Students' work with inferential statistics is foundational to using data to make decisions. Students must decide whether a result observed through data is consistent with a mathematical model of the process that generates the data (S-IC.2). For instance, students are asked to engage in a thought experiment and consider how many households use gardens as a source of food. If a student hypothesizes that 30 percent of the students at the school grow food at home, the estimate offers a mathematical model that gives them an idea of what proportion to expect in a sample. If they then survey five randomly chosen students, and all say they grow food at home, then the student should be able to reason as follows: If 30 percent of students grow food at home, then the chances of five randomly chosen students all being among those 30 percent of students is $(.3)^5 = 0.00243 = 0.243$ percent, or less than a quarter of 1 percent. Thus, the student might doubt—that is, they might reject—the 30-percent hypothesis. Students should have many experiences of simple situations like this to understand how decisions based on data rely on probability and are not guaranteed to produce correct answers to the original question.

Students should work with data that originate from four different methods of data production, including at least some student-generated questions and student-gathered data. These methods are

1. generating census data, which are data that contain measurements on every member of the target population (such as the database of crimes occurring in a given city in a given time frame, or rain-gauge data for a given location, which captures all precipitation at that location—census data is first encountered in early elementary grades);
2. administering surveys to random samples (to estimate population values, or parameters, for the surveyed quantities);
3. conducting randomized experiments (to compare treatments and demonstrate cause); and
4. conducting observational studies (to study characteristics or quantities when random selection or assignment is not possible) (S-IC.3).

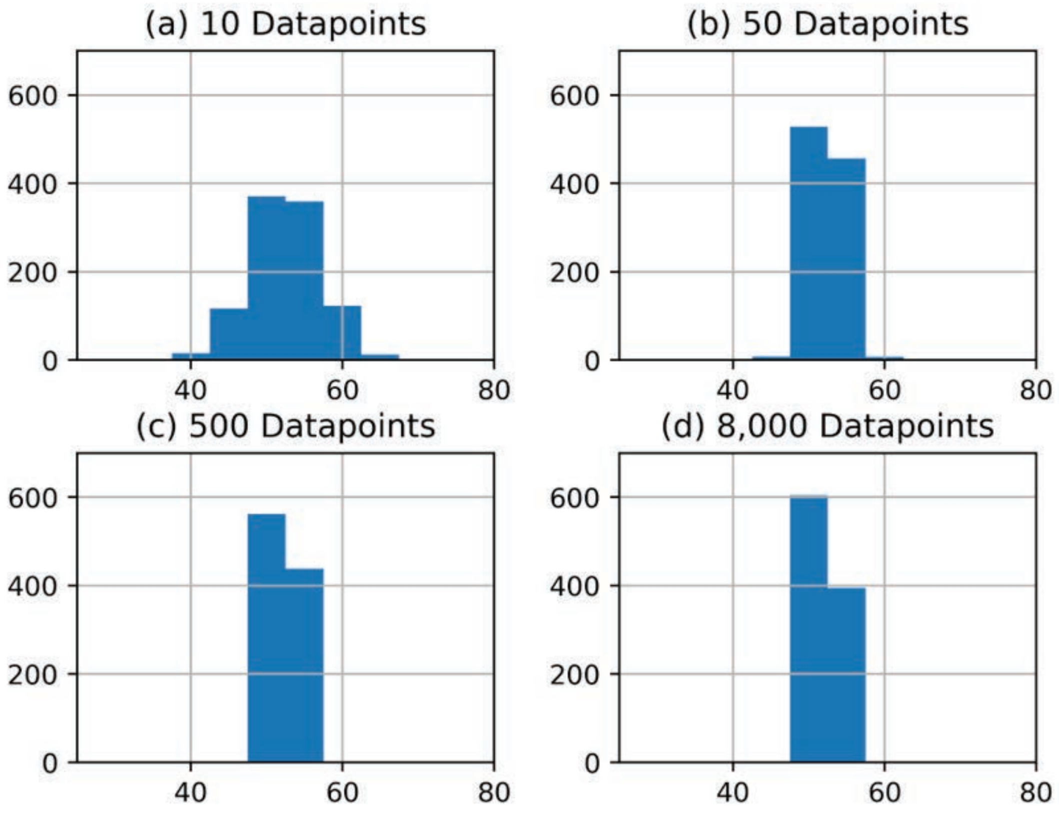
The "Statistics and Probability, High School" chapter of *Progressions for the Common Core State Standards for Mathematics* contains detailed examples describing the expectations in the standards (Common Core Standards Writing Team 2022).

Teaching with surveys and experiments must include a link between the random selection or assignment and the ability to reason probabilistically to make claims. With a survey, the random sampling allows generalizing to a population. With an experiment, the random assignment allows causal conclusions but not generalization to a broader population—unless the sample in the experiment was randomly selected from some larger population.

When using a sample mean or proportion to estimate a population mean or proportion, students use simulation models to estimate a margin of error, instead of using formulaic calculations. Briefly, the process is to use data simulation software to draw many random samples from a hypothetical population and to see how often a result is obtained that is as extreme as the sample mean or proportion. Doing this process for hypothetical populations with many different mean or proportion parameters helps students see there is a range of population parameters that often (more than 5 percent of the time) produce simulated sample means or proportions that are as extreme as (or more extreme than) the actual sample mean or proportion. This range of population parameters is the (simulation-based) confidence interval, given as a sample mean or proportion ± margin of error. Note the probabilistic argument here: If the population mean or proportion were outside of the confidence interval, then sample means or proportions as extreme as were obtained in the random sample would be rare. So, the true population mean or proportion is expected to be within the confidence interval (but one cannot be certain that it is!).

Figure 5.14 shows the results from a simulation that drew 1,000 random samples of different sizes from 10,000 normally distributed values with a mean of 50 and a standard deviation of 15. The first chart, "(a) 10 Datapoints," shows the distribution of sample means obtained from 1,000 random samples of 10 values. Although most sample means were close to 50, the sample means ranged from 38 to 67. The distribution of sample means becomes narrower as the samples get larger. Drawing 1,000 random samples of 50 values yields sample means between 43 and 62. Drawing 1,000 random samples of 500 values yields an even narrower range of sample means, from 48 to 57. And finally, drawing 1,000 random samples of 8,000 individuals only yields sample means that are very close to 50. What this exercise shows is that larger random samples are able to generate more precise estimates of population parameters, in this case the mean.

Figure 5.14: The Relationship Between Sample Size and the Shape of the Sampling Distribution



Source: National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2023)

Long description of figure 5.14

A similar process is used to evaluate confidence in a randomized experiment in which subjects are randomly assigned to two or more treatment groups. ("Treatment" could mean medical treatment, or assignment of different tasks, or being shown different motivational videos, and so on.) Some quantity is then measured for each subject, and the investigator then must decide from the results whether a treatment, say treatment A, produced any effect on the measured quantity. Simply having a different mean for each treatment group is not enough, as variation is expected in the measurement and thus between groups. In this case, all the treatment groups are pooled into a population and then resampled (randomly) many times to see how often the resampled mean or proportion is at least as extreme as the actual treatment A group difference. If such differences are rare, the experiment is taken as evidence that treatment A caused a change in the measured quantity.

In the following snapshot, students investigate real-world environmental data and health impacts. The snapshot demonstrates how the sequence provided them opportunities to use all three thematic topics (shown in parentheses within the snapshot).

Snapshot: Data on Environmental Threats to Health

In this example provided by Gerald Lieberman and K. Brown, students compared CalEnviroScreen data related to four environmental topics that are known to affect human health:

1. water (using data on groundwater threats, impaired water, and drinking water);
2. toxic chemicals (using data on pesticides, cleanups, and toxic releases);
3. air pollution (using data on the ozone, particulate matter [PM 2.5], diesel, and traffic); and
4. waste (using data on hazardous waste and solid waste) (2020).

The results were compared against environmental impacts, using data for asthma, low birth weight, and cardiovascular disease (California Health Education Standards 1.13.P; 2.3.P; 3.3.P, 3.4.P).

In preparation for the students' analysis and reporting, the teacher reviewed California's EP&Cs with students by asking them to identify one that is directly related to their environmental health problem. Based on their data analysis, students identified environmental health and environmental justice concerns related to water pollution in the local community and observed that they differentially affected various parts of the community (*understanding and describing variability in data and data distributions*). Their conclusion was that the key factors in the differential environmental health impacts were related to "Environmental Principle V: Decisions affecting resources and natural systems are based on a wide range of considerations and decision-making processes."

Depending on the focus of their individual environmental health study, students are encouraged to choose two variables to analyze, such as the impact of water quality on low birth weight, or the impact of toxic chemicals on the incidence of cardiovascular disease, or the impact of air quality on asthma (*data collection, sampling, and random processes*). After collecting the data for these variables, the students use technology to create a scatter plot of the data, fit a function to the data, and create a symbolic representation for the function (*understanding and describing variability in data and data distributions*). Students are able to connect the parameters of the symbolic representation to the context of the data. After a class discussion about the comparison of different variables, students should be guided to focus on the combinations of variables that make the most sense for their investigations (*comparing distributions and identifying associations between variables*).

Following their research and analysis, student teams report back to the class, summarizing their quantitative comparisons using charts to depict the results about water, toxic chemicals, air pollution, and waste (Health 4.1.P). In their presentation, they use graphs to compare the environmental effects they discovered from the environmental health impacts they analyzed (English Language Arts SL.9-12.1; SL.9-12.2; SL.9-12.4; SL.9-12.5; Health 5.3.P).

Several of the teams mention that they observed a pattern that relates to the socioeconomic conditions in the communities they compared. Some of the students mention that they see these issues as directly related to EP&C V because the places where waste, toxic chemicals, and manufacturing facilities are located depend on a variety of political, economic, and social factors. The teacher explains that differential environmental health impacts on communities with varied socioeconomic conditions is a major health topic identified as "environmental justice," a term that came into use in the 1980s when residents of an African American community in North Carolina protested the siting of a landfill to store soil contaminated with polychlorinated biphenyls (PCBs). These residents knew the health hazards associated with this toxin and responded by demanding that their health and well-being be protected by the government. The landfill proposal went forward, but the protests spurred the federal government to study the issue. The findings show that many of the nation's landfill sites are located in communities of color. The environmental justice movement has grown to focus on a more equitable distribution of environmental benefits and burdens. Since many of the students expressed a strong interest in this topic, they request a guest speaker from a community-based health organization to provide additional information and answer students' questions about environmental justice (Health 8.1.P.; 8.2.P).

Guidance for High School Data Science Courses

While all high school students should exercise and refine their understanding of data exploration, causal inference, and statistical reasoning using large, real-world data sets, many sophisticated approaches to working with rich, complex data sets can be left to an advanced statistics or data science course. (Chapter eight and appendix A provide further detail on the different mathematics pathways available to schools and students.)

With the rapid expansion of information available to all in the form of data, students may be interested in a data science course as a culminating high school mathematical science experience. In addition to recognizing the importance of the data science content—to twenty-first-century jobs and to a wide range of college majors—many students are more engaged with math classes that are taught in a spirit of open-ended exploration, drawing upon important mathematical principles and tools rather than more traditional teaching methods focused solely on procedures without motivation or context.

Effective data science courses consider how to help students with the following:

- Understand how data are used by professionals to address real-world problems
- Understand that data are used in all facets of modern life
- Understand how data support science to identify and tackle real-world problems in communities
- Learn about statistical variability
- Use appropriate tools and techniques to make sense of large data sets
- Create and analyze statistical graphics to identify patterns in data and to connect these patterns back to the real world
- Understand that by treating photos, words, numbers, and sounds as data, you can gain insight into the real world
- Learn to analyze data, including posing questions that can be answered by considering relations among variables in a data set, using collected data to generate hypotheses for future data collection, critically evaluating shortcomings and strengths in the data and the data-collection process, and informally evaluating hypotheses using data at hand
- Learn basic programming and use computer programs in the development and analysis of statistical models
- Learn about data ethics, including consideration of where data come from, who is collecting the data, and how the data are used
- Refine computational models to better represent the relationships among different elements of data that are collected and analyzed
- Design algorithms to solve computational problems by using and adapting existing algorithms and creating new ones

When designing or planning for a data science course, there are many different sequences and approaches. A course might actively engage students by exploring the meaning of data, the importance of communicating data visually, the role of cleaning data, exploratory data analysis, ethical issues around data, data dashboards, linear and nonlinear regression models, statistics, probability, and forecasting.

In addition to the mathematics content described above, exposure to some software and other technology tools is essential for those wishing to pursue a career in data science, and facility with such tools is increasingly valuable for a variety of professionals whose work involves basic data analysis. However, data science does not require any particular software package, and different data science courses may use different software and technology tools depending on the specifics of the course and school context. More important than the technology students use is that they learn to ask good questions and apply effective mathematical tools to help them answer their questions.

Equitable and Inclusive Instruction

Educators can offer social and emotional support to students by designing engaging lessons that allow students to connect in meaningful ways with content. Traditional mathematics lessons that have taught the subject as a set of procedures to follow have often resulted in widespread disengagement as many students see no relevance for their lives. The data science field provides multiple opportunities for students to pursue answers to real-world problems and their own wonderings and to see that they can excel in quantitative fields.

Important principles underlying the teaching of data science that will offer the greatest chance for social, emotional, and academic development include the following:

- *Convey Mindset and Belonging Messages*

  Informed by successful interventions in mindset and belonging strategies, teachers can remind students that struggle represents an important part of learning; all students struggle at times, and successful students respond to times of difficulty by using strategies they have developed and practiced over time. Teachers can share with students examples of successful people within the field that highlight gender and racial diversity.

- *Use Real Data*

  Modern computing provides an opportunity for students to question real sets of data, developing social awareness and investment in the solutions they discover. When working with secondary data sets (data obtained from others, rather than collected by students), teachers should choose meaningful content selected to create a connection with their learning and secure opportunities to hear the perspective of others. When teachers use local data sets, they can also help students feel like they are important members of their community—as they use real data to explore questions and find answers to local problems that they can help to address. Identifying problems and finding solutions will help students develop skills to make responsible decisions. Some teachers worry that they cannot provide culturally sustaining connections for their classes because they lack expertise in the cultures of all their students, but real data sets from different communities provide opportunities for students to bring their own knowledge and expertise to data-rich problems. There should also be times when students are invited to collect data from their own community and build their own data sets. Students can pose questions that are important to them, including those with cultural meaning, collecting data from their own lives and communities. As Django Paris describes, students will thereby be fostering and sustaining "linguistic, literate, and cultural pluralism" (2012). The act of collecting data provides an important learning opportunity for students to understand decisions that need to be made around the collection and organization of data and to understand how to deal with uncertainty in their data. Students will be the ones with important expertise in these investigations.

- *Focus on Collaboration and Communication*

  Meaningful collaborations typically reflect perspectives from diverse groups of students who come together to work effectively with different ideas being valued and developed. Creating opportunities for this kind of group work makes an environment where differences thrive and where students have the tools to work respectfully to reach solutions. For example, students may start their work in structured and unstructured conversations in which each group member shares their thoughts. Collaborative classrooms founded in engaged listening and the capacity to articulate verbally as students build on each other’s ideas are places where students feel valued and where they develop important relationship skills of communication, social engagement, and teamwork.

Connecting to the Drivers of Investigation and Content Connections

This chapter highlights three thematic topics central to data science that exist within the CA CCSSM K–12 progressions. Data investigations that leverage the SMPs provide students with opportunities to recognize that statistics is a problem-solving process that connects the mathematics they are doing in class to their lived experiences and to other content areas. The chapter also describes how investigations can provide students with needed opportunities to collect data within their classrooms and to consider the implications of their decisions (such as decisions about sampling). The grade band descriptions included examples of how students can use plots to develop an understanding of variability in data and also use plots as a communication tool to describe that variability. Additionally, as students pose their wonderings about the world around them or try to predict future events using data, there may be opportunities to explore probability and identify associations between variables. A focus on these thematic topics helps ensure that all students continue to grow in their abilities to work with data and lay the groundwork for pathways toward data science.

To help teachers who have been working with the CA CCSSM standards and progressions, this chapter focuses on the progressions of statistical and data science ideas across grade levels. Additionally, as described in chapter one, this framework encourages teachers to design instruction by using the big ideas of mathematics at each grade level as focal points for student investigations. Investigations are guided by the three Drivers of Investigation (DIs) that provide the "why" of learning mathematics, the eight SMPs that provide the "how" of learning mathematics, and the four Content Connections (CCs) that provide the "what" of learning mathematics.

To address the mathematical concepts as discussed in this chapter within such an approach, educators may want to begin with the kinds of questions they are asking to build mathematical understandings. The aim of the DIs is to ensure that there is always a reason to care about mathematical work—and that investigations provide opportunities for students to make sense, predict, and/or affect the world. Just as the three thematic topics progress across the K–12 band, early DI questions are primarily about description and begin with categorizing and counting, expanding into questions that present measurement situations (initially length/distance; later time, area, volume, and rates). As students progress through the grade bands, they begin to investigate relationships between two or more varying quantities and perform formal quantitative calculations to describe future events or probable outcomes. Figure 5.15 illustrates three different questions associated with the three different DIs, each of which enables elementary students to learn about and use data in ways described earlier in this chapter.

Figure 5.15: Using Drivers of Investigation to Frame Questions and Investigations About Weather

Driver of Investigation 1: Making Sense of the World (Understand and Explain)	Driver of Investigation 2: Predicting What Could Happen (Predict)	Driver of Investigation 3: Impacting the Future (Affect)
"What are different ways to describe our weather?"	*"What will the weather be like tomorrow?"*	*"How can we use weather data to make recommendations to visitors for packing or outdoor activities?"*

When the DIs are coupled with data from relevant contexts, students may be more likely to authentically engage in statistical problem-solving. Connecting DIs to contexts relevant to students' backgrounds and interests may also increase the inclusion of and participation among girls and students from racial and ethnic groups historically underrepresented in STEM fields.

When designing learning sequences, the DI will link CCs and one or more SMPs together. While CC1 (Reasoning with Data) is explicitly tied to data, earlier portions of this chapter note many places throughout the K–12 experience where data explorations might arise and thus support CCs 2, 3, and 4. Figure 5.16 uses the weather situation from figure 5.15 to show how investigations using weather data might support each of the four CCs. Within each row, the bolded information (labeled "Thematic Topic") links back to the primary thematic topic covered in this chapter.

Figure 5.16 Using Weather Data to Support Content Connections

Content Connection	Supporting Investigation
Content Connection 1: Reasoning with Data	Students have the opportunity to count, measure, and classify attributes such as temperature, time, precipitation, or cloud cover. These values can be expressed and explored graphically and then interpreted and shared with peers or the community. **Thematic Topic: Understanding and describing variability in data and data distributions**
Content Connection 2: Exploring Changing Quantities	The collection of weather data can provide a rich context for students to explore their numeracy and develop and apply operations (e.g., *What is the difference between the highest and lowest daily temperature for Tuesday?*); express values in terms of ratio or percent (e.g., *What percentage of days were cloudy for the month of March?*); and find ways to express patterns between quantities mathematically via multiple representations or express climate patterns through linear equations (e.g., *The temperature increase over time can be described by 1.2C + 22*). **Thematic Topic: Data collection**
Content Connection 3: Taking Wholes Apart, Putting Parts Together	Describing the weather requires students to decompose the investigation question into attributes that could and should be collected to address the question driving the investigation. Students might explore data on the average amount of different types of precipitation (e.g., rain versus snow) in each calendar month and then combine those data to develop a more detailed understanding of total annual precipitation. **Thematic Topic: Understanding and describing variability in data and data distributions**
Content Connection 4: Discovering Shape and Space	While there may be variation in daily temperatures, the shape of the plot for temperatures throughout the day or for particular seasons is overall quite predictable. Length of day or seasonal weather patterns are inversely related between the northern and southern hemispheres. Students might explore 2-D versus 3-D graphical representations and the role of geospatial data in weather forecasting. **Thematic Topic: Comparing distributions and identifying associations between variables**

Conclusion

As readers consider the three subsequent chapters of the framework, they will see ideas similar to the ones discussed in this chapter, organized to help them learn about and begin to use the big ideas approach. While the transition between standards domains and progressions discussed in this chapter and this new approach will not be straightforward for classroom teachers, both emphasize the central idea that students at all levels should have experiences that build their mathematical toolkits for making sense of their worlds.

Life in a data-rich world requires that California schools prepare all students to examine claims justified with data, to understand the probabilistic underpinning of drawing conclusions from samples, and to see the use of data as a tool for answering many questions of interest. Developing these abilities requires that students generate questions and work with data beginning in kindergarten (or before) and have experiences of increasing depth and complexity throughout their school careers. As discussed in chapter eight, students who wish to focus extra attention on data science should have an opportunity to pursue advanced courses late in their high school careers.

Additional Resources

As educators consider how to create rich lessons that integrate data and statistical investigations into their classrooms, the following resources, which influenced this document, may be helpful (listed in alphabetical order):

American Statistical Association (ASA): The following is an excerpt from the ASA K–12 web page: "The American Statistical Association is dedicated to and involved in enhancing statistics education at all levels, including providing resources for K–12 teachers and teacher educators. Here, you will find information about classroom resources, publications in statistics education, guidelines and reports, workshops and webinars for teachers, and student competitions."

https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link1

GAISE II: The Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) is a professional report from the ASA setting out guidelines for assessment and instruction in PreK–12 in statistics and data science (Bargagliotti et al. 2020). The GAISE II is an important resource for this area of mathematical science. It includes guidance and examples for skills and concepts at the elementary, middle, and high school levels.

https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link2

National Council of Teachers of Mathematics (NCTM): The NCTM published two books in its Essential Understanding series devoted to statistics: Developing Essential Understanding of Statistics: Grades 6–8 and a second volume for grades nine through twelve.

https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link3
https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link4

The Statistical Education of Teachers (SET): The following is an excerpt from the introduction:

> *The SET report outlines the content and conceptual understanding teachers need to know to assist their students develop statistical reasoning skills. SET is intended for everyone involved in the statistical education of teachers, both the initial preparation of prospective teachers and the professional development of practicing teachers. (Franklin et al. 2015)*

https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link5

Statistics Teacher: The following is an excerpt from the publication's website:

> *In 2016, the ASA/NCTM Joint Committee decided the Statistics Teacher Network newsletter should evolve to the Statistics Teacher online journal. The goal of Statistics Teacher remains to inform and support K–12 teachers. The new publication will continue to provide articles about successful classroom practice and announcements of important professional development opportunities. It will also more seamlessly integrate peer-reviewed lesson plans. Each issue will have dedicated technology and assessment columns. (American Statistical Association 2024)*

https://www.cde.ca.gov/ci/ma/cf/ch5.asp#link6

California Department of Education, October 2023