Chapter 8
Learning to Reason About Distribution

Statisticians look at variation through a lens which is "distribution".

(Wild, 2006, p. 11)

Snapshot of a Research-Based Activity on Distribution

Groups of three to four students are each given an envelope containing 21 different pieces of paper, each with a different histogram printed on it. Students sort the graphs into piles, so that the graphs in each pile have a similar shape. After sorting them into the piles (e.g., normal/bell-shaped, left-skewed, right-skewed, bimodal, and uniform), students choose one histogram from each pile that best represents that category, and these selections are shared and discussed as a class. Students use their own informal language to "name" each shape: bell-shaped, bunched to one side, like a ski slope, camel humps, flat, etc. These informal names are matched to the formal statistical terms such as normal, skewed, bimodal, and uniform. Finally, students consider which terms are characteristics (e.g., skewness) that can apply to graphs in more than one category (e.g., a distribution that is skewed and bimodal) vs. those that can only be labeled by one name (e.g., uniform or normal distribution).

Rationale for This Activity

Although this activity may seem like a game for elementary school students, the activity involves some important challenges and learning outcomes for high school and college students in a first statistics course. First of all, this activity helps students look at histograms as an entity, rather than as a set of data values and cases, which research has shown to be a key problem in reasoning about distributions. Secondly, students often fail to see the general shape of distributions, because of the effects of randomness (the "noise"); and expect to see perfect shapes like the models given in their textbooks. This activity helps them see that there are many types of "normal" distributions or skewed distributions. They learn to look beyond the individual features of the graph and see the more general or global characteristics. Finally, this activity focuses on the language used to describe distributions, which can often be confusing to students. The word "normal" in statistics refers to a bell-shaped curve

J.B. Garfield, D. Ben-Zvi, *Developing Students' Statistical Reasoning*,

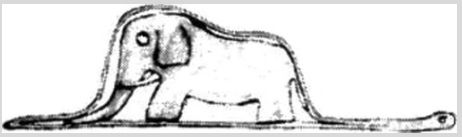
that has certain characteristics while in everyday life it means typical or not unusual. Students can take an informal name of a shape (e.g., ski slope for a right-skewed distribution) and map them to the correct statistical labels, which can then be used to help remind them of the statistical term (e.g., a teacher talking about a skewed curve can say, "remember it is like a ski slope").

The Importance of Understanding Distribution

We begin this section with a poignant illustration, offered by Bill Finzer to participants at the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy, the focus of which was on "Reasoning about Distribution" (Makar, 2005)

The Little Prince, by de Saint-Exupéry (2000) begins with this drawing.

To adults, the drawing looked exactly like a hat.
To the child artist who drew it and to the little prince, it was a drawing of a boa that had eaten an elephant.



If the little prince showed this picture to a statistician, he would say: "This represents a distribution of data."

We like this example because it shows how statisticians look at irregular shapes of data sets and look beyond the details to see a general shape and structure. This is usually a first step in any data analysis and leads to important questions about the

data to be analyzed, such as: What mechanism or process might have led to this shape? Are there any values that need to be investigated (e.g., possible outliers)?

A graph of a distribution reveals variation of a quantitative variable. According to Wild (2006), statisticians respond to the "omnipresence of variability" in data (Cobb & Moore, 1997) by investigating, disentangling, and modeling patterns of variation represented by distributions of data. He suggests that statisticians look at variation through a lens that is "distribution."

Students encounter two main types of "distribution" in an introductory statistics class. The first type is distributions of sample data that students learn to graph, describe, and interpret. These are *empirical* distributions of some particular measured quantity. The second type of distribution encountered is a *theoretical* one, e.g., normal or binomial distributions, which are actually probability models (Wild, 2006).

Although the two types share many common features (e.g., they can be described in terms of shape, center, and spread), it is important to help students distinguish between them because of the way in which we use them. The distinction that underlies *empirical* versus *theoretical* distributions relates to variation. When examining an empirical distribution, the focus is on description and interpretation of the message in the data, and thinking about what model may fit or explain the variation of the data. Theoretical distributions are models to fit to data, to help explain, estimate, or make predictions about the variability of empirical data. Yet, a third type of distribution, students encounter in a statistics course is a distribution made up of sample statistics, which again has both empirical and theoretical versions. These *sampling distribution*s are discussed in detail in Chapter 12.

The Place of Distribution in the Curriculum

Empirical distributions are the foundation of students' work in an introductory statistics course, either beginning a course or following a unit on collecting and producing data (experiments and surveys, Chapter 6). This chapter focuses mainly on teaching and learning issues related to *empirical* distributions, while Chapters 7 and 12, respectively, also discuss *theoretical* distributions. In addition, this chapter focuses on understanding a single distribution, primarily in the form of dotplots and histograms, while Chapter 11 examines these graphs along with boxplots in the comparison of two or more distributions.

The methods of Exploratory Data Analysis introduced by Tukey (1977) have had a big impact on the way distributions are taught in today's courses. Students use many ideas and tools to explore data and learn to think of data analysis as detective work. Students usually learn multiple ways to graph data sets by hand and on the computer. These methods include dotplots (also called line plots), stem and leaf plots, histograms, and boxplots. Students learn that different graphs of a data set reveal different characteristics of the data. For example, a histogram or dotplot gives a better idea of the shape of a data set, while a boxplot is often better at revealing an outlier. A stem-and-leaf plot or dotplot may give a better idea of where there are clumps or gaps in the distribution.

Current statistical software programs (e.g., *TinkerPlots, Fathom*) allow students to easily manipulate data representations, for example, to transform one graph of a data set to another, display several interlinked graphs of the same data set on one screen (a change in one will show in the others), and some allow students to highlight particular data values and see where they are located in each graph. These explorations are used to ask questions about the data: What causes gaps and clusters? Are outliers real data values or errors in data collection or coding? What factors may help explain the features revealed in a graph of a distribution? In most introductory courses along with learning how to graph distributions of data, students are taught to look for specific features of distributions and begin to describe them informally (e.g., estimate center and range) and then more formally (e.g., shape, center, and spread).

Distribution is one of the most important "big ideas" in a statistics class. Rather than introduce this idea early in a class and then leave it behind, today's more innovative curriculum and courses have students constantly revisit and discuss graphical representations of data, before any data analysis or inferential procedure. In a similar vein, the ideas of distributions having characteristics of shape, center, and spread can be revisited when students encounter theoretical distributions and sampling distributions later in the statistics course.

Review of the Literature Related to Reasoning About Distribution

The research literature provides a strong case that understanding of distributions, even in the simplest forms, is much more complex and difficult than many statistics teachers believe. Although little of the research includes college students, the results of studies on precollege level students and precollege level teachers demonstrate the difficulty of learning this concept, some common misconceptions, and incomplete or shallow understandings that we believe also apply to college students.

Much of the research on distribution emerged because of the consensus in the statistics education community that it is a basic building block for a web of key statistical ideas, such as variability, sampling, and inference (e.g., Garfield & Ben-Zvi, 2004; Pfannkuch & Reading, 2006). Other studies (e.g., Reading & Shaughnessy, 2004; Watson, 2004) focused on broader questions than how students reason about distribution, but yielded relevant results. For example, Chance et al. (2004) assert that the knowledge of distribution and understanding of histograms are necessary prerequisites to learning and understanding sampling distributions.

Developing an Aggregate View of Distribution

A major outcome of several studies on how students solve statistical problems is that they tend not to see a data set (statistical distribution) as aggregate, but rather

as individual values (e.g., Hancock, Kaput, & Goldsmith, 1992). Konold & Higgins (2003) claimed that, "students need to make a conceptual leap to move from seeing data as an amalgam of individuals each with its own characteristics to seeing the data as an aggregate, a group with emergent properties that often are not evident in any individual member" (p. 202). They explained this challenging transition in the following way:

> With the individuals as the foci, it is difficult to see the forest for the trees. If the data values students are considering vary, however, why should they regard or think about those values as a whole? Furthermore, the answers to many of the questions that interested students—for instance, Who is tallest? Who has the most? Who else is like me?—require locating individuals, especially themselves, within the group. We should not expect students to begin focusing on group characteristics until they have a reason to do so, until they have a question whose answer requires describing features of the distribution. (Konold and Higgins, 2003, p. 203)

To explore the emergence of second graders' informal reasoning about distribution, Ben-Zvi and Amir (2005) studied the ways in which three second grade students (age 7) started to develop informal views of distributions while investigating real data sets. They described what it may mean to begin reasoning about distribution by young students, including two contrasting distributional conceptions: "flat distribution" and "distributional sense". In the "flat distribution" students focused just on the values of distribution and did not refer at all to their frequencies, while students who started acquiring a "distribution sense" showed an appreciation and understanding that a distribution of a variable tells us what values it takes and how often it takes these values. The gradual transfer from the incomplete perception of a distribution towards the more formal sense of distribution presented an immense challenge to these students.

In a teaching experiment with older students (seventh grade students in Israel), Ben-Zvi and Arcavi (2001) show how students were able to make a transition from *local* to *global* reasoning, from *individual-based* to *aggregate-based reasoning*. The researchers found that carefully designed tasks (e.g., comparing distributions, handling outliers), teachers' guidance and challenging questions, along with motivating data sets and appropriate technological tools helped students to make this transition.

Konold, Pollatsek, Well, and Gagnon (1997) interviewed two pairs of high-school students who had just completed a year-long course in probability and statistics. Using software and a large data set students had used as part of the course, these students were asked to explore the data and respond to different questions about the data and to support their answers with data summaries and graphs. The results suggest that students had difficulty in thinking about distributions and instead focused on individual cases. They did not use the methods and statistics learned in the course when comparing two distributions, but instead relied on more intuitive methods involving comparisons of individual cases or homogeneous groups of cases in each group. Results were re-analyzed along with results from two other studies (Konold, Higgins, Russell, & Khalil, 2003) and the following types of responses were suggested as ways students reason about a distribution of data.

1. Seeing data as Pointers (to the larger event from which the data came).
2. Seeing data as Case-values (values of an attribute for each case in the data set).
3. Seeing data as Classifiers (giving frequency of cases for a particular value).
4. Seeing data as an Aggregate (the distribution as an entity with characteristics such as shape, center, and spread).

The authors note that although an important goal in statistics is to help students see a distribution as an aggregate, they feel it is important to pay attention to students' initial views of data and to carefully help them gradually develop the aggregate view (Konold et al., 2003).

Understanding the Characteristics of a Distribution

Several studies focused on how students come to conceive of shape, center, and spread as characteristics of a distribution and look at data with a notion of distribution as an organizing structure or a conceptual entity. For example, based on their analysis of students' responses on the National Assessment of Educational Progress (NAEP) over the past 15 years, Zawojewski and Shaughnessy (2000) suggest that students have some difficulty finding the mean and the median as well as difficulty selecting appropriate statistics. They explain that one of the reasons that students do not find the concepts of mean and median easy may be that they have not had sufficient opportunities to make connections between centers and spreads; that is, they have not made the link between the measures of central tendency and the distribution of the data set. Mokros and Russell (1995) claim that students need a notion of distribution before they can sensibly choose between measures of center and perceive them as "representatives" of a distribution.

Reasoning About Graphical Representations of Distributions

One of the difficulties in learning about graphical representations of distributions is confusion with bar graphs. In elementary school, students may use bars to represent the value of an individual case (e.g., number of family pets), or a bar can represent the frequency of a value (e.g., number of families with one pet). Today, some statistics educators distinguish between these two types of representations, referring to case-value plots as the graphs where a line or bar represents the value of an individual case, or student. In contrast, the bars of a histogram represent a set of data points in an interval of values. While case-value and bar graphs can be arranged in any order (e.g., from smallest to largest or alphabetical by label), bars in a histogram have a fixed order, based on the numerical (horizontal) scale. Furthermore, while the vertical scale of a histogram is used to indicate frequency or proportion of values in a bar (interval), the vertical scale for a bar graph may represent either a frequency or proportion for a category of categorical data, or it may represent magnitude (value of a case presented by that bar). These differences can cause confusion in students, leading them to try to describe shape, center, and

spread of bar graphs or to think that bars in a histogram indicate the magnitude of single values (Bright & Friel, 1998).

Establishing connections among data representations is critical for developing understanding of graphs; however, students cannot make these connections easily and quickly. To find instructional strategies that help learners understand the important features of data representations and the connections among them, Bright and Friel (1998) studied ways that students in grades 6, 7, and 8 make sense of information in graphs and connections between pairs of graphs. They report that students benefited from these activities by recognizing the importance of "the changing roles of plot elements and axes across representations", and, therefore, suggest that teachers need to "provide learners with opportunities to compare multiple representations of the same data set" (p. 87). They also suggest to promote rich discourse about distributions of data in the classroom to help students understand the important aspects of each representation.

Students' recognition of graphical aspects of a distribution as an entity was studied by Ainley, Nardi, and Pratt (2000). They observed young students (8–12 years) who collected data during ongoing simple experiments and entered them in spreadsheets. They noted that despite limited knowledge about graphs, students were able to recognize abnormalities (such as measurement errors) in graphs and to take remedial action by adjusting the graphs toward some perceived norm. The researchers have labeled this behavior, "normalizing," an activity in which children construct meanings for a trend in data and in graphs. Ainley and her colleagues claim that children gained this intuitive sense of regularity from everyday experience, experience gained during the activity, their sense of pattern, or from an emerging perception of an underlying mathematical model. The researchers recommend the use of computer-rich pedagogical settings to change the way in which knowledge about data graphs is constructed.

Helping Students to Reason with Graphs of Distributions

Students often see and use graphs as illustrations rather than as reasoning tools to learn something about a data set or gain new information about a particular problem or context (Wild & Pfannkuch, 1999; Konold & Pollatsek, 2002). Current research on students' statistical understanding of distribution (e.g., Pfannkuch, 2005a; Watson, 2005) recommends a shift of instructional focus from drawing various kinds of graphs and learning graphing skills to making sense of the data, for detecting and discovering patterns, for confirming or generating hypotheses, for noticing the unexpected, and for unlocking the stories in the data. It has been suggested that reasoning with shapes forms the basis of reasoning about distributions (Bakker, 2004a; Bakker and Gravemeijer, 2004).

Others refer to developing skills of visual decoding, judgment, and context as three critical factors in helping students derive meaning from graphs (Friel, Curcio and Bright, 2001). Reasoning about distributions is more than reasoning about shapes. It is about decoding the shapes by using deliberate strategies to

comprehend the distributions and by being cognizant of the many referents, which are bound within the distributions. Furthermore, students have to weigh the evidence to form an opinion on and inference from the information contained in the distributions (Friel et al., 2001). Such informal decision-making under uncertainty requires qualitative judgments, which are much harder than the quantitative judgments made by statistical tests (Pfannkuch, 2005a).

In one of the rare studies at the college level, delMas, Garfield, and Ooms (2005) analyzed student performance on a series of multiple-choice items assessing students' statistical literacy and reasoning about graphical representations of distribution. They found that college students, like younger peers in middle school described above, confused bar graphs and histograms, thinking that a bar graph of individual cases, with categories on the horizontal scale, could be used to estimate shape, center, and spread. They also thought that such a bar graph might look like the normal distribution. They tended to view flat, rectangular-shaped histograms as a time series plot showing no variation, when these graphs typically show much variation in values. The researchers also identified errors students make in reading and interpreting horizontal and vertical axes. Based on the difficulties students appeared to have reading and interpreting histograms, the authors questioned whether students should be taught to use only dotplots and boxplots to represent data sets. After questioning colleagues, they concluded that there were important reasons to keep histograms in the curriculum as a way of representing distributions of data, because of the need for students to understand the ideas of area and density required for understanding theoretical distributions, and because dotplots are not feasible for very large data sets.

Technological Tools to Develop the Concept of Distribution

Technology can play an important role in developing distributional reasoning by providing easy access to multiple representations and endless opportunities to interactively manipulate and compare representations of the same data set. However, this is not a simple task. Biehler (1997b) reports that despite using an innovative software tool to generate and move between different graphs of data, interpreting and verbally describing these graphs were profoundly difficult for high school and college students, unless they had a conceptual understanding of the foundational concepts.

To study the impact of technology on distributional understanding, Cobb (1999), McClain & Cobb (2001), and Bakker & Gravemeijer (2004) examined how a hypothetical learning trajectory, translated into a particular instructional sequence, involving the use of *Minitools* (Cobb et al., 1997) supported the development of students' statistical reasoning about distribution. *Minitools* are simple but innovative Web applets that were designed and used to assist students to develop the concept of distribution. Results of these teaching experiments suggest that students' development of relatively deep understandings of univariate distribution are feasible goals at the middle school level, when activities, discussion, and tools are used in particular ways.

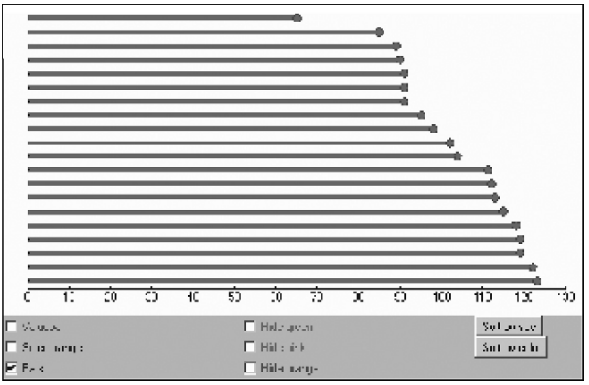


Fig. 8.1 Life span of batteries displayed by a case-value bar graph in *Minitool 1* (sorted by size). Each horizontal bar represents the life span in hours for a particular battery

For example, one aspect of distribution, shape, can be seen by looking at histograms or dotplots. To understand what dots in a dotplot represent, students need to realize that a dot corresponds to a value of a particular variable, and each dot represents one case that has that particular value. To help students develop this insight, a tool shows them case-value bars (*Minitool 1*, see Fig. 8.1). These bars seem to correspond to students' intuitive ways of organizing and displaying a set of data. Students then are helped to make a transition to a second *Minitool* (Fig. 8.2), which takes the end points of the case-value bars and stacks them in a dotplot. While each case in *Minitool 1* is represented by a bar whose relative length corresponds to the value of the case, each case in *Minitool 2* is represented by a dot in a dotplot. Figures 8.1 and 8.2 display just one data set at a time; however, all given data sets in *Minitools* include two groups (e.g., comparing two brands of batteries), which better help students develop distributional reasoning.

Minitool 2 has options to organize data in ways that can help students develop their understanding of distributions. For example, the dotplot can be divided into equal groups or into equal intervals, which support the development of an understanding of the median and quartiles, boxplot, density, and histogram, respectively

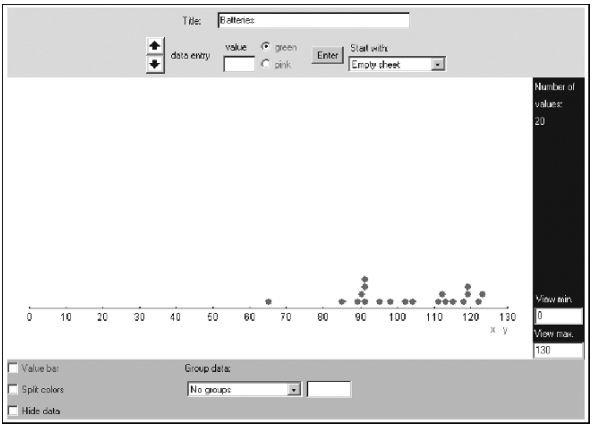


Fig. 8.2 The same data set of life span of batteries displayed by a dotplot in *Minitool 2*

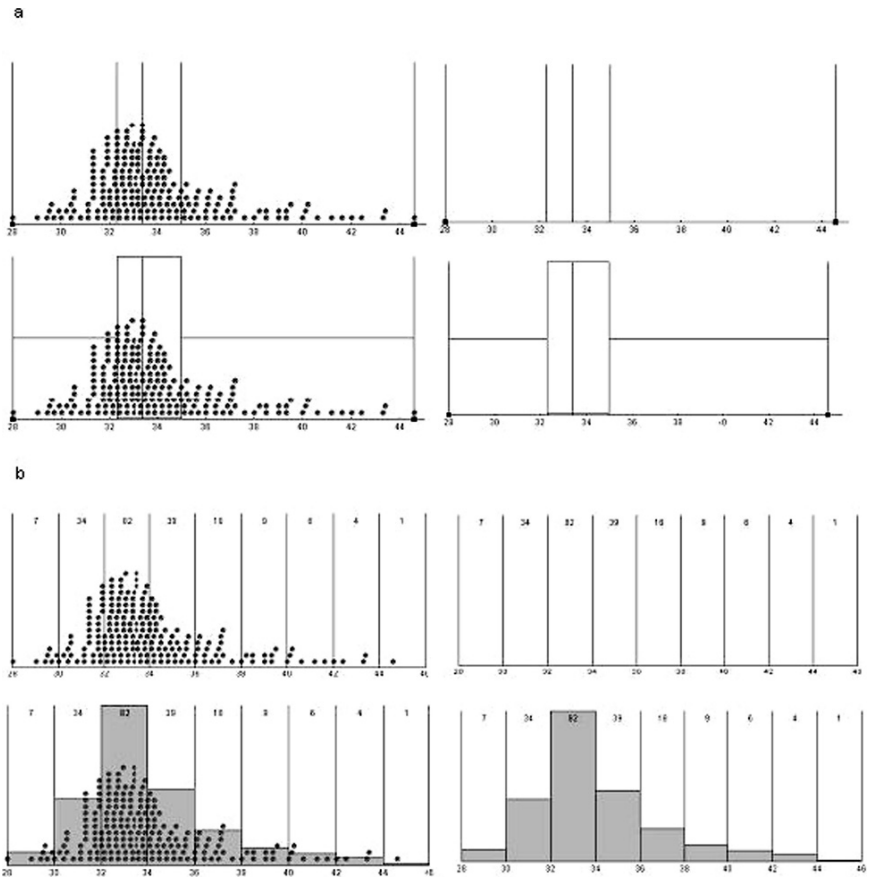


Fig. 8.3 (**a**) Four equal group and boxplot overlay option with and without data. (**b**) Fixed interval width and histogram overlay options with and without data

(see Fig. 8.3). Many of the features of the *Minitools* have been incorporated into the recently published *TinkerPlots* software (Konold & Miller, 2005).

The combination of the two *Minitools* graphs was found to be useful in helping students develop the idea of distribution. Bakker and Gravemeijer (2004) identified patterns of student answers and categorized an evolving learning trajectory that had three stages: Working with graphs in which data were represented by horizontal bars (*Minitool 1*, Fig. 8.1), working with dotplots (*Minitool 2*, Fig. 8.2), and focusing on characteristics of the data set, such as bumps, clusters, and outliers using both *Minitools*.

Based on their research, Bakker and Gravemeijer (2004) suggest several promising instructional heuristics to support students' aggregate reasoning of distributions: (1) Letting students invent their own data sets could stimulate them to think of a data set as a whole instead of individual data points. (2) *Growing samples*, i.e.,

letting students reason with stable features of variable processes, and compare their conjectured graphs with those generated from real graphs of data. (3) Predictions about the *shape* and location of distributions in hypothetical situations. All these methods can help students to look at global features of distributions and foster a more global view.

Implications of the Research: Teaching Students to Reason About Distribution

The results of these studies suggest in general that it takes time for students to develop the idea of distribution as an entity, and that they need repeated practicing in examining, interpreting, discussing, and comparing graphs of data. It is important to provide opportunities for students to build on their own intuitive ideas about ways to graph distributions of data. Some of the research suggests that students use their own informal language (e.g., talking about 'bumps' and 'clumps' of data) before learning more formal ones (e.g., mode, skewness). The research also suggests that teachers begin having students use graphical representations of data that show all the data values (e.g., dotplots or stem-and-leaf plots) and carefully move from these to more abstract and complex graphs that hide the data (e.g., histograms and boxplots), showing how different graphs represent the same data. Several studies suggest a sequence of activities that leads students from individual cases (case-value bars) to dotplots to groups of data points (clusters in intervals) to histograms. This sequence can later be used to develop the idea of a boxplot (see Chapter 11). New computer tools (e.g., *Minitools, TinkerPlots*, and *Fathom*) show promise for helping to guide students through this process and to allow them to connect different graphical representations of distributions.

Teaching Students the Concept of Distribution

The strongest message in the research on understanding graphs and distributions is that statistics teachers need to be aware of the difficulties students have developing the concept of distribution as an entity, with characteristics such as shape, center, and spread. While most textbooks begin a unit on descriptive statistics with graphs of data, when to use them, how to construct them and how to determine shape, estimate center and spread, we believe that there are some important steps to precede this. We think it is best to begin data explorations with case-value bars that represent individual cases, a type of graph students are very familiar with and that is more intuitive for them to understand and interpret. Then, this type of graph can be transformed to a dotplot using diagrams or a tool such as *TinkerPlots*. For example, a diagram of case-value bars such as the one shown below (Fig. 8.4), for a set of students test scores, can be converted by the students to a dotplot (Fig. 8.5), taking the end point of each case-value and plotting it on a dotplot.

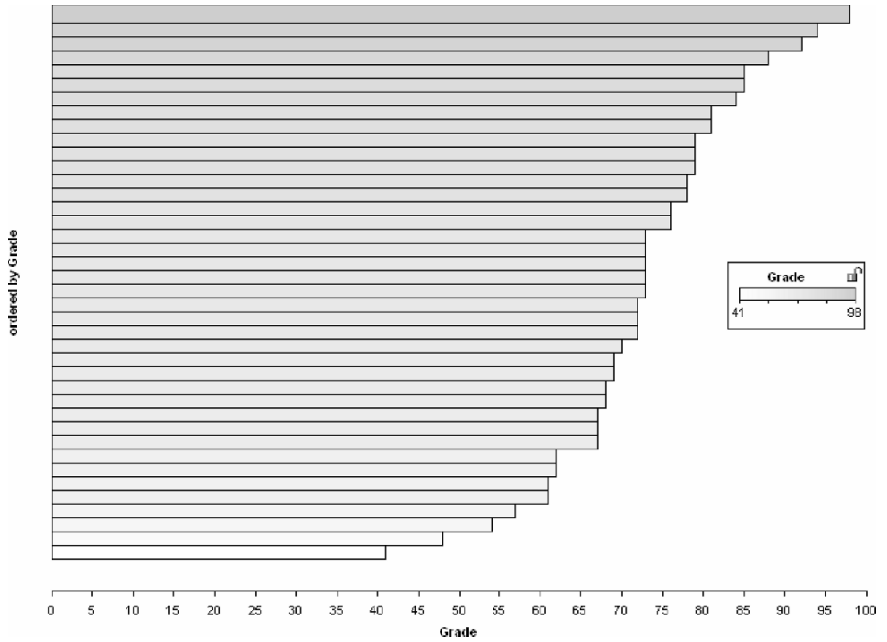


Fig. 8.4 Exam scores displayed by a case-value bar plot in *TinkerPlots*

We then suggest having students talk about categories that are useful for grouping the data, such as students who scored in the 40s, 50s, 60s, etc. (Fig. 8.6). These groupings could lead to bars of a histogram (Fig. 8.7). Students should be encouraged to compare the three types of graphs of the same data, discussing what each graph does and does not show them, how they compare, and how they are different. We also suggest moving from small samples of data to large samples, to continuous curves drawn over these graphs to help students see that plots often have some common shapes. This can also be done by giving students sets of graphs to sort and classify, as described in the snapshot of an activity in the beginning of this chapter, so that students can abstract general shapes for a category of graphs of distributions.

Another way to help students develop ideas of distribution is to help them discover characteristics of distribution. Giving them sets of graphs to compare can help

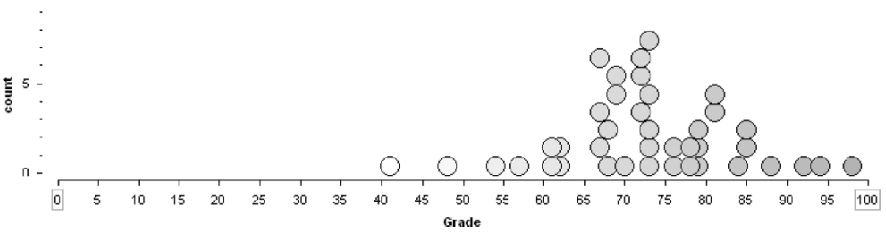


Fig. 8.5 The same exam scores displayed by a dotplot in *TinkerPlots*

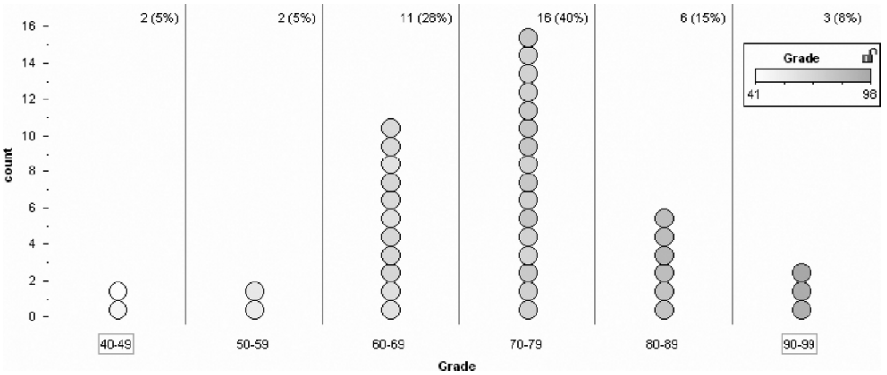


Fig. 8.6 Exam scores grouped by bins of 10 in *TinkerPlots*

do this. For example, an activity originally developed by Rossman et al. (2001) giving students a set of graphs that are similar in shape and spread, but have different centers can allow students to discover these characteristics. This can be repeated with distributions that have similar shapes and centers but differ in spread. We provide examples of these activities in Lesson 1. We focus on characteristics of distribution first using dotplots, which are easier for students to read and interpret than histograms.

The research suggests that having students make conjectures about what a set of data might look like for a particular variable and sample, can help students develop their reasoning about distribution. Rossman and Chance (2005) have also built on these ideas in their activities that have students match different dotplots to variables, forcing students to reason about what shape to expect for a particular variable, (e.g., a rectangular distribution for a set of random numbers or a skewed distribution for a set of scores on an easy test). We think it is important to have students do both activities: Draw conjectured distributions for variables and match distributions to variable descriptions. After students have studied measures of center and spread,

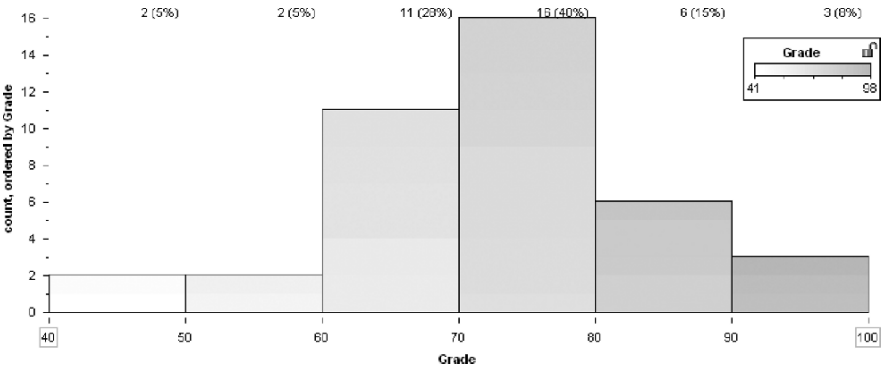


Fig. 8.7 Exam scores displayed by a histogram in *TinkerPlots*

it is suggested that they try to match graphs of distributions to a set of summary statistics, as in the histogram matching activity by Scheaffer, Watkins, Witmer and Gnanadesikan (2004b, pp. 19–21). An even more challenging version of this activity is to have students match two different representations of the same data set, such as histograms to boxplots (also found in Scheaffer et al., 2004b, pp. 21–22), which we include in Chapter 11 on Comparing Groups.

Finally, the research suggests that technology can be used to help students see the connections between different graphical representations of data, helping students to build the idea of distribution as an entity. We see three important uses of technology:

1. To visualize the transition from case-value graphs to dotplots to histograms, all based on the same data set.
2. To illustrate the ways that different graphs of the same data reveal different aspects of the data, by flexibly having multiple representations on the screen at the same time, allowing students to identify where one or more cases is in a graph.
3. To flexibly change a graph (e.g., making bins wider or narrower for histograms) so that a pattern or shape is more distinct, or to add and remove values to see the effect on the resulting graph.

Fortunately, today's software tools and Web applets readily provide each of these different types of functions.

Progression of Ideas: Connecting Research to Teaching

Introduction to the Sequence of Activities to Develop Reasoning about Distribution

Although most statistics textbooks introduce the idea of distribution very quickly, amidst learning how to construct graphs such as dotplots, stem and leaf plots, and histograms, the research literature suggests a learning trajectory with many steps that students go through in order to develop a conceptual understanding of distribution as an entity. Students need to begin with an understanding of the concept of variable, and that measurements of variables yield data values that usually vary. A set of data values can be visually displayed in different ways, the most intuitive way being individual cases, e.g., case-value graphs. After students understand how to interpret case-value graphs, they can be guided to understand dotplots, and then histograms, as one representation is mapped or transformed to the next, showing their correspondence. In this process, it is important to emphasize the role of grouping the data into different intervals (bin size), which has an effect on the graph shape, and the possible interpretations that can be drawn from it.

When first looking at dotplots and histograms, students should be encouraged to use their own language to describe characteristics, and to move from small samples of data to larger ones, as they move from empirical distributions (dotplots) to theoretical distributions (density curves). Throughout each of these phases, students should make

conjectures about data sets and graphs for particular sets of data where they can reason about the context, and then should be allowed to test these conjectures by comparing their graphs to real graphs of data. The ideas of distribution are explicitly revisited in every subsequent unit in the course: in studying measures of center and spread, in reasoning about boxplots when comparing groups, when studying the normal distribution as a model for a univariate data set, when studying sampling distributions, understanding P-values, and reasoning about bivariate distributions.

Table 8.1 shows the steps of this learning trajectory and what corresponding activities may be used for each step.

Table 8.1 Sequence of activities to develop reasoning about distribution[1]

Milestones: ideas and concepts	Suggested activities
Informal ideas prior to formal study of distribution	
• Understand that variables have values that vary and can be represented with graphs of data	• Variables on Back Activity (Lesson 1, Data unit, Chapter 6)
• Understand simple graphs of data where each case is represented with a bar (e.g., case-value graphs)	❖ An activity where students summarize and interpret data sets that are of interest to them, such as class survey data given in case-value plots. Have students arranged the points on the horizontal scale in different orders. (The symbol ❖ indicates that this activity is not included in these lessons.)
• A distribution is a way to collect and examine statistics from samples	• Gettysburg Address Activity (Lesson 3, Data unit, Chapter 6)
• A distribution can be generated by simulating data	• Taste Test Activity (Lesson 4, Data unit, Chapter 6)
• Understanding a dotplot	❖ An activity where students see how the data can be represented in a dotplot, and how this plot gives a different picture than a case value plot
Formal ideas of distribution	
• Characteristics of shape, center, and spread for a distribution	• Distinguishing Distributions Activity (Lesson 1: "Distributions")
• Features of graphs, clustering, gaps, and outliers of data	• Distinguishing Distributions Activity (Lesson 1)
• A continuous curve as representing a distribution of a large population of data	• Growing a Distribution Activity (Lesson 1)
• An understanding of histogram by changing one data set from a dotplot to a histogram, by forming equal intervals of data. Recognizing the difference between these two types of graphs	• What is a Histogram Activity (Lesson 2: "Reasoning about Histograms")
• The abstract idea of shape of histogram and recognition of some typical shapes	• Sorting Histograms Activity (Lesson 2)

[1] See page 391 for credit and reference to authors of activities on which these activities are based.

Table 8.1 (continued)

Milestones: ideas and concepts	Suggested activities
• Understand that histograms may be manipulated to reveal different aspects of a data set	• Stretching Histograms Activity (Lesson 2)
• Recognize where majority of data are, and middle half of data	❖ An activity where students describe graphs in terms of middle half of data and overall spread
• Recognize the difference between bar graphs of categorical data, case value graphs, and histograms of quantitative data	❖ An activity where students examine and compare these three types of graphs that all use bars in different ways
• Only certain types of graphs (e.g., dotplots and histograms) reveal the shape of a distribution	• Exploring Different Representations of the Same Data Activity (Lesson 2)
• Reason about what a histogram/dotplot would look like for a variable (integrate ideas of shape, center, and spread) given a verbal description or sample statistics	• Matching Histograms to Variable Descriptions Activity (Lesson 2)
Building on formal ideas of distribution in subsequent topics	
• Idea of center of a distribution and how appropriate measures of center depend on characteristics of the distribution	• Activities in Lessons 2 (Center Unit, Chapter 9)
• Idea of variability of a distribution and how appropriate measures of variability depend on characteristics of the distribution	• Activities in Lessons 1 and 2 (Variability Unit, Chapter 10)
• How a boxplot provides a graphical representation of a distribution	• Activities in Lessons 1, 2, 3, and 4 (Comparing Groups unit, Chapter 11)
• How boxplots and histograms reveal different aspects of the same distribution	• Matching Histograms to Boxplots Activity (Lesson 3, Comparing Groups Unit, Chapter 11)
• Probability distribution as a distribution of a random variable that has characteristics of shape, center, and spread	• Coins, Cards, and Dice Activity (Lesson 2, Modeling Unit, Chapter 7)
• The normal distribution as a model of univariate data that has specific characteristics of shape, center, and spread	• Activities in Lesson 3, The Normal Distribution as Model (Models Unit, Chapter 7)
• The idea of sampling distribution as distributions of sample statistics that can be described in terms of shape, center, and spread	• Activities in Lessons 1, 2, and 3 (Samples and Sampling Unit, Chapter 12)
• How statistical inferences may involve comparing an observed sample statistic to a sampling distribution	• Activities in Lessons 1 and 2, (Inference Unit, Chapter 13)
• Bivariate distribution as represented in a scatterplot	• Activities in Lesson 1 (Covariation Unit, Chapter 14)
• Characteristics of a bivariate distribution such as linearity, clusters, and outliers	• Activities in Lesson 1 (Covariation Unit, Chapter 14)

Introduction to the Lessons

The two lessons on distribution contain many small activities, which together lead students from exploring one set of data in a simple, intuitive form, to more sophisticated activities that involve comparing and matching graphs. While it is not at all "traditional" to spend two full class sessions on the idea of distribution and basic graphs, we feel strongly that unless students understand this idea early on, they will not understand most of the subsequent course material at a deep level.

Lesson 1: Distinguishing Distributions

In the first activity of this lesson, students are given several different groups of dotplots and asked to determine the distinguishing feature that distinguishes each of the dotplots in a group. In this way, the students discover the characteristics of shape, center, and spread, and features such as clusters, gaps, and outliers. The activity also helps students see distributions as a single entity with identifiable characteristics. The second activity has students make predictions about graphs for a variable measured on their class survey and then make and test predictions about what would happen if the sample size were increased. Student learning goals for this lesson include:

1. To develop the idea of a distribution as a single entity rather than individual points.
2. To recognize different characteristics of a distribution and understand these characteristics in an intuitive, informal way.
3. To recognize differences between graphs of small and large samples, and how graphs of distributions stabilize as more data from the same population is added.
4. To develop an understanding of a density curve as it represents a population.

Description of the Lesson

The lesson begins with a discussion of the term "distribution" and how this differs in everyday usage and in statistics. Students reason about and discuss pattern, what we need to know to draw a reasonable graph without knowing the data values.

Next, in the *Distinguishing Distributions* activity, students are given a series of dotplots that depict the distributions of hypothetical exam scores in various classes. For example, they are asked:

> For classes A, B, and C, what is the main characteristic that distinguishes these three graphs from each other? What might explain this difference? (Fig. 8.8)

Each set of graphs reveals a different characteristic of distribution: e.g., center, spread, shape, and outliers. Students are then asked what features are important to examine when describing distributions, what to look for, features that are always present such as shape, center and spread, and ones that might be present or absent, depending on the data set.

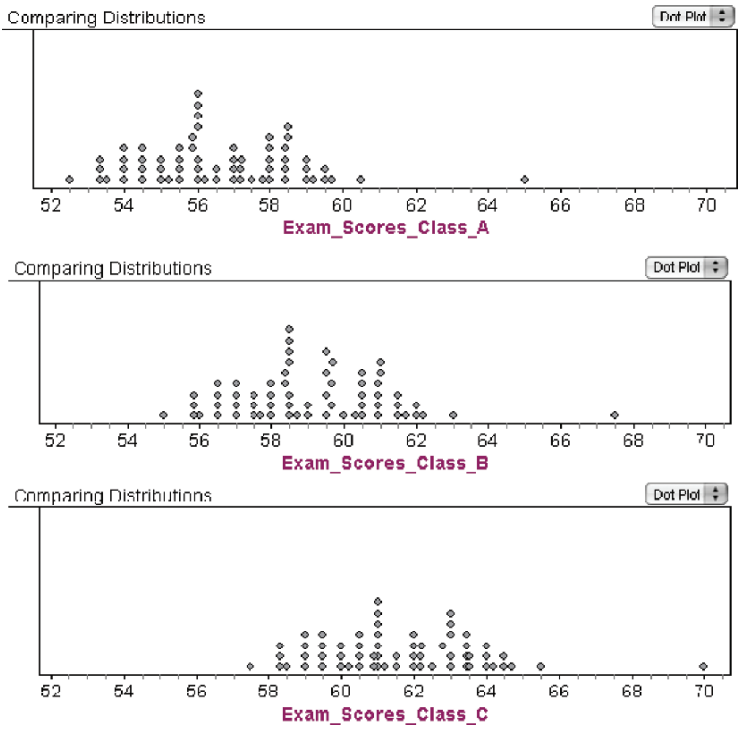


Fig. 8.8 Distributions of hypothetical exam scores in various classes

The following activity (*Growing a Distribution*) has students making predictions about number of hours students in their class study each week, both in terms of a typical value and a range of values. Next, they gather a sample of five data values from students in the class and graph this using a dotplot. More data are gathered (e.g., the entire class) and new dotplots are generated and described. Then students are asked to imagine if four classes of students were combined, and to draw their imagined plot of study hours per week. Finally, they consider all students at the university, and draw a smooth curve to represent this population. The students repeat this by looking at dotplots of three other variables based on student survey data, and then draw a smooth curve to represent the distribution of all students at their school for these variables.

A wrap-up discussion focuses on differences between dotplots and smooth curves. Students discuss what we mean by the term distribution, what are some of the common characteristics of a distribution of quantitative data, what information a graph of a distribution provides and what information a histogram provides that is not revealed by looking at a bar graph or case value graph (e.g., shape, center, and spread). The students consider when a histogram might be a better representation of data than a dot plot, what information can be determined by looking at a dotplot

more easily than in a histogram, and what information is lost or not shown by a histogram.

Lesson 2: Exploring and Sorting Distributions

This lesson begins with a data set of body measurements gathered from a set of students that includes kneeling heights. After students make some predictions about this variable, they examine and describe a dotplot of class data on kneeling heights. The students are led through a transformation of this dot plot into a histogram, using *TinkerPlots* software. They compare different representations of the same data to see how different features are hidden or revealed in different types of graphs.

In the third activity, students sort a set of histograms into different piles according to general shape, which leads students to recognize and label typical shapes, guiding them to see these distributions as entities, rather than as sets of individual values. The students further develop the idea of shape by changing bin widths on histograms using *Fathom* software to manipulate and reveal how the size of the bins used affects the stability of the shape. In the fifth and final activity, students match histograms to variable descriptions, reasoning about the connections between visual characteristics of distributions and variable contexts.

Student learning goals for this lesson include:

1. To understand how a distribution is represented by a histogram and that a histogram (or dotplot) allows us to describe shape, center, and spread of a quantitative variable in contrast to a bar graph (bar chart).
2. To understand the differences between case value graphs, bar graphs (case-value bars) of individual data values and graphs displaying distributions of data such as histograms.
3. To understand how graphical representations of data reveal the variability and shape of a data set.
4. To recognize and label typical shapes of distribution, using common statistical terms (normal, skewed, bimodal, and uniform).
5. To understand that the shape of a graph may seem different depending on the graphing technique used, so it may be important to manipulate a graph to see what the shape seems to be.

Description of the Lesson

In the *What is a Histogram* activity, students begin by making conjectures about what they would expect to see in the distribution of kneeling height data. This data set is then used to help students develop an understanding of a histogram, by using *TinkerPlots* to sort the data into sequential intervals, and then fuse the intervals into bars.

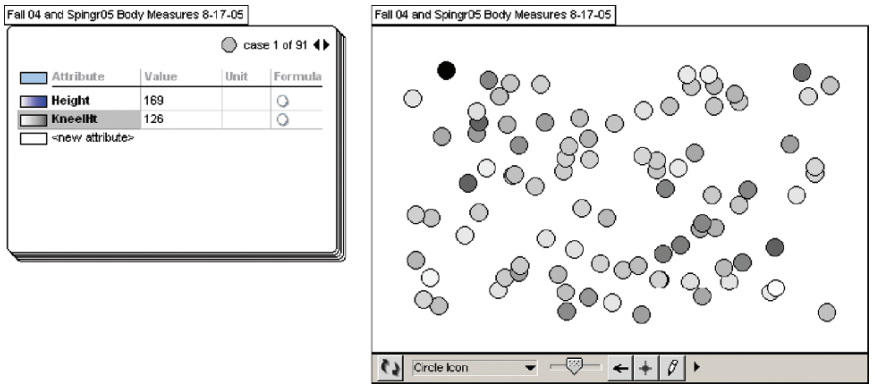


Fig. 8.9 Kneeling heights scattered dots in *TinkerPlots*

They begin with a graph of the data values, scattered as shown in Fig. 8.9, and after playing with and examining the data, are guided to arrange the dots into case-value bars, to sort the smaller values of kneeling heights as shown in Fig. 8.10.

Students are then guided to use the *Separate* operation in *TinkerPlots* to separate the cases into intervals that are then fused into a histogram, as shown in Figs. 8.11 and 8.12.

In the *Stretching Histograms* activity, students use a histogram applet to examine the effect of bin width on shape, seeing that larger and smaller bin widths may obscure shape and details (such as gaps, clusters, and outliers). Students are then encouraged to think about the difference between the different types of plots they have seen (*Exploring Different Representations of Data* activity). These plots

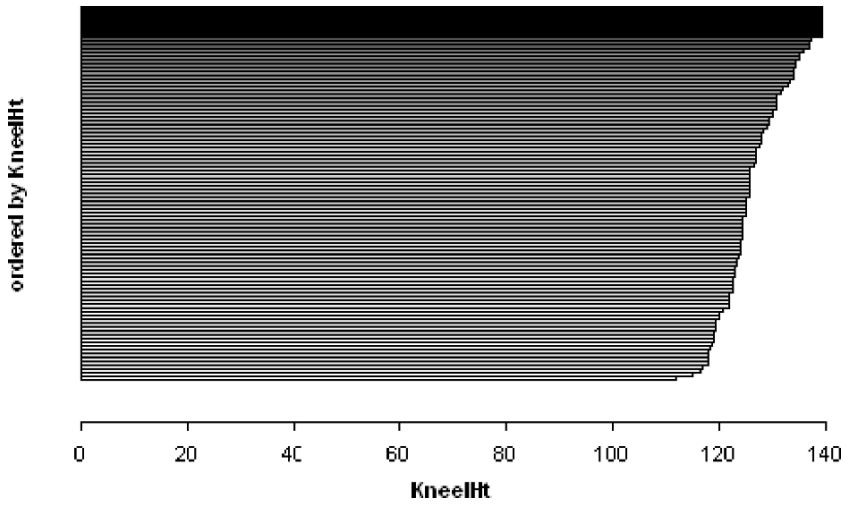


Fig. 8.10 A case-value graph ordered by kneeling heights

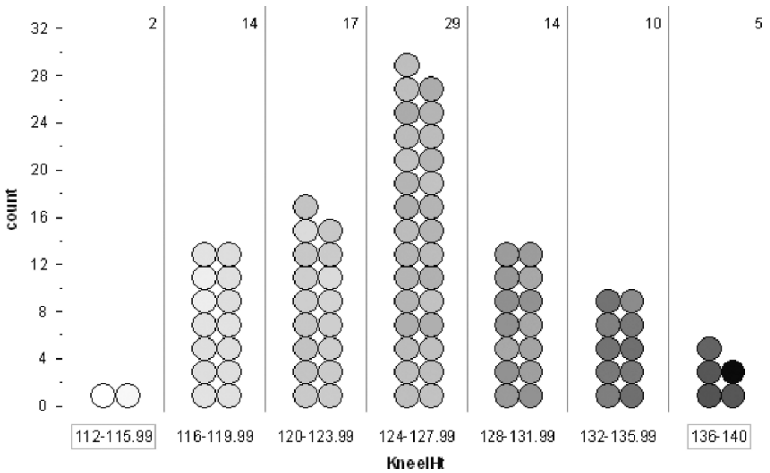


Fig. 8.11 Stacked kneeling height data with frequencies noted in each interval

include the dotplot, the case-value graph (value bar graph), and the histogram. They are then asked to consider which graphs help them better estimate the lowest five values, where "most" of the kneeling heights are clustered, the middle or typical value of kneeling heights, the spread, and the shape of the data. Students discuss how to determine the best representation to answer a particular question and why one representation may be better than another.

In the next activity (*Sorting Histogram*), students work in small groups to sort a set of 21 different graphs, representing different data sets. Students sort the stack of graphs into piles, according to those that look the same or similar, then describe what is similar about the graphs in each group, pick one representative graph for

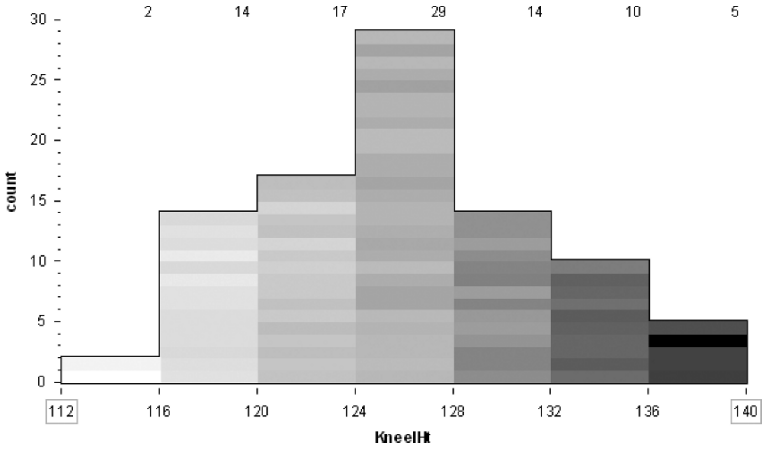


Fig. 8.12 A histogram presenting the kneeling height in seven bins

each group, and come up with their own term to describe the general shape of each group. A whole class discussion follows where group results are compared and then the correct statistical terms for the graphs (uniform, normal, right- and left-skewed, bimodal) can be introduced or reinforced. Models (uniform, normal) are described in terms of symmetry and shape (bell-shape or rectangular). Other distributions that do not fit these models can be described in terms of their characteristics (skewness, bimodality or unimodality, etc.). A discussion of which descriptors can and cannot go together may follow. For example, normal and skewed cannot go together.

During this discussion, some important points are addressed:

- Ideal shapes: density curves vs. histograms (theoretical vs. empirical distributions)
- Different versions of ideal shapes
- Idea of models, characteristics of distributions
- Statistical words vs. informal descriptors
- Other ways to describe a distribution
- Why is it important to describe a distribution?
- Normal, skewed, uniform, bimodal, and symmetric: which can be used together? How well do they fit the graphs? Which fit best?

Next, students work in groups to discuss and sketch what they expect for the general shape for some new data sets, and use the statistical terms to describe the shape of each. For example, the salaries of all persons employed at the University or grades on an easy test. There is another whole class discussion to compare answers and explanations.

The final activity (*Matching Histograms to Variable Descriptions*) has students match descriptions of variables to graphs of distributions, helping to develop students' reasoning about behavior of data in different contexts and how this is related to different types of shapes of graphs. A wrap-up discussion focuses on the characteristics of distributions and how they are revealed in different types of graphs.

Summary

Understanding the idea of distribution is an important first step for students who will encounter distributions of data and later distributions of sample statistics as they proceed through their statistics course. While most textbooks ask students to look at a histogram or stem plot and describe the shape, center, and spread, many students never understand the idea of distribution as an entity with characteristics that reveal important aspects of the variation of the data. The focus of the activities described in this chapter is on developing a conceptual understanding of distribution, and we have not included activities where students learn to construct different graphs, a topic well covered in most textbooks. We encourage instructors to repeatedly have students interpret and describe distributions as they move through the course, whether plots of sample data or distributions of sample statistics.