Chapter 7
Learning to Reason About Statistical Models and Modeling

One of the most overworked words in statistics education and mathematics education is "model." Appearing in a variety of dissimilar contexts, its usage is at best unclear, and at worst, inappropriate.

(Graham, 2006, p. 194)

Snapshot of a Research-Based Activity

Class begins with a discussion about the "One-Son Policy" that was proposed for families in China to keep the birth rate down, but to allow each family to have a son. This policy allowed a family to keep having children until a son was born, at which point no more children were allowed. Students are asked to speculate about what would happen to the ratio of boys to girls if this policy was introduced and about what they would predict the average family size to be. Most students think that this policy would result in more boys or a higher ratio of boys to girls. Others think it might result in more girls, because a family might have several girls before they have a boy. Some students think that the average number of children might also increase under this policy.

After students make and explain their conjectures, they discuss how to *model* this process so that they can *simulate* data to estimate what would be the ratio of boys to girls and average family size if the One-Son Policy were implemented. First, students are introduced to the idea of making small tokens labeled "Boy" and "Girl" to model the problem. They are guided to put equal numbers of these tokens in a container, assuming boy and girl babies are equally likely. The students then draw from tokens, one at a time from the container with replacement, writing down the outcomes, for example, B, GB, B, GGGB, etc.

The students are asked to consider other ways that they might model this problem and generate data, without labeling tokens "Girl" and "Boy", and use a coin. A suggestion is made to simulate data by tossing coins, with a Head representing a Girl and a Tail representing a Boy. The students discuss their assumptions for this simulation, such as, assuming the coin tosses are equally likely to land Tails up or Heads up, and that the result is unpredictable, a *random outcome*. Preliminary data

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

are gathered and examined, and students are surprised to see that the ratio of boys to girls is close to 1 and that the average family size is smaller than predicted. They predict what they think would happen if more data are gathered.

Students are then guided to use the *Fathom* software (Key Curriculum Press, 2006) to simulate larger data sets to answer this question. They discuss the appropriateness of the coins and *Fathom* simulations for modeling birth rates, and students bring in their own knowledge of boys being more likely to be born than girls, so that it is not exactly 50% boys and 50% girls. However, they note that the coin model was helpful in providing data that approximates real results and helps estimate an answer to the original questions about what would happen if the One-Son Policy were adopted. The students also see that it can be easier to simulate data by finding a useful *model* than to try to work out a complex probability problem.

Rationale for This Activity

The One-Son Modeling activity can take place on early in a class, even on the second day (after the first lesson plan for the topic of Data, from Chapter 6). We have introduced it this early in a course because of the importance of introducing the ideas of random outcome, model, and simulation. These ideas are interconnected and nicely illustrated in this first activity. The students can see that the results of a coin toss are random, but that repeated tosses yield predictable patterns (e.g., half Tails, half Heads). They see that coin tosses can be used as a model for birth rates, but that the model is not perfect, just useful. In other words, students see that statistical models can be useful for simulating data to answer real world questions, but they are not perfect fits to reality. They are also exposed to simulation of data as a way to examine chance phenomena, and this sets the stage for simulations as a helpful process that is used throughout the course. The research basis for this activity comes from some of the research on understanding models and modeling described in this chapter that suggests the need for students to create simple models for chance events, and to use chance devices such as coins or cards to simulate data before having a computer produce larger amounts of simulated data. We have added on the importance of using language about models and drawing students' attention to the fact that what they are doing is creating a model to represent a problem and using the model to produce data to solve the problem.

The Importance of Understanding Statistical Models and Modeling

> It is the job of statisticians to represent the data taken from the real world with theoretical models.
>
> (Graham, 2006, p. 204)

Statisticians use models in different ways, and some of these uses appear in introductory statistics courses. Two main uses of statistical models are (see schematic illustration in Fig. 7.1):

1. Select or design and use appropriate *models* to simulate data to answer a research question. Sometimes, the model is as simple as a random device, sometimes it takes the form of a statement (such as a null hypothesis) that is used to generate data to determine if a particular sample result would be surprising if due to chance. As George Cobb states: Reject any model that puts your data (the investigated sample) in its tail (2005). Sometime a data set is used to simulate (bootstrap) more data, creating a simulated population distribution to use in making statistical inferences.
2. Fit a statistical model to existing data or data that you have collected through survey or experiment in order to explain and describe the variability. This may be as simple as fitting the normal distribution to a data set, particularly when checking assumptions when using a particular procedure. It may involve fitting a linear model to a bivariate data set to help describe and explain a relationship between two variables. In advanced courses, data modeling is an essential technique used to explore relationships between multiple variables. In all of these cases, we examine the fit of a model to data by looking at deviations of the data from the model.

Figure 7.1 shows the commonalities and differences between these two uses of statistical models as well as their role in the ongoing cycle of statistical investigative work.

Why should students in introductory statistics class need to learn about statistical models? Because models are a foundational part of statistical thinking, working with models is a big element of the real work of statisticians. It is surprising; therefore, that little explicit attention is paid to the use of models in most introductory courses. The words "model" or "modeling" hardly appear in most introductory statistical textbooks. When these words are used, it is often in the context of a linear model for bivariate data or the normal distribution as a mathematical model. Sometimes, the term is used in connection with probability, as in probability models. Although many statisticians talk about the importance of data modeling and fitting models to data, most students can take an introductory statistics course and never understand what a statistical model is or how it is used.

The models statisticians use are actually mathematical models. David Moore (1990) describes the role of mathematical models in the data analysis process: "Move from graphic display to numerical measures of specific aspects of the data to compact mathematical models for the overall pattern" (Page 104). The two most commonly used mathematical models in an introductory statistics class are the normal distribution and the linear regression models. Although these two topics involve the use of mathematical models that are fitted to sample data, students rarely see these topics as related, or as examples of two different kinds of statistical models that help us analyze data.

The normal curve, which is perfectly symmetric and does not reflect the irregularities of a data set, is an idealized model that nicely fits many distributions of real data, such as measurements and test scores. Moore (1990) comments that moving from observations of data to an idealized description (model) is a substantial

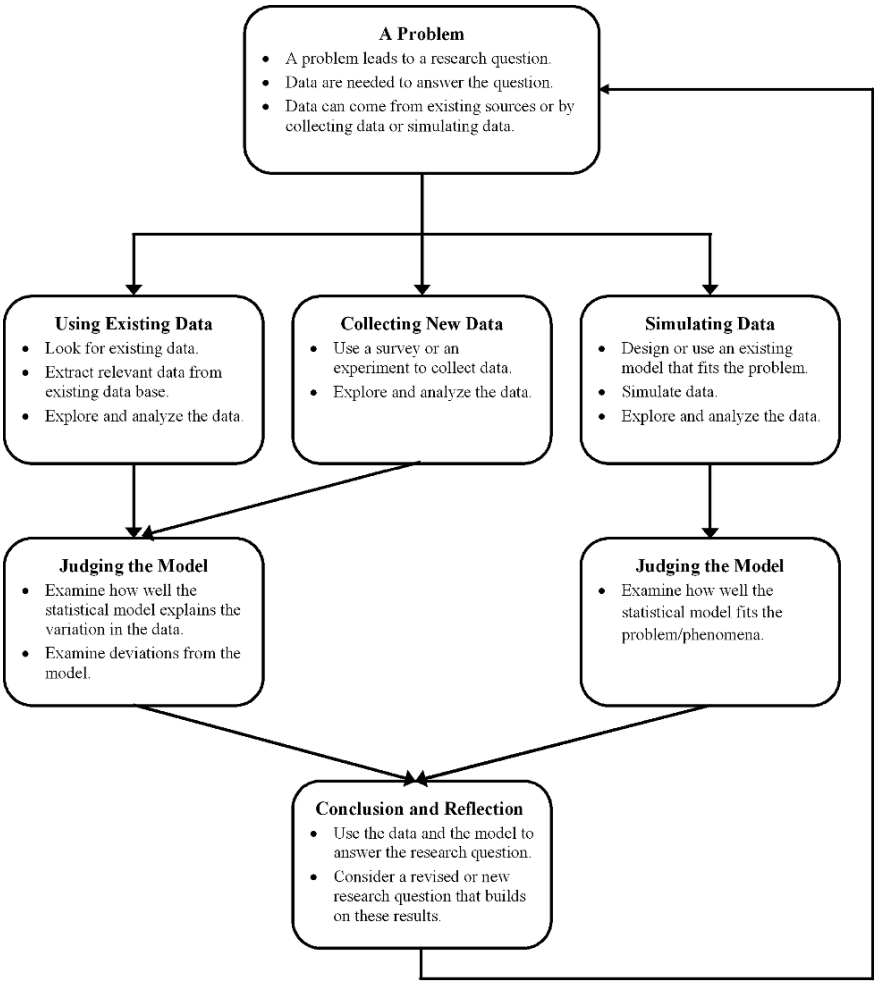

Fig. 7.1 Uses of models in statistical analysis

abstraction, and the use of models such as the normal distribution and the uniform distribution is a major step in understanding the power of statistics. Some software packages, such as *Fathom*, allow students to superimpose a model of the normal distribution on a data set. This feature helps students judge the degree of how well the model fits the data and develops students' understanding of the model fitting process.

The Place of Statistical Models and Modeling in the Curriculum

We find the lack of explicit attention to statistical models (other than probability models, in a mathematical statistics class) surprising. While there are examples of

ways to model probability problems using concrete materials and or simulation tools (see Simon, 1994; Konold, 1994b), these do not appear to be part of most introductory statistics classes, and do not appear to be part of an introduction to the use of models and modeling in statistics. We have thought carefully about how to incorporate lessons on statistical models and modeling into an introductory course. Rather than treat this topic as a separate unit, we think that activities that help students develop an understanding of the idea and uses of a statistical model should be embedded throughout the course, with connections made between these activities.

The One-Son Modeling activity described earlier is a good way to introduce the related ideas of model, random outcome, and simulation. In the following unit on data (see Lesson 4, Chapter 6), we revisit the idea of modeling and simulation after students conduct a taste test and want to compare their results to what they would expect due to chance or guessing (the null model). The normal distribution is informally introduced as a model in the unit on distribution (Chapter 8) and revisited in the units on center (Chapter 9), variability (Chapter 10), and comparing groups (Chapter 11). After completing the topics in data analysis, the topic of probability distribution can be examined as a type of distribution based on a model. The normal distribution is then introduced as a formal statistical model (probability distribution) and as a precursor to the sampling unit (this activity is described in the end of this chapter). The sampling unit (Chapter 12) revisits the normal distribution as a model for sampling distributions. In the unit on statistical inference (Chapter 13), models are used to simulate data to test hypotheses and generate confidence intervals. Here a model is a theoretical population with specified parameters. Statistical models are used to find P-values if necessary conditions are met. The final model introduced after the unit on statistical inference is the regression line as a model of a linear relationship between two quantitative variables (see Chapter 14). This model is also tested by using methods of statistical inference and examining deviations (residuals). We find that the idea and use of a statistical model is explicitly linked to ideas of probability and often to the process of simulation. Therefore, we briefly discuss these related topics as well in this chapter.

Review of the Literature Related to Reasoning About Statistical Models and Modeling

> All models are wrong, but some are useful.
>
> (George Box, 1979, p. 202)

Models in Mathematics Education

Several researchers in mathematics education have applied mathematical modeling ideas to data analysis (e.g., Horvath & Lehrer, 1998). Lehrer and Schauble (2004) tracked the development of student thinking about natural variation as elementary grade students learned about distribution in the context of modeling plant growth at the population level. They found that the data-modeling approach assisted children

in coordinating their understanding of particular cases with an evolving notion of data as an aggregate of cases. In another study by the same researchers, four forms of models and related "modeling practices" were identified that relate to developing model-based reasoning in young students (Lehrer & Schauble, 2000). They found that studying students' data modeling, in the sense of the inquiry cycle, provided feedback about student thinking that can guide teaching decisions, an important dividend for improving professional practice.

A related instructional design heuristic called "emergent modeling" is discussed by Gravemeijer (2002) that provides an instructional sequence on data analysis as an example. The "emergent modeling" approach was an alternative to instructional approaches that focus on teaching ready-made representations. Within the "emergent modeling" perspective, the model and the situation modeled are mutually constituted in the course of modeling activity. This gives the label "emergent" a dual meaning. It refers to both the process by which models emerge and the process by which these models support the emergence of more formal mathematical knowledge.

Models in Statistical Thinking

> Statisticians . . . have a choice of whether to access their data from the real world or from a model of the real world.
>
> (Graham, 2006, p. 204)

How students understand and reason about models and modeling processes has received surprisingly little attention in statistics education literature. This is surprising since statistical models play an important part in statistical thinking. The quote by Box, "All models are wrong, but some are useful" (1979, p. 202), is a guiding principle in formulating and interpreting statistical models, acknowledging that they are ideal and rarely match precisely real life data. The usefulness of a statistical model is dependent on the extent that a model is helpful in explaining the variability in the data.

Statistical models have an important role in the foundations of statistical thinking. This is evident in a study of practicing statisticians' ways of thinking (Wild & Pfannkuch, 1999). In their proposed four-dimensional framework for statistical thinking, "reasoning with statistical models" is considered as a general type of thinking, as well as specific "statistical" type of thinking, which relates, for example, to measuring and modeling variability for the purpose of prediction, explanation, or control. The predominant statistical models are those developed for the analysis of data.

While the term "statistical models" is often interpreted as meaning regression models or time-series models, Wild and Pfannkuch (1999) consider even much simpler tools such as statistical graphs as statistical models since they are statistical ways of representing and thinking about reality. These models enable us to summarize data in multiple ways depending on the nature of the data. For example, graphs, centers, spreads, clusters, outliers, residuals, confidence intervals, and P-values are

read, interpreted, and reasoned with in an attempt to find evidence on which to base a judgment.

Moore (1999) describes the role of models to describe a pattern in data analysis as the final step in a four-stage process.

> When you first examine a set of data, (1) begin by graphing the data and interpreting what you see; (2) look for overall patterns and for striking deviations from those patterns, and seek explanations in the problem context; (3) based on examination of the data, choose appropriate numerical descriptions of specific aspects; (4) if the overall pattern is sufficiently regular, seek a compact mathematical model for that pattern (p. 251).

Mallows (1998) claims that too often students studying statistics start from a particular model, assuming the model is correct, rather than learning to choose and fit models to data. Wild and Pfannkuch (1999) add that we do not teach enough of the mapping between the context and the models. Chance (2002) points out that, particularly, in courses for beginning students, these issues are quite relevant and often more of interest to the student, and the "natural inclination to question studies should be rewarded and further developed."

Reasoning About a Statistical Model: Normal Distribution

There is little research investigating students' understanding of the normal distribution, and most of these studies examine isolated aspects in the understanding of this concept. The first pioneering work was carried out by Piaget and Inhelder (1951, 1975), who studied children's spontaneous development of the idea of stochastic convergence. The authors analyzed children's perception of the progressive regularity in the pattern of sand falling through a small hole (in the Galton apparatus or in a sand clock). They considered that children need to grasp the symmetry of all the possible sand paths falling through the hole, the probability equivalence between the symmetrical trajectory, the spread and the role of replication, before they are able to predict the final regularity that produces a bell-shaped (normal) distribution. This understanding takes place in the "formal operations" stage (13- to 14-year-olds).

In a study of college students' conceptions about normal standard scores, Huck, Cross, and Clark (1986) identified two misconceptions: On the one hand, some students believe that all standard scores will always range between -3 and $+3$, while other students think there is no restriction on the maximum and minimum values in these scores. Others have examined people's behavior when solving problems involving the normal distribution (Wilensky, 1995, 1997). In interviews with students and professionals with statistical knowledge, Wilensky asked them to solve a problem by using computer simulation. Although most subjects in his research could solve problems related to the normal distribution, they were unable to justify the use of the normal distribution instead of another concept or distribution, and showed a high "epistemological anxiety," the feeling of confusion and indecision that students experience when faced with the different paths for solving a problem.

In recent empirical research on understanding the normal distribution, Batanero, Tauber, and Sánchez (2004) studied students' reasoning about the normal distribution in a university-level introductory computer-assisted course. While the analysis suggests that many students were able to correctly identify several elements in the meaning of normal distribution and to relate one to another, numerous difficulties understanding normal distributions were identified and described. The main conclusion in this study is that the normal distribution is a very complex idea that requires the integration and relation of many different statistical concepts and ideas. The authors recommend the use of appropriate activities and computer tools to facilitate the learning of basic notions about normal distributions (see causeweb.org for some of these resources).

Understanding Ideas Related to Probability Models

The research reviewed in Chapter 2 along with literature reviews by Falk and Konold (1998); Shaughnessy (2003); Jones (2005) and others illustrate the conceptual difficulties students have in understanding basic ideas of probability such as randomness. For example, Falk & Konold, (1994, 1997) found that people attempting to generate random number sequences usually produce more alternations of heads and tails than expected by chance. A related research result is that students tend to think that all models of random events are ones with equally likely outcomes. Lecoutre (1992) refers to this misconception as the "equiprobability bias," which she found students to use in solving different types of probability problems.

Reasoning About the Use of Models to Simulate Data

Studies on the use of simulations cover many different topics, such as how students and teachers understand statistical models in simulating data (e.g., Sánchez, 2002), the use of simulation to illustrate abstract concepts such as sampling distributions (e.g., Saldanha & Thompson, 2002) and learning to formulate and evaluate inferences by simulating data (e.g., Stohl & Tarr, 2002). While there is a strong belief that physical simulations should precede computer simulations, this has not yet been a topic of empirical study. Helpful guidelines for secondary school students on how to select models to simulate problems are included in the Art and Techniques of Simulation – a volume in the Quantitative Literacy Series (Gnanadesikan, Scheaffer, & Swift, 1987).

Biehler (1991) has presented an extensive analysis of the capabilities and limitations of simulation techniques in teaching statistics; he points out that "the different roles, goals and pedagogical perspectives for simulations have not yet been clearly analyzed and distinguished." He suggests a basic distinction between "the use of simulating as a method for solving problems, similar to the professional use outside school and the use of simulation to provide model environments to explore, which compensate for the 'lack of experience" (p. 183).

To study teachers' opinions about the instructional use of models in simulating data, Sánchez (2002) interviewed six high school teachers who participated in a workshop of simulation activities using *Fathom*. The analysis of their responses included four general aspects: the role of simulation in teaching; the different steps to follow in a simulation; the complexity of starting situations; and the statistical concepts that take part in simulation activities. The results show that teachers deem as important only certain aspects of simulation, but neglect the fundamental concepts of randomness and distribution. Sánchez noted that the teachers have centered their attention in certain modeling aspects like formulation of a model and its simulation, but neglected other aspects like the analysis of results and the validation of the model.

Implications of the Research: Teaching Students to Reason About Statistical Models and Modeling

Other than some literature on reasoning about the normal distribution or a linear relationship in bivariate data, there is little research illuminating how students come to learn and use statistical models. Therefore, we are speculative in putting together a research-based sequence of activities for this topic.

The literature reviewed implies that it is important to make models an explicit topic in the introductory statistics course, and to help students develop this idea and its multiple meanings and uses through experiences with real statistical problems and data, rather than through a formal study of probability. The literature also suggests that understanding the idea of randomness is difficult and that carefully designed activities should be used to help students understand the ideas of random outcomes, random variables, random sampling, and randomization.

There are also implications from the literature about the role of probability in an introductory college statistics course. It is suggested that in order for students to develop a basic understanding of statistics in a first college course, they only need the basic ideas of probability that were introduced above. As Garfield and Ahlgren wrote in 1988 "useful ideas of inference can be taught independently of technically correct probability." Therefore, we propose helping students understand the ideas of random outcomes and a probability distribution. They do not need to learn the language and laws of probability in such a course. Indeed, research shows that even if students encounter these topics in probability in an introductory statistics course, few students understand and can reason about this topic (see reviews by Garfield & Ahlgren, 1988; Hawkins & Kapadia, 1984; Shaughnessy, 2003).

We agree with David Moore's (1997) claim that mathematical probability is a "noble and useful subject" and should be part of the advanced coursework in statistics, and instead of being part of the introductory course, should be learned in a separate course that is devoted to this topic. The literature also implies that in order to develop a deep understanding of the basic ideas outlined in Chapter 3, the traditional course should be streamlined. Following the lead of Moore (1997),

our candidate for the guillotine is formal probability. Moore recommends that an informal introduction to probability is all that is needed and that this begins with experience with chance behavior, usually starting with physical devices and moving to computer simulations that help demonstrate the fundamental ideas such as the Law of Large Numbers.

Technological Tools to Help Students Develop Reasoning About Models

Many Web applets are available to help students see and use the normal distribution (e.g., http://www.rossmanchance.com/applets/NormalCalcs/ NormalCalculations.html) and fit a line to data (e.g., http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html). Applets can also be used to simulate data for a probability problem (see rossmanchance.com). The simulation tool *Probability Explorer* (Stohl, 1999–2005; http://www.probexplorer.com/) enables school students and teachers to design, simulate, and analyze a variety of probabilistic situations. Model Chance is a new program that is now being developed as part of the *TinkerPlots* project, to help student model problems and simulate data to estimate answers to these problems (http://cts.stat.ucla.edu/projects/info.php?id=4).

Progression of Ideas: Teaching Students to Reason About Statistical Models and Modeling

Introduction to a Sequence of Activities to Develop Reasoning about Statistical Models and Modeling

While most textbooks do not introduce the term "model" until the normal distribution and may not use it again until regression, we believe that the term should be introduced early in a course and used frequently, to demonstrate and explain how models are used in statistical work. Since there is no empirical research on an optimal approach for helping students develop the idea of model in an introductory statistics class, we offer one of several possible sequences of activities that we have found useful in our courses. These activities develop the idea of a model to simulate data, and the idea of a model to fit to a given set of data.

We believe that the idea of a model can be presented informally in the first few days of class using a fairly simple context. Physical simulations using devices such as coins can be used to model a random variable as part of solving a problem, and once data have been simulated using coin tosses, the computer can be used to quickly produce large amounts of simulated data. It is important that students be made aware of the use of a model, to represent key features of the event and to produce simulated data.

In the Taste Test activity in the Data topic (see Chapter 6), when the notion of an experiment is introduced, the idea of a null model (what would happen if results were only due to chance) can be reintroduced. In this context, the model is used to simulate data to informally determine a P-value, indicating whether or not an individual's identification of soda brand in a blind taste test may not just be due to chance. The formal idea of model may be encountered in the context of random variables. Students can model random outcomes such as coin tosses, dice throwing, and drawing of cards. Each time students can describe the model, e.g., for a standard deck of Poker cards, where the outcome is a Heart, this can be modeled as a binomial variable with a probability of $1/4$. These models can be used to simulate data, which leads to examining probability distributions. Different probability problems can also be modeled first using coins or dice, and then using applets or software to simulate data. In this case, the problem is modeled by the coins or dice, and then this model is replicated by a technology tool. After generating empirical probability distributions for binomial random variables, the probability distribution for the normal distribution can be introduced as a model that is often used to describe and interpret real data. This model is seen again as students begin to examine distributions of sample means and the Central Limit Theorem. Null models are encountered again in the unit of inference to generate distributions of sample statistics to run tests of significance. In our proposed sequence of activities, the final model presented is that of the linear model, used in the unit on covariation, to model the relationship of two quantitative variables.

What is unique about the lessons in this chapter is that unlike all the other chapters in part II of this book, we are not offering a unit on models but rather providing several activities that illustrate and use the idea of statistical models throughout an introductory course. While the informal ideas of model are introduced at the beginning of a course, the formal ideas are encountered midway through the course, and then revisited at additional times in other topic areas, rather than in a set of sequential lessons in a unit of their own. Table 7.1 shows a suggested series of ideas and activities that can be used to guide the development of students' reasoning about models and modeling.

Introduction to the Lessons

Three lessons are designed to help students develop an understanding of the importance and use of statistical models. Other activities involving models are interspersed throughout other topics in this book. The first lesson has students model birth outcomes in a situation where there is a "One-Son Policy." A probability model is used to simulate data that is summarized to answer some research questions. The second activity in this lesson has students use first a physical and then a computer simulation to model the "Let's Make a Deal" game, using the simulated data to find the best strategy for playing the game. The second lesson has students create binomial models for different random devices (coins, dice, and cards) and use these

Table 7.1 Sequence of activities to develop reasoning about statistical models and modeling[1]

Milestones: ideas and concepts	Suggested activities
Informal ideas prior to formal study of statistical models	
• Models can be used to portray simple random outcomes. Random devices and computers can be used to simulate data to answer a question about this context	• One-Son Modeling Activity (Lesson 1: "Using Models to Simulate Data")
• A random outcome, unpredictable, but giving a predictable pattern over the long run. The more data there is, the more stable is the pattern	• Let's Make a Deal Simulation (Lesson 1)
• Designing and using a model can help to answer a statistical question	• Let's Make a Deal Simulation (Lesson 1)
• The idea and importance of random samples (revealing the predictable pattern of random outcomes)	• The Gettysburg Address Activity (Lesson 3, Data Unit, Chapter 6)
• Models can be used to generate data to informally test an experimental result to provide evidence about whether or not this result is due to chance	• Taste Test Activity (Lesson 4, Data Unit, Chapter 6)
• Distinguish between the model, the simulated data, and the sample data	• Taste Test Activity (Lesson 4, Data Unit, Chapter 6)
• The normal distribution as a model for some distributions of real world data	• Sorting Histograms Activity (Lesson 2, Distribution Unit, Chapter 8)
• The mean is a good summary of the center of a normal distribution	• Choosing an Appropriate Measure of Center Activity (Lesson 2, Center Unit, Chapter 9)
• The mean and standard deviation are good summaries for a normal distribution	• How do Students Spend their Time Activity (Lesson 4, Comparing Groups Unit, Chapter 11)
Formal ideas of statistical models	
• Random variables and random outcomes	• Coins, Cards, and Dice Activity (Lesson 2: "Modeling Random Variables")
• Equally likely model does not fit all random outcomes	• Coins, Cards, and Dice Activity (Lesson 2)
• A probability distribution as a model	• Coins, Cards, and Dice Activity (Lesson 2)
• Probability problems can be modeled using random devices and simulation tools	❖ Activity where cards are used to model a problem, such as Random Babies activity in Chance and Rossman (2006) (The symbol ❖ indicates that this activity is not included in these lessons.)

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

Table 7.1 (continued)

• Characteristics of normal distribution as a model	• What is Normal? (Lesson 3: "The Normal Distribution as a Model")
• What does normal data look like?	• What is Normal? (Lesson 3)
• Using the normal distribution as a Model	• Normal Distribution Applications (Lesson 3)
Building on formal ideas of models in subsequent topics	
• How and why the sampling distribution of means can be modeled by the normal distribution	• Central Limit Theorem Activity (Lesson 3, Samples and Sampling Unit, Chapter 12)
• The null hypothesis as model to which we compare sample data	• Balancing Coins Activity (Lesson 1, Statistical Inference Unit, Chapter 13)
• When testing a hypothesis, it is often important to check the condition of normality of the sampling distribution	• Research Questions Involving Statistical Methods (Lesson 5, Statistical Inference Unit, Chapter 13)
• The regression line is a useful model of bivariate relationships between quantitative variables	• Diamond Rings Activity (Lesson 2, Covariation Unit, Chapter 14)
• Checking the fit of a model to data, by examining residuals from a regression line	• da Vinci and Body Measurements Activity (Lesson 2, Covariation Unit, Chapter 14)

to generate and then simulate data (on the computer) to examine and compare probability distributions (each having a different shape and expected value). The third lesson introduces the normal distribution as a model of a probability distribution. This model is used to fit to samples of data (e.g., do the data sets appear to have a normal distribution?) and to demonstrate when it is appropriate to use this model in analyzing data.

Lesson 1: Choosing Models to Simulate Data

The first lesson begins with the *One-Son Modeling* activity described at the beginning of this chapter. Next, students consider the chances of wining on the game show *Let's Make a Deal* to determine whether one strategy has a higher chance of wining than another, modeling the game first with cards and then with a Web applet. The student learning goals for this lesson include:

1. Be able to use simulation as a tool for answering statistical questions
2. Be able to develop models to simulate data
3. Examine probability as an indication of how likely is an event to happen.
4. Realize that their intuitions about probabilities may be misleading.

Description of the Lesson

The students are told about the "One-Son policy" that was proposed for China as a way to decrease their birthrate, but still allow families to produce a son.

> In the early 1990's China considered adopting a "One-Son Policy", to help reduce their birthrate by allowing families to keep having children until they had a son. Under this plan a family has a child. If it is a son, they stop having children. If it is a daughter, they can try again. They can keep trying until they have a son and then they stop having children.

The following questions are posed to students to engage them in reasoning:

1. *If a country adopted a policy that let families have children only until they had a boy, and then they had to stop, what would you expect to happen?*
2. *What would the average number of children per family be? Would you expect more boys or more girls overall?*

Students present their answers and reasoning. Simulation is introduced as a tool statisticians use to generate data to estimate an answer to this type of question.

Working in pairs, students take a yogurt container that contains two slips of paper. One is labeled "B" for boys. One is labeled "G" for Girls. They randomly draw one slip of paper from the container and that will be the first child born in a simulated family. If they draw a "B", then they are told to stop; the family is done having children. They enter the data on a chart indicating the result for the first family. If the result is a "G", they draw again. They keep drawing until they draw a "B", then they stop and enter the data on the chart for that family. Students repeat this process for five simulated families, getting results such as: GB, B, GGB, B, GB, etc.

After this first round of collecting simulated data, the number of children in each family is counted along with the number of girls and the number of boys. Students examine this small set of data and are asked if the results confirm their original predictions. Next, they consider what they would get if they did not have slips of paper, but instead only had a coin. Students figure out how to do this same simulation using only one coin that is tossed. They discuss and write out the process so that another group could run the simulation by following their directions. Students usually determine that one outcome (heads) will represent a Boy and the other outcome will represent a Girl. They describe how to repeatedly toss a coin until it lands heads, recording the data each time to simulate families. Next they simulate this data using the coin, generating data for five more families.

Now student groups have data for 10 simulated families based on the two sets of simulations. They tally the total number of girls, the total number of boys, find the ratio of boys to girls, and find the average number of children per family. Again, they compare these results to their initial conjectures. These results are also shared with the class and graphed on the board.

Next, students are asked what they would expect to find if they repeated this simulation many more times. They are shown how to use the *Fathom* to run this simulation and to gather data for more simulated families, as shown in Fig. 7.2.

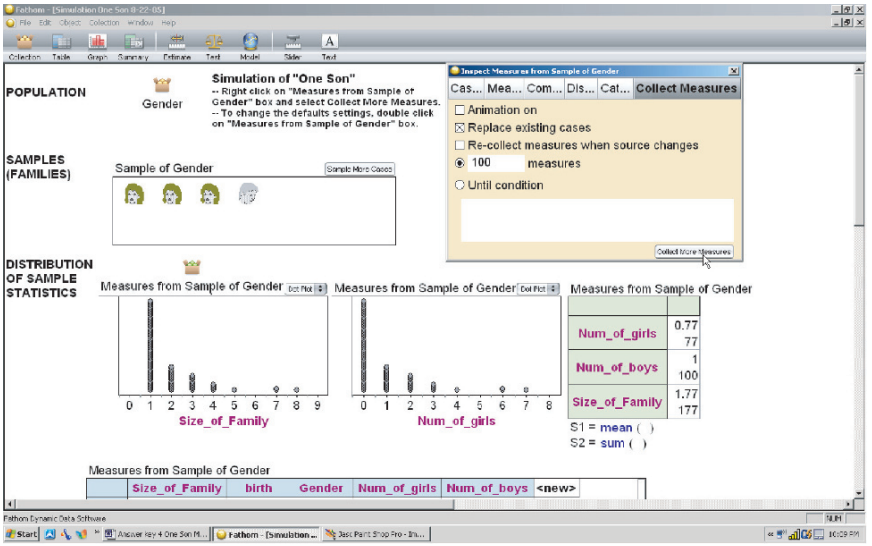

Fig. 7.2 Running the One-Son simulation in *Fathom*

A discussion of these data reveals that the results tend to stabilize as more data are collected. Students are asked what they would expect if more data were generated, and this leads to an informal discussion of the Law of Large Numbers. Next, a discussion of choosing models to simulate data leads students to consider real life factors to consider using a coin or an equally likely binomial model to simulate births. They suggest ways in which the model of equally likely outcomes may not perfectly fit the data, such as the fact that the probability of having a boy is actually more than .5. They consider how these factors could affect simulated results. A wrap-up discussion includes issues involved in selecting and using models (first the coin, and then the computer simulation of a binomial model with $p = .5$) to simulate a real world phenomenon (birth rates of boys and girls).

Students are next introduced to an activity (*Let's Make a Deal Simulation*), where they will check their intuitions about chance events by using simulation to determine probabilities in a game show setting. *Let's Make A Deal* is introduced as a popular TV game show from the 1970s.

> The task: *Suppose you are given three doors to choose from. Behind one door there is a big prize (a car) and behind the other two, there are goats. Only Monty Hall knows which door has the prize. You are asked to select a door, and then Monty opens a different door, showing you a goat behind it. Then you are asked the big question: Do you want to stay with your original door, or switch to a different door? What would you do?*

Students discuss in small groups whether they think there is a higher chance of winning the prize if they stay with their first door selection or should they switch to the remaining door and why. A show of hands in the class indicates that most students think that either the contestant should stay with the first choice or that

it does not make a difference; staying or switching are equally likely to result in winning a prize.

Next, students think about how they could test these conjecture. They design a simulated game using index cards. Each group is given three index cards, and they write lightly on each one of the following outcomes: "Goat", "Goat", and "Car". One person is the game host (Monty Hall) and knows where the car is when the cards placed face down in order. The other group members take turns being contestants. A recording sheet is used to write down what strategy was used each time (stay or switch doors) and what the outcome was ("Win" or "Lose"). After a few rounds, the data are pooled for the class, and it appears that "switch" is resulting in more wins. But is that just due to chance or a stable result? Students then use a Web applet that simulates this problem to quickly simulate large amounts of data, revealing that the chance of winning when they switch doors is about 2/3, and the chance of winning when they stay with their first choice about 1/3. Students are asked about the correctness of their original intuitions and why they were incorrect. A quick explanation may be given about why this happens, or students may be given a written explanation about why it pays to switch doors when playing this game.

The *Let's Make a Deal Simulation* activity concludes with a discussion about the use of models and simulations to easily generate data to estimate answers to statistical questions, and when to trust results of simulations (e.g., when the model is a good fit to the problem and when there are enough simulations to generate a stable result).

Lesson 2: Modeling Random Variables

This lesson engages students with modeling random variables using coins, cards, and dice. Students construct probability distributions to represent each of the scenarios and make predictions on probabilities of other events based on these histograms. They first generate data with concrete materials and then move to *Fathom* to simulate larger data sets and see the stable trend emerge. Student learning goals for this lesson include:

1. Understand use of models to represent random variables and simulate data.
2. Understand how to interpret visual representations for probability.
3. Use a simulation to generate data to estimate probabilities
4. Gain an informal understanding of probability distributions as a distribution with shape, center, and spread.

Description of the Lesson

Students begin by considering different random devices they have encountered, such as coins, cards, and dice (the *Coins, Cards, and Dice* activity). Then they make conjectures about the expected number of:

- *Heads on five tosses of a fair coin*?
- *Hearts in five draws from a poker deck*?
- *Twos in five rolls of a fair die*?

They then reason about what is the same and what is different about these three experiments. Next, students are led to define three random variables as follows:

X: Number of heads in five tosses
Y: Number of hearts in five cards dealt
Z: Number of 2's in five rolls

They now create models of each variable to generate (by using the actual devices) and then simulate data using the computer, which can be graphed and summarized to compare the distributions for each random variable.

Students begin with the Coin variable. They toss a coin five times and count the number of heads and then repeat this 10 times, making a frequency distribution for the number of heads that show up on each toss of five coins. Each time they toss the five coins, they check the value of X (Number of heads in five tosses) in the table, then find the relative frequency probability for each value of X. Next, students open a *Fathom* file that generates data based on this model of equally likely outcomes as shown in Fig. 7.3.

Students are asked how the simulated data compare to the data they generated by physically tossing coins, and if they expect these results to be similar for all students in the class. This leads to a discussion of the idea of two aspects of a random outcome: (1) that an individual outcome is unpredictable but (2) that you can predict patterns of outcomes for many repetitions, such as the proportion of

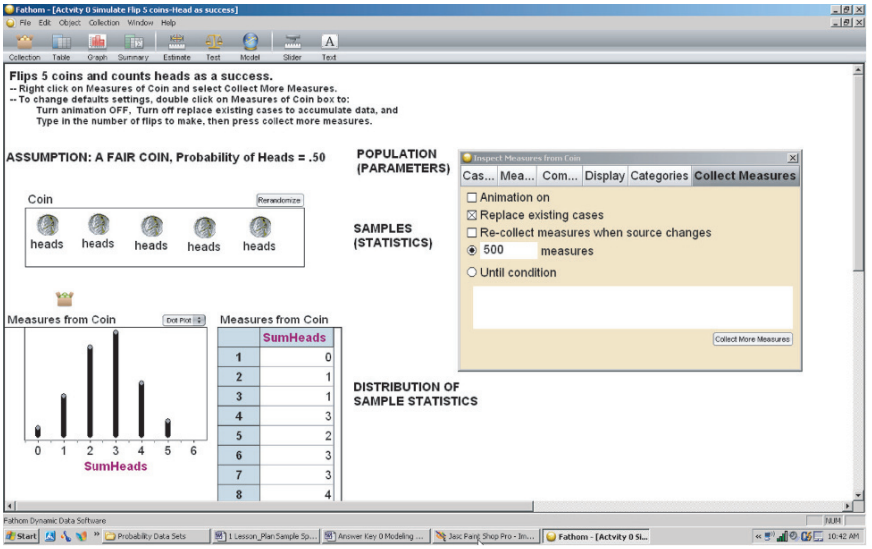

Fig. 7.3 Simulating the number of heads in five tosses of a fair coin in *Fathom*

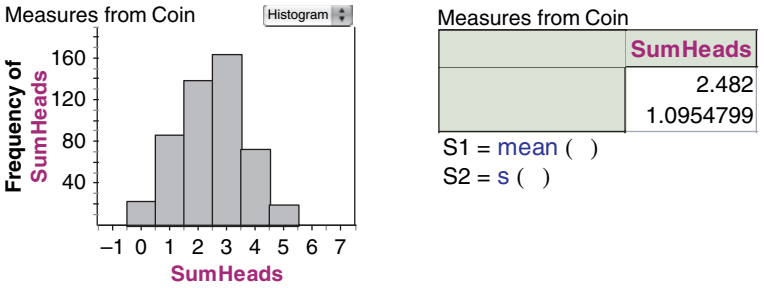

Fig. 7.4 The results of simulating the number of heads in five tosses of a fair coin in *Fathom*

heads for many tosses of a fair coin. The simulation is repeated for 500 trials, and relative frequencies are found for the six possible outcomes (0–5 heads). Students again discuss whether they think the results will vary from student to student and see that the larger sample size has much less variability. They graph the data and informally describe shape, center, and spread. They see that the expected number of heads for five tosses would be 2–3 heads, but that results can vary from 0 to 5 heads. They can generate the mean and standard deviation and interpret these values for the distribution (see Fig. 7.4).

The next part of the *Coins, Cards, and Dice* activity has students take a deck of cards, shuffle, draw a card, replace it, draw, and replace, five times. They count the number of hearts that showed up in their sample of five draws (with replacement). A discussion of the term "replacement" challenges students to reason about why they would replace the card drawn each time and how that would affect the results of the experiment so it is not similar to the coin tosses. This is an informal introduction to the idea of *independent events* without going into probability theory. Next, students use *Fathom* to simulate this experiment. They discuss what the model will be, that it is no longer one of equally likely outcomes, but that the chance of getting one heart when randomly drawn from a deck is now one-fourth. The students run a new simulation based on a binomial model with $p = .25$, graph the results, and find measures of center and spread, which are different from those in the Coin example (Fig. 7.5).

This activity is repeated a third time for the Dice example. First, students roll a dice five times counting the number of 2s that show up. They record the data and compare it with the class. Then they simulate the data on *Fathom*, first discussing how the model must be changed, based on the new probability of getting a two, which is one-sixth.

The results of the three experiments are compared. Students consider what was the same and what was different across the three experiments and how the differences in the experiments reveal themselves in the histograms of data. The expected value is also compared for each experiment, as well as where the same value (e.g., 4) appears in each histogram. This can lead to a discussion on how likely this outcome is (or is not) for each experiment, based on whether it is in the tail of the distribution, a precursor of the P-value concept.

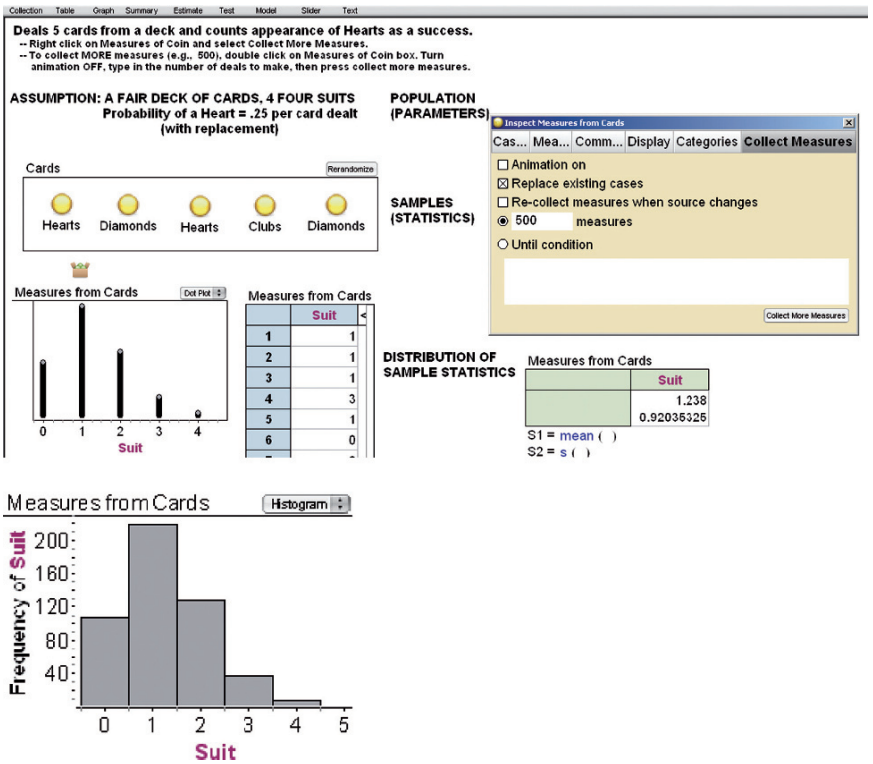

Fig. 7.5 The *Fathom* simulation of the number of hearts in cards

Lesson 3: The Normal Distribution as a Model

This lesson introduces students to the formal model of the normal distribution. The characteristics are examined, as students make and test conjectures about whether data gathered on different variables have a normal distribution. The unique characteristics of the mean and standard deviation in a normal distribution are used to examine percentages of data within one, two, and three standard deviations of the mean. The idea of the standard normal z-score is introduced and used to locate different areas of the distribution, using a Web applet. Student learning goals for this lesson include:

1. Understand and reason about the normal (and standard normal) distribution as a *model*
2. Understand and reason about the important characteristics of this distribution, e.g., the percentages of data within 1, 2 and 3 standard deviations of the mean.
3. Use standard deviations and z-scores to measure variation from the mean.

Description of the Lesson

Students are asked what they think when they hear the word "normal" and how they could tell if something is not normal. They contrast a normal data value (e.g., body temp of 98.6 F) in different contexts. They discuss how they think statisticians use this word and how their use is different from everyday use. Note, they have used the term *normal curve* informally in the unit on distribution (Chapter 8) when describing the shape of a data set.

In the *What is Normal* activity, students consider again the body measurements gathered earlier in the course (for which some may show a symmetric, bell-shaped pattern and others do not).

- Height
- Hand span
- Hand length
- Kneeling height
- Arm span
- Head circumference

The students make conjectures about which of these variables would have data sets that when graphed appear to have a normal distribution and why they predict those variables to have a normal distribution. They consider how to test their conjectures, and use the computer to generate graphs in *Fathom*. Next they select one distribution that looks approximately normal, to draw a picture of the graph on their handout and label the axes, the mean, and two standard deviations in each direction. Then they mark their own data value for this measure (e.g., their height) on the graph. They describe the location of their value in the overall graph as follows: *Are you close to center? In the tails? An outlier?* Next students find the z-score for their body measurement for that variable and explain what this z-score tells about the location of their body measurement relative to the class mean.

Students are instructed to open a Web applet that gives areas under the curve for a normal distribution. They use this applet to find the proportion of the distribution that is *more* than their value (e.g., what is the proportion of the curve representing values higher than their height) and then *less* than their value. Students discuss whether the obtained results from the applet makes sense to them and why or why not.

Next, students find the value that is one standard deviation above the mean and one standard deviation below the mean. They use the Applet to find the proportion of the distribution that is between these two values, and then repeat this for two and three standard deviations above and below the mean.

A class discussion focuses on which of the body measurements seemed to be normal and how they can tell how well the *normal model* fits a data set. Statisticians fit models to data, and this is illustrated by drawing a curve over the plot of a data set. Students consider and discuss how good a fit there is of the model to the data.

Next, students re-examine the use of the Web applet. They see that when they found the proportions of the distribution that were above and below their own

data value, it was from a normal distribution that has the class mean and standard deviation as the class data. They used the model of the normal curve to estimate proportions. Students see that it depends on how well the data fit the model and that when they use z-scores to find percentages (probabilities) using the normal curve, they are using a statistical model. The use of models to explain, describe, estimate, or predict is revisited, recalling the earlier use of models to simulate births of boys and girls.

In the *Normal Distribution Applications* activity, students apply the use of intervals used with the normal distribution (i.e., the middle 68, 95, 99.7%, referred to as the Empirical Rule) to real data sets. They explore and discuss when it is appropriate to use this rule (a model) to describe or make inferences about a particular set of data. They explore when and how to use the model to solve the problems in context. A last part of the activity is to see a *Fathom* demo on "What are Normal Data?"

In a final wrap-up discussion, students are given this popular quote by Box, "all models are wrong, but some are useful" (1979, p. 202) and discuss how this quotation applies to the normal distribution as a model.

Summary

Models are one of the most important and yet least understood ideas in an introductory statistics course. This chapter has tried to make the case that the idea of statistical model should be made explicit and used repeatedly in an introductory statistics course, so that students become familiar with the importance of models and modeling in statistical work. We believe that ideas of probability are best introduced in this context, without having to go into the formal rules and vocabulary that are better saved for a course in mathematical statistics or probability. We also encourage the explicit discussion of how models are used to simulate data, from informal uses early in the course to formal uses as part of tests of significance later in the course. When introducing and using the normal distribution as a model of certain univariate data sets or the regression line as a model of certain bivariate data sets, we hope instructors will describe the importance and use of these modes, and fitting models to data, modeling important aspects of both statistical practice and statistical thinking for the students to see.