

CHAPTER 1

A CASE STUDY OF STATISTICS IN ACTION



Were older workers discriminated against during a company's downsizing? When an older worker felt he was unfairly laid off, his lawyers called on a statistician to help evaluate the claim.

Robert Martin turned 55 in 1991. Earlier in that same year, the Westvaco Corporation, which makes paper products, decided to downsize. They laid off several members of their engineering department, where Robert Martin worked, and he was one of those who lost their jobs. Later that year, he sued Westvaco, claiming he had been laid off because of his age. A major piece of Martin's case was based on a statistical analysis of the ages of the employees at Westvaco.

In the two sections of this chapter, you will get a chance to try your hand at two very different kinds of statistical work, called exploration and inference. **Exploration** is an informal and open-ended examination of data for patterns. Your goal will be to uncover and summarize patterns in data from Westvaco that bear on the *Martin* case. You will try to formulate and answer basic questions like, "Were those who were laid off older on average than those who weren't laid off?" You can use any tools—graphs, averages, and so on—that you think might be useful. **Inference**, which you'll get to in the second section, is quite different from exploration in that it follows strict rules, and its focus is to judge whether the patterns you found are the sort you would expect. You'll use inference to decide whether the patterns you find in the Westvaco data are the sort you would expect from a company that does not discriminate on the basis of age or whether further investigation is needed into possible age discrimination.

The purpose of this first chapter is to give you a head start with the ideas of statistical thinking, before you get involved with the details of learning the methods. It is very easy to get caught in the trap of doing rather than understanding, of asking how rather than why. You can't *do* statistics unless you learn the methods, but you must not get so caught up in the details of the methods that you lose sight of what they mean. Doing and thinking, method and meaning, will compete for your attention throughout the course.

In this chapter you will learn the basic ideas of

- exploring data—uncovering and summarizing patterns
- making inferences from data—deciding whether or not an observed feature of the data could reasonably be attributed to chance

These will remain key components of statistical ideas that you'll develop and study throughout this book.

1.1 Discrimination in the Workplace: Data Exploration

At the beginning of 1991, Robert Martin was one of 50 people working in the engineering department of Westvaco's envelope division. That spring, Westvaco's management went through five rounds of planning for a reduction in their workforce. In Round 1, they decided to eliminate 11 positions, and in Round 2, 9 more. By the time the layoffs ended, after all five rounds, only 22 of the 50 workers had kept their jobs, and the average age in the department had fallen from 48 to 46.

Display 1.1 shows the data provided by Westvaco to Martin's lawyers.¹ Each row corresponds to one worker, and each column corresponds to a characteristic of the workers: job title, whether hourly or salaried, the date of birth, the date of hire, and age as of the first of January 1991 (shortly before the layoffs). The next-to-last column (RIF) tells how the worker fared in the downsizing: a 1 means chosen for layoff in Round 1 of planning for the reduction in force; a 2 means chosen in Round 2, and so on for Round 3, 4, or 5; and a 0 means "not chosen for layoff."

The subjects (or objects) of statistical examination are often called **cases**. Here, in the rows, the cases are individual Westvaco employees. Their characteristics, in the columns, are **variables**. If you pick a row and read across, you get information about a single case. (For example, Robert Martin, in Row 44, was salaried, was born in September 1937, was hired in October 1967, was chosen for layoff in Round 2, and was 54 on January 1, 1991.) Although reading across may seem the natural way to read the table, in statistics you will often find it useful to pick a column and read down. This gives you information about a single variable as you range through all the cases. For example, pick *Age*, read down the column, and notice the *variability* in the ages. It is variability like this—the fact that individuals differ—that can make it a challenge to see patterns in data and to figure out what they mean.

Just imagine: If there had been no variability—if all the workers had been of just two ages, say, 30 and 50, and Westvaco had laid off all the 50-year-olds and kept all the 30-year-olds—the conclusion would be obvious and there'd be no need for statistics. But real life is more subtle than that. The ages of the laid-off workers varied, as did the ages of the workers retained. *Statistical methods were designed to cope with such variability*. In fact, you might define statistics as the science of learning from data in the presence of variability.

Although the bare fact that the ages vary is easy to see in the data table, the pattern of the ages is not so easy to see in a column of numbers. This pattern—what the values are and how often each occurs—is called their **distribution**. In order to see that pattern, a graph is better. The **dot plot** in Display 1.2 shows the distribution of the ages of the 14 hourly employees who worked in the engineering department just before the layoffs began.



Variables provide information about cases.

Variability is what statistics is all about.

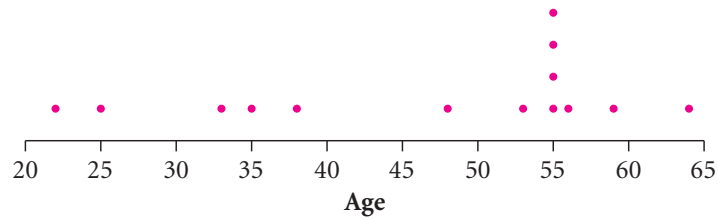
A distribution is a record of variability.

¹The statistical analysis in the lawsuit used all 50 employees in the engineering department of the envelope division, with separate analyses for exempt (salaried) and nonexempt (hourly) workers.

Row	Job Title	Pay	Birth		Hire		RIF	Age 1/1/91
			Mo	Yr	Mo	Yr		
1	Engineering Clerk	H	9	66	7	89	0	25
2	Engineering Tech II	H	4	53	8	78	0	38
3	Engineering Tech II	H	10	35	7	65	0	56
4	Secretary to Engin Manag	H	2	43	9	66	0	48
5	Engineering Tech II	H	8	38	9	74	1	53
6	Engineering Tech II	H	8	36	3	60	1	55
7	Engineering Tech II	H	1	32	2	63	1	59
8	Parts Crib Attendant	H	11	69	10	89	1	22
9	Engineering Tech II	H	5	36	4	77	2	55
10	Engineering Tech II	H	8	27	12	51	2	64
11	Technical Secretary	H	5	36	11	73	2	55
12	Engineering Tech II	H	2	36	4	62	3	55
13	Engineering Tech II	H	9	58	11	76	4	33
14	Engineering Tech II	H	7	56	5	77	4	35
15	Customer Serv Engineer	S	4	30	9	66	0	61
16	Customer Serv Engr Assoc	S	2	62	5	88	0	29
17	Design Engineer	S	12	43	9	67	0	48
18	Design Engineer	S	3	37	6	74	0	54
19	Design Engineer	S	3	36	2	78	0	55
20	Design Engineer	S	1	31	3	67	0	60
21	Engineering Assistant	S	6	60	7	86	0	31
22	Engineering Associate	S	2	57	4	85	0	34
23	Engineering Manager	S	2	32	11	63	0	59
24	Machine Designer	S	9	59	3	90	0	32
25	Packaging Engineer	S	3	38	11	83	0	53
26	Prod Spec—Printing	S	12	44	11	74	0	47
27	Proj Eng—Elec	S	9	43	4	71	0	48
28	Project Engineer	S	7	49	9	73	0	42
29	Project Engineer	S	8	43	4	64	0	48
30	Project Engineer	S	6	34	8	81	0	57
31	Supv Engineering Serv	S	4	54	6	72	0	37
32	Supv Machine Shop	S	11	37	3	64	0	54
33	Chemist	S	8	22	4	54	1	69
34	Design Engineer	S	9	38	12	87	1	53
35	Engineering Associate	S	2	61	9	85	1	30
36	Machine Designer	S	2	39	4	85	1	52
37	Machine Parts Cont—Supv	S	10	28	8	53	1	63
38	Prod Specialist	S	9	27	10	43	1	64
39	Project Engineer	S	7	25	9	59	1	66
40	Chemist	S	12	30	10	52	2	61
41	Design Engineer	S	4	60	5	89	2	31
42	Electrical Engineer	S	11	49	3	86	2	42
43	Machine Designer	S	3	35	12	68	2	56
44	Machine Parts Cont Coord	S	9	37	10	67	2	54
45	VH Prod Specialist	S	5	35	9	55	2	56
46	Printing Coordinator	S	2	41	1	62	3	50
47	Prod Dev Engineer	S	6	59	11	85	3	32
48	Prod Specialist	S	7	32	1	55	4	59
49	VH Prod Specialist	S	3	42	4	62	4	49
50	Engineering Associate	S	8	68	5	89	5	23

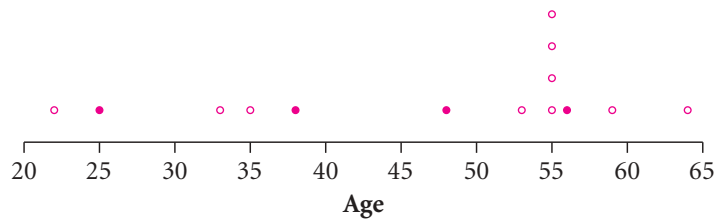
Display 1.1 The data in *Martin v. Westvaco*.

Source: *Martin v. Envelope Division of Westvaco Corp.*, CA No. 92-03121-MAP, 850 Fed. Supp. 83 (1994).



Display 1.2 Ages of the hourly workers. (Each dot is a worker; the ages are shown by the position of the dots along the scale below them.)

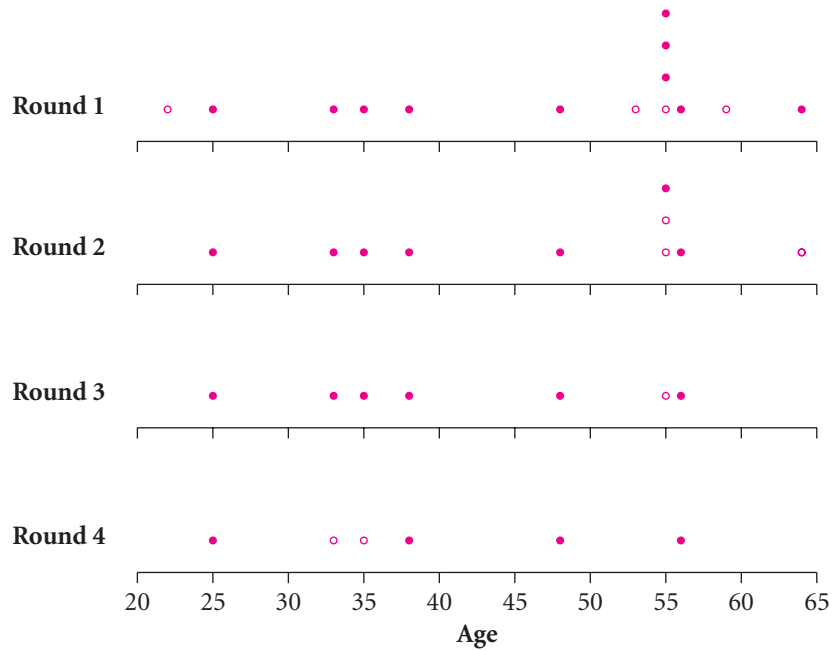
Display 1.2 provides some useful information about the variability in the ages but by itself doesn't tell anything about possible age discrimination in the layoffs. For that, you need something like Display 1.3 to distinguish between those hourly workers who lost their jobs and those who didn't.



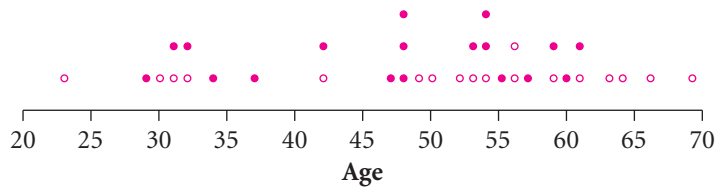
Display 1.3 Hourly workers: Ages of those laid off (open circles) and those retained (solid dots).

Discussion: Exploring the *Martin v. Westvaco* Data

- D1. Suppose you were on a jury in the *Martin v. Westvaco* case. How would you use the information in Display 1.1 to decide if Westvaco tended to lay off older workers (for whatever reason)?
- D2. Does the dot plot in Display 1.3 show a clear-cut case of age discrimination in layoffs of hourly workers at Westvaco, a possible case of age discrimination, or no discrimination?
- D3. The dot plots in Display 1.4 show the ages of the hourly workers laid off and retained by round. For example, in the top dot plot, the open circles show the ages of the hourly workers laid off in Round 1 and the solid dots show the ages of the hourly workers whose jobs survived Round 1. (There is no plot for Round 5 because no hourly workers were chosen for layoff in that round.) Compare the round-by-round information you get in Display 1.4 with the summary for all rounds in Display 1.3. Which display provides stronger support for Martin's claim that Westvaco discriminated against older workers?
- D4. The dot plot in Display 1.5 is like Display 1.3 except that it gives data for the salaried workers. Compare the plots for the hourly and salaried workers. Which provides stronger evidence in support of a claim of age discrimination?

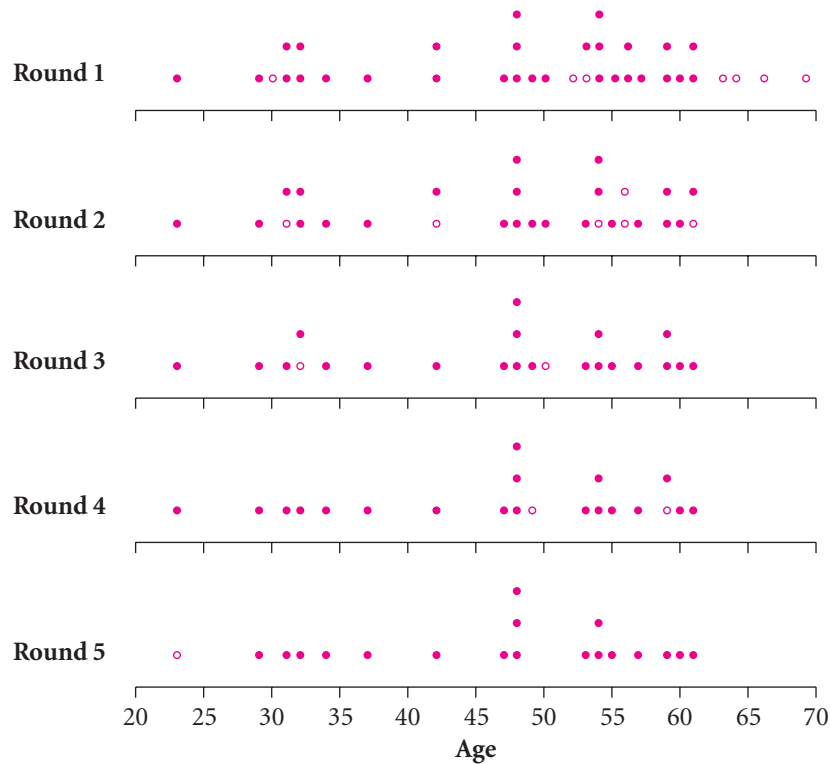


Display 1.4 Hourly workers: Ages of those laid off (open circles) and those retained (solid dots) in each round.



Display 1.5 Salaried workers: Ages of those laid off (open circles) and those retained (solid dots).

D5. Display 1.6 is the counterpart of Display 1.4. It shows, for the salaried workers, the ages of those laid off and those retained in each of the five rounds. Compare the pattern for the salaried workers with the pattern for the hourly workers in Display 1.4.



Display 1.6 Salaried workers: Ages of those laid off (open circles) and those retained (solid dots) in each round.

D6. The summary table shown here classifies salaried workers using two yes/no questions: Under 40? and Laid off? (In employment law, 40 is a special age because only those 40 or older belong to what is called the “protected class,” the group covered by the law against age discrimination.)

		Laid Off?		Total	% Yes
		Yes	No		
Under 40?	Yes	4	5	9	44.4
	No	14	13	27	51.9
	Total	18	18	36	50.0

- Does the pattern in this table support Martin’s claim of age discrimination? Why or why not?
- Construct a similar table for salaried workers, but this time use 50 instead of 40 to divide the ages. (Your two age groups will be those under 50 and those 50 or older.) Does the evidence in this new table provide stronger or weaker support for Martin’s case? Explain.
- How do you account for the different messages from the two tables? Both provide evidence; how do you judge the evidence from the two tables taken together?

- D7. Whenever you think you have a message from data, you should be careful not to jump to conclusions. The patterns in the Westvaco data might be “real”—they reflect age discrimination on the part of management. On the other hand, the patterns might be the result of chance—management wasn’t discriminating on the basis of age but simply by chance happened to lay off a larger percentage of older workers. What’s your opinion about the Westvaco data: Do the patterns seem “real”—too strong to be explained by chance?
- D8. You may feel as if the analysis so far ignores important facts like worker qualifications. That’s true. However, the first step is to decide if, based on the data in Display 1.1, older workers were more likely to be laid off. If not, Martin’s case fails. If so, it is then up to Westvaco to justify its actions. List several specific considerations that might justify the fact that older workers were more likely to be laid off.

■ Practice

- P1. Construct a dot plot, similar to the one in Display 1.3, comparing the ages of hourly workers who lost their jobs at some point during the first three rounds to the ages of hourly workers who still had their jobs at the end of Round 3. How do the ages differ?
- P2. In D6, you looked at a pair of summary tables for salaried workers.
- Construct similar tables for the hourly workers. Which of your tables (using age 40 or age 50) provides stronger evidence of age discrimination?
 - In what ways is the pattern in these tables similar to the patterns in D6 for salaried workers?

Summary 1.1: Data Exploration

Data exploration, or exploratory analysis, is a purposeful investigation to find patterns in data, using such tools as tables and graphs to display those patterns, and statistical concepts such as distributions and averages to summarize them.

- Table displays, with cases in rows and variables in columns, help you look at how variables differ from case to case.
- The distribution of a variable tells you the set of values that the variable takes on, together with how often each value occurs.
- Dot plots, which show the values of a variable along a number line, provide you with a visual display of the distribution of a variable.
- Plots of distributions give you a sense of how large or small the values are, which values occur most often, how spread out the values are, and whether there are any values that appear to be unusually large or small.

Remember that statistics involves coping with variability, but you have to understand that variability before you can use it intelligently to draw conclusions. All the features of data exploration that you have investigated here will be important when you move on to the inference phase of a statistical investigation.

■ Exercises

- E1. Explore whether hourly workers at Westvaco were more likely than salaried workers to lose their jobs. Start by constructing a table to summarize the relevant data. What do you conclude?
- E2. Twenty-two workers kept their jobs. Explore whether the age distributions are similar for the hourly and salaried workers who kept their jobs.
- Show the two age distributions in a pair of dot plots on the same scale. How do these distributions differ?
 - Do the dot plots support a claim that Westvaco was more inclined to keep older workers if they were salaried rather than hourly?
- E3. It may seem natural to think that cases will always be individuals, but it is also possible to have cases that are groups of individuals.
- Use Display 1.1 to create a data table whose five cases are Round 1 through Round 5 and whose three variables are the *total number* of employees laid off in that round, the *number* of employees laid off in that round who were 40 or older, and the *percentage* laid off in that round who were 40 or older.
 - Describe any patterns you find in the table and what you think they might mean.
- E4. “Last hired, first fired” is shorthand for “When you have to downsize, start by laying off the newest person, then the person hired next before that, and work back in reverse order of seniority.” (The person who’s been there longest will be the last to be laid off.) Examine the Westvaco data: How was seniority related to the decisions about layoffs in the engineering department at Westvaco? What explanation(s) can you suggest for the patterns you find?
- E5. Many tables in the newspaper and elsewhere are arranged with cases as rows and variables as columns. Pick one of these displays and tell what the cases and variables are.
- A table of major league baseball standings
 - A list of the day’s activity on the New York Stock Exchange
 - The nutritional summary on a food package



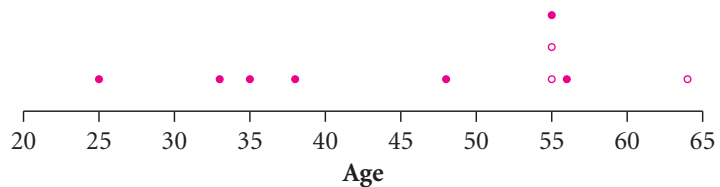
1.2 Discrimination in the Workplace: Inference

Overall, the exploratory work on the Westvaco data set in Section 1.1 shows that older workers were more likely than younger ones to be laid off, and they were laid off earlier. One of the main arguments in the court case, along the lines set out in D7, was about what those patterns mean:

- Can we infer from them that Westvaco has some explaining to do?
- Or are the patterns of the sort that might happen even if there was no discrimination?

A comprehensive analysis of *Martin v. Westvaco* will have to wait for its reappearance among the case studies of Chapter 12, when you'll know more of the concepts and tools of statistics. For now, though, you can get a pretty good idea of how the analysis goes by working with only a subset of the data.

The ages of the ten hourly workers involved in Round 2 of the layoffs, arranged from youngest to oldest, were 25, 33, 35, 38, 48, 55, 55, 55, 56, and 64. The three who were laid off were age 55, 55, and 64. Display 1.7 shows the same data in a dot plot.



Display 1.7 Hourly workers: Ages of those laid off (open circles) and those retained (solid dots) in Round 2.

Use a summary statistic to “condense” the data.

To simplify the statistical analysis to come, it will help to “condense” the data into a single number, called a **summary statistic**. One possible summary statistic is the average, or mean, age of the three who lost their jobs:

$$\frac{55 + 55 + 64}{3} = 58 \text{ years}$$

What to make of the data requires balancing two points of view. On one hand, the pattern in the data is pretty striking. Of the five people under age 50, all kept their jobs. Of the five who were 55 or older, only two kept their jobs. On the other hand, the number of people involved is pretty small: just three out of ten. Should you take seriously a pattern involving so few people? Listen to two imaginary people taking sides in an argument that was at the center of the statistical part of the *Martin* case.

Martin: Look at the pattern in the data. All three of the workers laid off were much older than the average age of all workers. That’s evidence of age discrimination.

Westvaco: Not so fast! You're looking at only ten people total, and only three positions were eliminated. Just one small change and the picture would be entirely different. For example, suppose it had been the 25-year-old instead of the 64-year-old who was laid off. Switch the 25 and the 64 and you get a totally different set of averages:

Actual data: 25 33 35 38 48 **55** **55** 55 56 **64**
 Altered data: **25** 33 35 38 48 **55** **55** 55 56 64

See! Just one small change and the average age of the three who were laid off is *lower* than the average age of the others.

	Average Age	
	Laid Off	Retained
Actual data	58.0	41.4
Altered data	45.0	47.0

Martin: Not so fast, yourself! Of all the possible changes, you picked the one that is most favorable to your side. If you'd switched one of the 55-year-olds who got laid off with the 55-year-old who kept his or her job, the averages wouldn't change at all. Why not compare what actually happened with *all* the possibilities that might have happened?

Westvaco: What do you mean?

Martin: Start with the ten workers, treat them all alike, and pick three at random. Do this over and over, to see what typically happens, and compare the actual data with these results. Then we'll find out how likely it is that their average age would be 58 or more.

Discussion: Picking Workers at Random

The dialog describes one age-neutral method for choosing which workers to lay off: Pick the three completely at random, with all sets of three having the same chance to be chosen.

- D9. If you pick three of the ten ages at random, do you think you are likely to get an average age of 58 or more?
- D10. If the probability of getting an average age of 58 or more turns out to be small, does this favor Martin or Westvaco?

Activity 1.1 shows you how to estimate the probability that if you choose three workers at random, just by chance you will get an average age of 58 years or more. To do that, you use **simulation**, a procedure in which you set up a model of a chance process (drawing three ages out of a box) that copies—or simulates—a real situation (selecting three employees at random to lay off).

Simulation requires a chance model.

Activity 1.1 By Chance or by Design?

What you'll need: paper or 3×5 cards, a box or other container

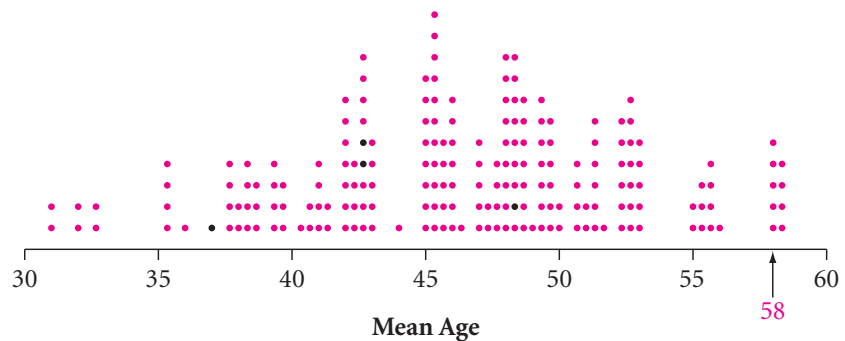
Let's test the process suggested by Martin's advocate.

1. *Create a model of a chance process.* Write each of the ten ages on identical pieces of paper or 3×5 cards, and put the ten cards in a box. Mix them thoroughly, draw out three (the ones to be laid off), and record the ages.
2. *Compute a summary statistic.* Compute the average of the three numbers in your sample to one decimal place.
3. *Repeat the process.* Repeat steps 1 and 2 nine more times.
4. *Display the distribution.* Pool your results with the rest of your class and display the summary statistics in a dot plot.
5. *Estimate the probability.* Calculate the number of times your class got an average age of 58 years or more. Estimate the probability that simply by chance the average age of those chosen would be 58 years or more.
6. *Interpret your results.* What do you conclude from the magnitude of your class's probability estimate?

Your simulation was completely age-neutral: All sets of three workers had exactly the same chance of being selected for layoff, regardless of age. The simulation tells you what sorts of results are reasonable to expect from that sort of age-blind process.

Here are the first 4 of 1000 repetitions from such a simulation. (The ages in red are those selected for layoff.) The average ages of these groups—42.7, 48.0, 42.7, and 37.0—are highlighted by the blue dots in the distribution in Display 1.8.

										Average age
25	33	35	38	48	55	55	55	56	64	42.7
25	33	35	38	48	55	55	55	56	64	48.0
25	33	35	38	48	55	55	55	56	64	42.7
25	33	35	38	48	55	55	55	56	64	37.0



Display 1.8 Results of 1000 repetitions: The distribution of the average age of the three chosen for layoff by chance alone. (Each dot represents 5 points.)

A distribution records variability in a chance process.

The simulation tells what kind of data to expect if workers are selected at random for layoff.

Here we take a closer look at the logic.

Display 1.8 is a dot plot that shows the distribution of average ages for 1000 repetitions of the process you used in Activity 1.1. This is the second important use of distributions to show variability: In the last section, you used distributions to show the set of values of one variable, *age*. Here, you have a distribution that shows the variability in a process, as you go from one repetition to the next.

Out of 1000 repetitions, only 46, or almost 5%, gave an average age of 58 or older. So it is not at all likely that just by chance you'd pick workers as old as the three Westvaco picked. Did the company discriminate? There's no way to tell from the numbers alone. However, if your simulations had told you that an average of 58 or larger is easy to get by chance alone, then the data would provide no evidence of discrimination. If, on the other hand, it turns out to be very unlikely to get a value this large simply by chance, as in this case, statistical logic says to conclude that the pattern is more than simple coincidence. It is then up to the company to explain why their decision-making process led to such a large average age for those laid off.

It might help your understanding of how this logic applies to *Martin v. Westvaco* if we imagine a realistic argument between the advocates for each side.

Martin: Look at the pattern in the data: All three of the workers laid off were much older than average.

Westvaco: So what? You could get a result like that just by chance. If chance alone can account for the pattern, there's no reason to look for any other explanation.

Martin: Of course you *could* get this result by chance. The question is whether it's easy or hard to do. If it's easy to get an average as large as 58 by drawing at random, I'll agree that we can't rule out chance as one possible explanation. But if an average that large is really hard to get from random draws, we agree that chance alone can't account for the pattern. Right?

Westvaco: Right.

Martin: Here are the results of my simulations. If you look at the three hourly workers laid off in Round 2, the probability of getting an average age of 58 or more by chance alone is only 5%. And if you do the same computations for the entire engineering department, the probability is a lot less, about 1%. What do you say to that?

Westvaco: Well . . . I'll agree that it's really hard to get patterns that extreme simply by chance, but that by itself still doesn't *prove* discrimination.

Martin: No, but I think it leaves you with some explaining to do!

In the actual case, Martin and Westvaco reached a settlement out of court before the case went to trial.

The logic you've just seen is basic to all statistical inference. But it's not easy to understand. In fact, it took mathematicians centuries to come up with it! It wasn't until the 1920s that a brilliant British biological scientist and mathematician, R. A. Fisher, realized that results of agricultural experiments should be analyzed carefully to see if observed differences could be attributed to

chance alone or to treatment. Calculus, in contrast, was first understood in 1665! Precisely because it *is* so important, you will see the logic of using randomization as a basis for statistical inference over and over again throughout this book. You'll have lots of time to practice with it.

Key Steps in a Simulation

Once again, a simulation goes like this:

1. *Model.* Set up a model of a process where chance is the only cause of being selected.

In Activity 1.1 related to the *Martin v. Westvaco* case, the model for an age-neutral chance process was to draw three numbers *at random* from the set of ten ages. You put the ages of the workers who could be laid off in a box and selected three by random draw.

2. *Repetition.* Gather data by repeating the process.

In the activity, you repeated the process of drawing ages many times.

3. *Distribution.* Display the distribution of outcomes using a summary statistic, and determine how likely the actual result would be.

In the activity, you used the average age to summarize the results, although other summaries also could be used. For each repetition, you computed the average age of those laid off and added that average to your dot plot. Repeated simulation showed that the chance of an average age of 58 or older was only about .05.

4. *Conclusion.* If the probability is small (and the definition of “small” will vary depending on the situation), conclude that chance may not be the only cause. If the probability isn't small, conclude that you can reasonably attribute the result to chance alone.

In the *Martin v. Westvaco* case, it turned out that the probability was small enough (1 chance out of 20) that Westvaco had some explaining to do. However, it wasn't small enough to serve as evidence of discrimination in a court case (which requires a probability of .025 or smaller).

Discussion: The Logic of Inference

- D11. Why must we estimate the probability of getting an average age of 58 *or more* rather than the probability of getting an average age of 58?
- D12. **How unlikely is “too unlikely”?** The probability you estimated in Activity 1.1 is in fact exactly equal to .05. In a typical court case, a probability of .025 or less is required to serve as evidence of discrimination.
- a. Did the Round 2 layoffs of hourly workers in the *Martin* case meet the court requirement?
 - b. What if the probability in the *Martin* case had been .01 instead of .05? Or .10 instead of .05? How would that have changed your conclusions?

- D13. **A trustworthy friend?** A friend wants to bet with you on the outcome of a coin toss. The coin looks fair, but you decide to do a little checking. You flip the coin and it lands heads. You flip again: also heads. A third flip: heads. Flip: heads. Flip: heads. You continue to flip, and the coin lands heads 19 times in 20 tosses.
- Explain why the evidence—19 heads in 20 tosses—makes it hard to believe the coin is fair.
 - Design and carry out a simulation to estimate how unusual this result would be if the coin were fair.

■ Practice

- P3. Suppose that three workers were laid off from a set of ten whose ages were the same as in the *Martin* case. However, this time, the ages of those laid off were 48, 55, and 55.

25 33 35 38 **48** 55 **55** **55** 56 64

- Use the dot plot in Display 1.8 to estimate the probability of getting an average age as large as that of those laid off here.
 - What would your conclusion be if Westvaco had laid off workers of these three ages?
- P4. At the end of Round 3, only six hourly workers were left. Their ages were 25, 33, 35, 38, 48, and 56. In Round 4, the 33- and 35-year-olds were chosen for layoff. Think about how you would repeat Activity 1.1 using the data from Round 4.
- What is the average age of the workers actually laid off?
 - Describe a simulation for finding the distribution of the average age of two workers laid off at random. Use your calculator or a statistical program to do the simulation. Repeat your simulation 20 times.
 - What is your estimate of the probability of getting an average age of 34 or more if two workers are picked at random for layoff in Round 4? Does this one part of the data (Round 4, hourly) provide evidence in Martin's favor?

Summary 1.2: Statistical Inference

Inference is a statistical procedure that involves deciding whether an event can reasonably be attributed to chance or whether you should look for—and perhaps investigate—some other explanation. In the *Martin* case, you used inference to determine whether the relatively high average age of the laid-off employees could reasonably be due to chance.

Simulation is a useful device for inference.

- First you set up a *model* of a process in which chance is the only factor influencing the outcomes.

- The next stage is *repetition*—you gather data by repeating the process in order to determine the likelihood of different outcomes.
- Then you plot the *distribution* of outcomes, using a summary statistic in order to determine how likely the actual result would be.
- Finally, you reach—or infer—a *conclusion*, evaluating the likelihood of getting your actual data in light of the chance-generated data.

If the probability of getting your actual data is small, conclude that chance may not be the only cause for getting the data. If the probability isn't small, conclude that you can reasonably attribute the result to chance alone. In the *Martin* case, the probability was about .05, which was considered small enough to warrant asking for an explanation from Westvaco but not small enough to present in court as clear evidence of discrimination.

■ Exercises

- E6. Revisit the idea of the simulation in Activity 1.1, this time for all 14 hourly workers and using a different summary statistic. Because the age class protected by law is those 40 or older, use as your summary statistic the number of hourly workers laid off who were 40 or older. The ages shown here are of the hourly workers, with those laid off in red. Note that, out of 10 hourly workers laid off by Westvaco, 7 were in the protected class.
- 22 25 33 35 38 48 53
55 55 55 55 56 59 64
- Write the 14 ages on 14 slips of paper and draw 10 at random for layoff. How many of the 10 were in the protected class of 40 or older? Repeat your simulation a total of 20 times using your calculator or statistical software, and make a dot plot of the number who were laid off in each repetition.
 - What is your estimate of the probability that, just by chance, seven or more of the ten hourly workers who were laid off would be in the protected class?
 - Do you conclude that seven out of ten could reasonably be due to chance alone, or should Westvaco be asked for an explanation?
 - Discuss the advantages and disadvantages of using actual ages rather than just the information about whether a person is 40 or older.
- E7. Ten hourly workers were left after Round 1. Their ages were
- 25 33 35 38 48 55 55 55 56 64
- The ages of the four workers laid off in Rounds 2 and 3 are in red type. They have an average age of 57.25. In the questions, consider the combined layoffs in Rounds 2 and 3.
- Describe how to simulate the chance of getting an average age of 57.25 or more using the methods of Activity 1.1 and your calculator or statistical program.
 - Repeat your simulation a total of ten times and make a dot plot of the average age of the four hourly workers laid off in Rounds 2 and 3. What is your conclusion?

- E8. **Which summary statistic?** Activity 1.1 asked you to use the average age to summarize the set of three ages of the workers chosen for layoff. Here are some other possible choices.
- Sum of the ages of the three who were laid off
 - Average age of those laid off minus average age of those retained
 - Number of employees 55 or older who were laid off
 - Age of the youngest worker who was laid off
 - Age of the oldest worker who was laid off
 - Middle (median) of the ages of the three who were laid off
- a. Are any of these possible summary statistics equivalent to the average age?
- b. Which of these possible summary statistics would it be reasonable to use?

- E9. **Snow in July?** You have spent some time in Oz. You think the date is July 4 back in Kansas, but you can't be sure because days may not have the same length in Oz as on Earth. A friendly tornado puts you and your dog Toto down in Kansas. However, you see snow in the air (data). Which of the following is the inference you should make?
- If this is Kansas, it is very unlikely to be snowing on July 4. Therefore, this probably isn't Kansas.
 - If it is July 4, it is very unlikely to be snowing in Kansas. Therefore, this probably isn't July 4.
 - If it is snowing in Kansas on July 4, it is time to go back to Oz.
 - If this is Kansas and it is July 4, it probably isn't really snowing.

E10. For some situations, instead of using simulation, it is possible to find exact

probabilities by counting equally likely outcomes. Suppose only two out of the 10 hourly workers had been laid off in Round 2 and that those two workers were 55 and 64, with an average age of 59.5 years. It is straightforward, though tedious, to list all possible pairs of workers who might have been chosen. Here's the beginning of a systematic listing. The first nine outcomes all include the 25-year-old and one other. The next eight outcomes all include the 33-year-old and one other but not the 25-year-old because the pair {25, 33} was already counted.

Count	Pair Chosen (red = laid off)										Average Age
1	25	33	35	38	48	55	55	55	56	64	29.0
2	25	33	35	38	48	55	55	55	56	64	30.0
3	25	33	35	38	48	55	55	55	56	64	31.5
.
.
9	25	33	35	38	48	55	55	55	56	64	44.5
10	25	33	35	38	48	55	55	55	56	64	34.0
11	25	33	35	38	48	55	55	55	56	64	35.5
...

- a. How many possible pairs are there? (Don't list them all!)
- b. How many pairs give an average age of 59.5 years or older? (Do list them.)
- c. If the pair is chosen completely at random, then all possibilities are equally likely, and the probability of getting an average age of 59.5 or older equals the number of possibilities with an average of 59.5 or more divided by the total number of possibilities. What is the probability for this situation?
- d. Is the evidence of age discrimination strong or weak?

Statistics in Action: What Next?

In this chapter, you have explored the data from the *Martin* case, looking for evidence you consider relevant. After that, you saw how to use statistical reasoning to test the strength of the evidence: Are the patterns in the data solid enough to support a conclusion of age discrimination, or are they the sort that you would expect to occur even if there was no discrimination?

The next two chapters deal with data exploration. Chapter 2 looks at distributions, much as you did in Section 1.1. What graphs and summaries are useful for finding and showing patterns? Chapter 3 does the same sort of thing for exploring relationships between pairs of variables. For example, how are employee age and seniority at Westvaco related?

Chapter 4 deals with the role of random samples in surveys and random assignment of treatments to subjects in experiments. With exploration, what you see is all you get. Often, though, you want more—you want to generalize to an entire population. When the Gallup Organization interviews a sample of 1500 carefully chosen people, the goal is not just to learn about those 1500 people; it's to learn about the whole adult population in the country. If you pause and think, it's not obvious that you can learn much about 200 million people just from talking to 1500, but the surprising fact is that you can—if the 1500 were selected at random.

Chapters 5, 6, and 7 deal with sampling distributions and probability. Instead of using a simulation, you will learn to use probability theory to find the characteristics of distributions like the one you generated in Activity 1.1.

Chapters 8 through 11 deal with inference—how to extend the logic you used in Section 1.2 to answer a great variety of questions. When the news media report the results of a poll and give the margin of error, what does that mean and where does that number come from? Do patients with AIDS-related complex (ARC) do better when given the drug AZT alone or AZT combined with acyclovir? Does living in a city reduce the ability of your lungs to get rid of harmful particles in the air you breathe? Can special exercises help a baby learn to walk sooner?

Finally, Chapter 12 is a review presented as a set of case studies. This last chapter tells the stories of several investigations, integrating all three elements of statistical thinking—collecting data, exploration, and inference. You'll have a chance to apply what you've learned to study the salaries of professional athletes, data from experiments to compare treatments for producing bigger and better flowers, and then end with a final look at the *Martin* case.

Review Exercises

E11. People with asthma often use an inhaler to help open up the lungs and breathing passages. A drug company has come up with a new compound to put in the inhaler that, they believe, will open up the lungs of the user even more than the standard compound tends to do. The new compound B and the standard compound A are each tested on five volunteers with asthma, with the ten volunteers being randomly split into the two treatment groups. The measurements are the increase in lung capacity (in liters) one hour after the use of the inhaler.



Compound A	Compound B
1.03	1.11
0.45	1.01
0.32	0.44
0.64	1.41
1.29	1.04

- By simply studying the data in the table, do you think compound B does better than compound A in increasing lung capacity?
 - Construct a dot plot, plotting the A's and B's with different symbols. Does it now appear that B tends to give larger measurements than A?
- Find the average increase for compound A and for compound B. When you compare these means, does it look to you as if B is better than A at opening up the lungs?
- E12. If you are studying the effects of poverty and plan to construct a data set whose cases are villages in Bolivia, what would be some meaningful variables?
- E13. Refer to the scenario of E11. If there really is no difference between compound A and compound B, then the apparent differences between the two data sets are due to chance alone. (Recall that the treatments were randomly assigned to the volunteers.) Your task, now, is to see if the observed difference in the means of each treatment group could be attributed to chance alone.
- Place the ten data values on separate slips of paper and mix them in a bag. Select five at random to play the role of the A treatment group; the other five play the role of the B treatment group. This time you will use as your summary statistic the difference between the means of each treatment group. Calculate this difference for your samples.
 - Repeat the procedure from part a until you have at least 20 simulated differences between means for the two groups. (You may use a calculator or computer to simulate the choices for participants for the two groups.) Plot the simulated differences on a dot plot.
 - Compute the difference between the means for the actual data. Mark this difference on the dot plot. How many simulated differences exceed this one? In light of this (small) simulation, do you think the actual difference could be due to chance alone? Explain your reasoning.

E14. In this exercise, you will follow the same steps as in E10 to find the probability of getting an average age of 58 or more when drawing three hourly workers at random in Round 2. The number of ways to pick three different workers from ten to lay off is

$${}_{10}C_3 = \binom{10}{3} = 120$$

- List the ways that give an average age of 58 or more.
- Compute the probability of getting an average age of 58 or more when three workers are selected for layoff at random.
- How does this number compare with the results of your class simulation in Activity 1.1? Why do the two probabilities differ (if they do)?

E15. How would your reasoning and conclusions change if the five oldest workers among the entire group of ten hourly workers in Round 2 were all age 55 (so that the ages of the ten were 25, 33, 35, 38, 48, 55, 55, 55, 55, 55) and the three chosen for layoff were all 55? Is the evidence of age discrimination stronger or weaker than in the actual case?