Figure 3.2. LOOKING AT DATA: Picturing Distributions of Data

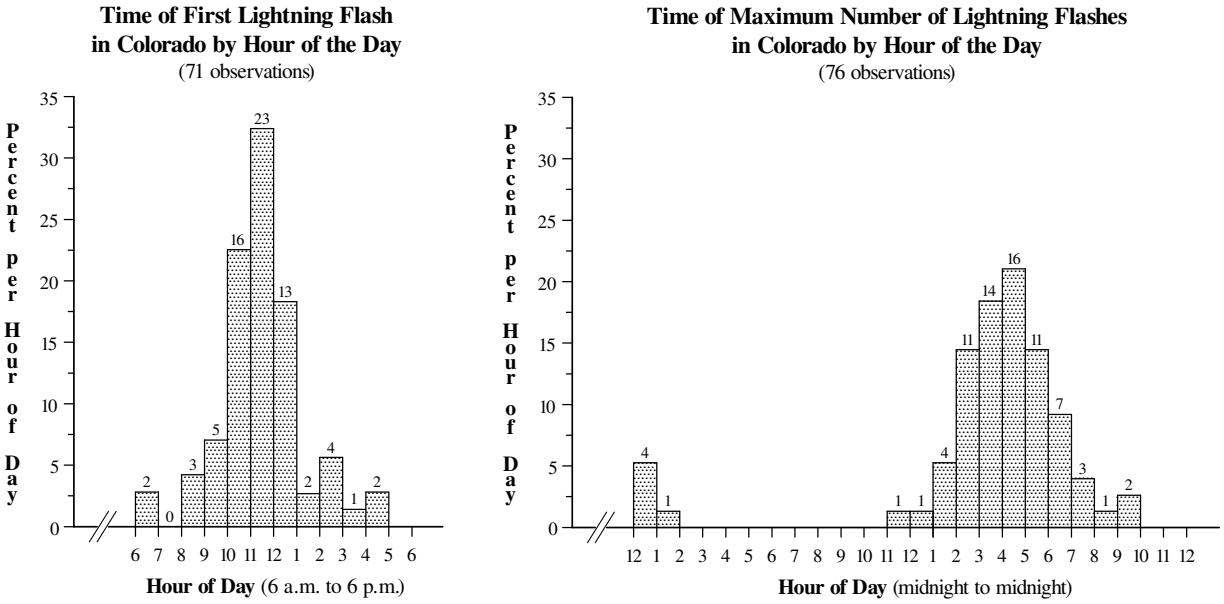
Program 2 in: *Against All Odds: Inside Statistics*

A *variable* is any numerical characteristic of a person, animal, or thing. The values of a variable vary when measurements are made on different people or things, or at different times. The *distribution* of a variable describes its pattern of variation by giving the values of the variable and how often each value occurs. This program shows how graphs can be used to help us see the major features of the distribution of a single variable.

You should look first for the *overall pattern* of the distribution, and then for major *deviations* from that pattern. The shape of the distribution – *symmetric* or *skewed* – is an important aspect of the overall pattern. The *centre* and *spread* of the distribution are also important. Deviations from the pattern may take the form of *outliers* or of gaps and other irregularities in the distribution. Distributions can be pictured by *stemplots* of the values or by *histograms* when the values are grouped into classes.

The video begins with the search for a pattern in lightning strikes. A histogram of lightning flashes in Colorado shows that the distribution of the hour of the first flash in a day is surprisingly regular. It is quite symmetric and centred near noon. A histogram of the hour with the maximum number of flashes is centered in the late afternoon, with a few outliers in the early morning. These distributions suggest explanations for lightning storms based on local wind patterns. When you make a histogram, you must first group your observations into classes. The video shows graphically how either too many or too few classes produce an uninformative histogram.

Often we want to compare several distributions. You watch the program director of a television station compare the age distribution for the viewers of two programs in the process of deciding which program to air. These distributions are skewed. Then you see a comparison of the number of hysterectomies performed by male and female gynecologists. This example shows how stemplots completely display the distribution of a small number of observations, and how *back-to-back* stemplots show comparisons clearly.



1 In each of the histograms above, explain what the figures on the tops of the bars indicate.

- Suggest an improved method for constructing the scale on the vertical axis of each histogram; make it clear *why* your suggestion would be an improvement.
  - Suggest a possible reason why this improvement was *not* used for these histograms in the video.

2 Explain briefly what can be inferred from the two histograms above about the following characteristics of electrical storms in Colorado:

- their starting time;
- their time of maximum intensity;
- their duration.

(*continued overleaf*)

Numbers of hysterectomies performed by 26 Swiss gynecologists:

The stemplot on the *left* shows the data for all 26 gynecologists; the back-to-back stemplots on the *right* show the data subdivided by the sex of the gynecologist.

				Male		Female	
0	057	(3)	(–)		0	057	(3)
1	0488	(4)	(–)		1	0488	(4)
2	05557 89	(7)	(5)	87550	2	59	(2)
3	11334 67	(7)	(5)	76431	3	13	(2)
4	4	(1)	(1)	4	4		(–)
5	09	(2)	(2)	90	5		(–)
6		(–)	(–)		6		(–)
7		(–)	(–)		7		(–)
8	56	(2)	(2)	65	8		(–)

3 Outline the most important feature(s) of the hysterectomy data that are indicated by the back-to-back stem plots; *quantify* your comments where appropriate.

- Suggest a plausible explanation for each feature you identify.
- Comment briefly on the statement: *The back-to-back stemplots show that five of the 15 male gynecologists performed more hysterectomies than any of the 11 females.*

1995-04-20