discontinued in the following series of sessions. The communication treatment seemed to make little difference in Series 1, and the subsequent sessions were conducted under conditions of full communication.

The results of Series 1 clustered around the common prediction of models 1–8, and 16 and seem to decisively reject models 9–15 for this committee institution and environment. Series 2 was designed to isolate models 5 and 7 from 1–4, 6, and 16. The results did not support models 5 and 7 and clustered around the common prediction of the other models. Series 3 sessions were designed so core and voting equilibrium did not exist; only the obvious point (model 16) and min-max set (model 4) were defined. The committee outcomes were dispersed around the min-max set, though the explanatory power of the set was not high. On the other hand, absence of the core did not result in complete dispersion of the outcomes over the blackboard as some theories have predicted.

Fiorina and Plott (1978) is a good example of a seminal experiment conducted with little more than paper, pencil, and chalkboard for equipment and facilities. In this, as in any other good experiment, most of the work goes into defining the critical issues, identifying the relevant theories and facts, and designing critical experiments before any subjects are recruited. The published paper includes detailed instructions and parameters to enable the reader to replicate their research. Instructions have been reproduced in Appendix II.

7

Data analysis

Imagine that you have just assembled the raw data from your recent experiments on market efficiency. You gaze at sheets of paper covered with numbers specifying which subjects did what and when they did it. Do the data support the efficient-markets hypothesis or not? You could stare at the raw data for hours and be none the wiser. It is time to begin your *data analysis*. You will transform and process the raw data in various ways to find out what they have to say. Think of data analysis as a form of interrogation. But be gentle – coax the data to tell their own story. You will learn very little if you torture the data until they confess.

This chapter introduces the basic tools for analyzing experimental data. Many experimentalists prefer a two-phase approach. The first phase is qualitative or descriptive and is intended to give an overview of what the data have to say. The tools are graphs and summary statistics. The second phase is more quantitative and is intended to give specific answers to specific questions. Here the tools are inferential statistics.

Experimental data and happenstance data raise the same general issues and require mostly the same analytical techniques, but there are notable differences in emphasis. Experimental data often come from newly created environments and are unlikely to be familiar to most readers, so the descriptive phase is particularly important. In most respects, statistical inference is quite straightforward for data obtained in well-designed experiments. Some subtleties do arise, which we discuss in Section 7.2. Section 7.3 should provide helpful perspectives on statistical tests of experimental data and a quick review of several specific tests, but for systematic training you will have to consult texts such as Box et al. (1978), Conover (1980), or Kirk (1982). Finally, after summarizing our advice on data analysis, we illustrate the main ideas while reviewing some of the literature on first price auctions.

7.1 Graphs and summary statistics

Day 18 of session Das2 was about as simple and straightforward as a trading period can be in a double-auction asset-market experiment. There were 8 traders divided equally into two types, each trader initially endowed with three shares. At the beginning of each 2-minute trading period ("Day"), all traders were notified of their per share payouts for the Day; for Day 18 the payouts were 25 cents for type 1 traders and 75 cents for type 2 traders. (Some Days type 1 traders get a payout of $1.95 and some Days type 2 traders get a payout of $1.65 in this session.) You might like to know whether all shares were acquired by the traders who valued them most highly (type 2), whether prices approached the fundamental value of 75 cents, whether prices were volatile, whether convergence was fast or slow or nonexistent, and so on.

Table 7.1 provides a complete record of all activity in the trading period, about 100 events (bids, asks, etc) in all. Look at Table 7.1 for a minute or two. Do these raw data answer your questions clearly? Now look at Figure 7.1, where the same data are plotted. (The upper step function is the best ask price, the lower step function is the best bid price, and stars indicate transaction prices. The horizontal dashed line is the equilibrium price, $0.75 per share. The realized payouts (1B, 2B) for the two trader types are indicated in the upper left corner, and the final allocation of shares is indicated in the lower right corner.) You can see at a glance in Figure 7.1 exactly what happened on Day 18 of session Das1. After about 10 seconds the traders had begun to digest the bad news (the low payouts to type 1 and 2 traders are indicated in the upper left corner of the graph by the notation 1B, 2B). Bids rose quickly to near the fundamental value of 75 cents and asks gradually declined toward that value, taking about 60 seconds to converge. By this time traders transacted 6 times, all accepted bids. Accepted asks were common in the 8 later transactions. Except in the first 30 seconds, the pace of trade was quite steady and all transactions prices were between 70 and 75 cents per share. By the end of the trading period, all 24 shares were held by the right type (2) of traders.

Summary statistics can be very useful in conjunction with graphs or even on their own. The final allocations shown in the lower right corner of the graph are summary statistics. Another example not shown explicitly is the mean transaction price. It is $0.725, a −2.5 cent deviation from equilibrium. This single number summarizes much of the information in Table 7.1 relevant to testing equilibrium theory.

Table 7.1 *Data from trading period Day 18 of double-auction asset market Das2*

period	subper	time	id	cpid	event	price	qty	bbid	qty	bask	qty
18	1	81	5	-	ASK	1.65	1.00	0.00	1.00	1.65	1.00
18	1	87	6	-	ASK	2.00	1.00	0.00	1.00	1.65	1.00
18	1	98	2	-	BID	0.70	1.00	0.70	1.00	1.65	1.00
18	1	104	3	-	ASK	2.00	1.00	0.70	1.00	1.65	1.00
18	1	107	0	-	ASK	1.50	1.00	0.70	1.00	1.50	1.00
18	1	120	7	-	ASK	1.70	1.00	0.70	1.00	1.50	1.00
18	1	126	6	-	ASK	1.55	1.00	0.70	1.00	1.50	1.00
18	1	133	0	-	BID	0.70	1.00	0.70	1.00	1.50	1.00
18	1	166	7	-	ASK	1.80	1.00	0.70	1.00	1.50	1.00
18	1	175	0	-	BID	0.74	1.00	0.74	1.00	1.50	1.00
18	1	189	5	-	ASK	1.45	1.00	0.74	1.00	1.45	1.00
18	1	217	2	-	ASK	1.49	1.00	0.74	1.00	1.45	1.00
18	1	235	4	-	ASK	1.00	1.00	0.74	1.00	1.00	1.00
18	1	240	0	-	ASK	1.40	1.00	0.74	1.00	1.00	1.00
18	1	248	6	-	ASK	2.00	1.00	0.74	1.00	1.00	1.00
18	1	254	3	-	BID	0.74	1.00	0.74	1.00	1.00	1.00
18	1	259	2	-	ASK	1.45	1.00	0.74	1.00	1.00	1.00
18	1	289	0	-	ASK	0.95	1.00	0.74	1.00	0.95	1.00
18	1	292	5	0	SOLD	0.74	1.00	0.74	1.00	0.95	1.00
18	1	296	0	-	CANBID	0.74	--	0.74	1.00	0.95	1.00
18	1	301	6	-	ASK	0.99	1.00	0.74	1.00	0.95	1.00
18	1	306	5	3	SOLD	0.74	1.00	0.74	1.00	0.95	1.00
18	1	311	3	-	CANBID	0.74	--	0.70	1.00	0.95	1.00
18	1	317	2	-	BID	0.74	1.00	0.74	1.00	0.95	1.00
18	1	337	0	-	BID	0.74	1.00	0.74	1.00	0.95	1.00
18	1	345	2	-	BID	0.70	1.00	0.74	1.00	0.95	1.00
18	1	375	4	-	BID	0.20	1.00	0.74	1.00	0.95	1.00
18	1	380	2	-	BID	0.74	1.00	0.74	1.00	0.95	1.00
18	1	395	6	-	ASK	0.94	1.00	0.74	1.00	0.94	1.00
18	1	415	4	-	CANASK	1.00	--	0.74	1.00	0.94	1.00
18	1	427	1	-	BID	0.65	1.00	0.74	1.00	0.94	1.00
18	1	457	2	-	ASK	0.93	1.00	0.74	1.00	0.93	1.00
18	1	468	0	-	CANBID	0.74	--	0.74	1.00	0.93	1.00
18	1	473	4	-	CANASK	0.70	--	0.74	1.00	0.93	1.00
18	1	473	4	0	SOLD	0.74	1.00	0.74	1.00	0.93	1.00
18	1	491	5	2	SOLD	0.74	1.00	0.74	1.00	0.93	1.00
18	1	494	2	-	CANBID	0.74	--	0.65	1.00	0.93	1.00
18	1	495	5	-	CANASK	1.45	--	0.65	1.00	0.93	1.00
18	1	506	0	-	BID	0.74	1.00	0.74	1.00	0.93	1.00
18	1	510	6	-	ASK	0.90	1.00	0.74	1.00	0.90	1.00
18	1	531	2	-	ASK	0.91	1.00	0.74	1.00	0.90	1.00
18	1	538	0	-	ASK	0.85	1.00	0.74	1.00	0.85	1.00
18	1	547	0	-	CANBID	0.74	--	0.65	1.00	0.85	1.00
18	1	552	4	-	CANASK	0.72	--	0.65	1.00	0.85	1.00
18	1	552	4	0	SOLD	0.74	1.00	0.65	1.00	0.85	1.00
18	1	571	2	-	ASK	0.88	1.00	0.65	1.00	0.85	1.00
18	1	582	0	-	BID	0.74	1.00	0.74	1.00	0.85	1.00
18	1	609	0	-	CANBID	0.74	--	0.65	1.00	0.85	1.00

Table 7.1 (*cont.*)

period	subper	time	id	cpid	event	price	qty	bbid	qty	bask	qty
18	1	614	6	-	CANASK	0.74	--	0.65	1.00	0.85	1.00
18	1	614	6	0	SOLD	0.74	1.00	0.65	1.00	0.85	1.00
18	1	620	4	-	ASK	0.74	1.00	0.65	1.00	0.74	1.00
18	1	624	2	-	BĭD	0.70	1.00	0.70	1.00	0.74	1.00
18	2	655	2	-	BID	0.66	1.00	0.66	1.00	0.74	1.00
18	2	674	0	-	BID	0.73	1.00	0.73	1.00	0.74	1.00
18	2	683	7	-	ASK	0.80	1.00	0.73	1.00	0.74	1.00
18	2	692	6	-	ASK	0.74	1.00	0.73	1.00	0.74	1.00
18	2	735	0	-	CANBID	0.73	--	0.66	1.00	0.74	1.00
18	2	739	7	-	CANASK	0.70	--	0.66	1.00	0.74	1.00
18	2	739	7	0	SOLD	0.73	1.00	0.66	1.00	0.74	1.00
18	2	745	4	-	CANASK	0.74	--	0.66	1.00	0.74	1.00
18	2	751	2	-	BID	0.67	1.00	0.67	1.00	0.74	1.00
18	2	764	6	-	ASK	0.70	1.00	0.67	1.00	0.70	1.00
18	2	772	0	-	CANBID	0.74	--	0.67	1.00	0.70	1.00
18	2	777	6	-	CANASK	0.70	--	0.67	1.00	0.85	1.00
18	2	777	0	6	BOUGHT	0.70	1.00	0.67	1.00	0.85	1.00
18	2	787	4	-	ASK	0.71	1.00	0.67	1.00	0.71	1.00
18	2	816	0	-	CANBID	0.74	--	0.67	1.00	0.71	1.00
18	2	824	4	-	CANASK	0.71	--	0.67	1.00	0.85	1.00
18	2	824	0	4	BOUGHT	0.71	1.00	0.67	1.00	0.85	1.00
18	2	830	7	-	ASK	0.70	1.00	0.67	1.00	0.70	1.00
18	2	870	0	-	CANBID	0.74	--	0.67	1.00	0.70	1.00
18	2	872	7	-	CANASK	0.70	--	0.67	1.00	0.85	1.00
18	2	873	0	7	BOUGHT	0.70	1.00	0.67	1.00	0.85	1.00
18	2	878	6	-	ASK	0.75	1.00	0.67	1.00	0.75	1.00
18	2	911	2	-	ASK	0.73	1.00	0.67	1.00	0.73	1.00
18	2	915	7	-	ASK	0.80	1.00	0.67	1.00	0.73	1.00
18	2	917	6	-	ASK	0.74	1.00	0.67	1.00	0.73	1.00
18	2	932	0	-	CANBID	0.74	--	0.67	1.00	0.73	1.00
18	2	938	2	-	CANASK	0.73	--	0.67	1.00	0.74	1.00
18	2	938	0	2	BOUGHT	0.73	1.00	0.67	1.00	0.74	1.00
18	2	946	6	-	ASK	0.72	1.00	0.67	1.00	0.72	1.00
18	2	962	0	-	BID	0.70	1.00	0.70	1.00	0.72	1.00
18	2	973	0	-	CANBID	0.70	--	0.67	1.00	0.70	1.00
18	2	974	7	-	CANASK	0.70	--	0.67	1.00	0.72	1.00
18	2	974	7	0	SOLD	0.70	1.00	0.67	1.00	0.72	1.00
18	2	991	2	-	BID	0.71	1.00	0.71	1.00	0.72	1.00
18	2	1005	1	-	BID	0.70	1.00	0.71	1.00	0.72	1.00
18	2	1028	0	-	BID	0.70	1.00	0.71	1.00	0.72	1.00
18	2	1049	2	-	ASK	0.72	1.00	0.71	1.00	0.72	1.00
18	2	1076	0	-	CANBID	0.74	--	0.71	1.00	0.72	1.00
18	2	1079	6	-	CANASK	0.72	--	0.71	1.00	0.72	1.00
18	2	1079	0	6	BOUGHT	0.72	1.00	0.71	1.00	0.72	1.00
18	2	1086	2	-	ASK	0.73	1.00	0.71	1.00	0.73	1.00
18	2	1129	0	-	CANBID	0.74	--	0.71	1.00	0.73	1.00
18	2	1131	2	-	CANASK	0.73	--	0.71	1.00	0.85	1.00
18	2	1131	0	2	BOUGHT	0.73	1.00	0.71	1.00	0.85	1.00
18	2	1165	0	-	BID	0.73	1.00	0.73	1.00	0.85	1.00
18	2	1196	2	-	ASK	0.84	1.00	0.73	1.00	0.84	1.00

Time is measured in tenths of a second from the beginning of the trading period. Traders with I.D.'s 0–3 are type 1 and have payout $0.25 per share on Day18. Traders with I.D.'s 4–7 are type 2 and have payout $0.75. The counterparty in a transaction appears in the cpid column. In this session the quantity traded (qty) is always 1.0, i.e., trades are for single indivisible shares. The best bid and best ask are denoted bbid and bask.

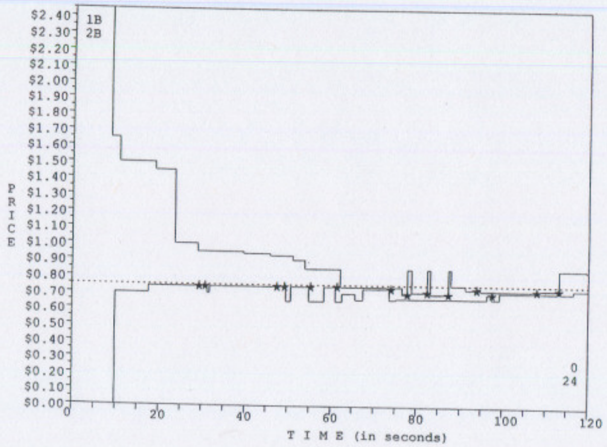

Fig. 7.1 Time graph for Day 18 of Das2. The upper step function is the best ask price, the lower step function is the best bid price, and stars indicate transaction prices. The horizontal dashed line is the equilibrium price, $0.75 per share. The realized payouts (1B, 2B) for the two trader types are indicated in the upper left corner, and the final allocation of shares is indicated in the lower right corner.

For another example, consider the risky choice experiments reported in Kachelmeier and Shehata (1992). Their raw data are certainty equivalents (selling prices elicited via the Becker-DeGroot-Marschak procedure mentioned in Section 4.2) from various subjects for various lotteries with differing probabilities of winning a fixed cash prize. With 50 trials for each of 20 subjects in their first session, the raw data consists of 1,000 numbers. Their main summary statistic is called CE ratio, the ratio of the certainty equivalent to the expected value, usually averaged over subjects. Figure 7.2 reproduces their Figure 1. You can see at a glance that subjects demanded a substantial premium before they were willing to sell the low-probability lotteries, but the premium decreased as the win probability increased and when a high cash prize was substituted for the low cash prize.

How can you choose a good summary description of your data? Perhaps the best advice is to look at past work for an effective presentation, and modify it to deal with special features of your own data. The tradition behind Figure 7.1, for example, goes back at least to Smith (1962). But

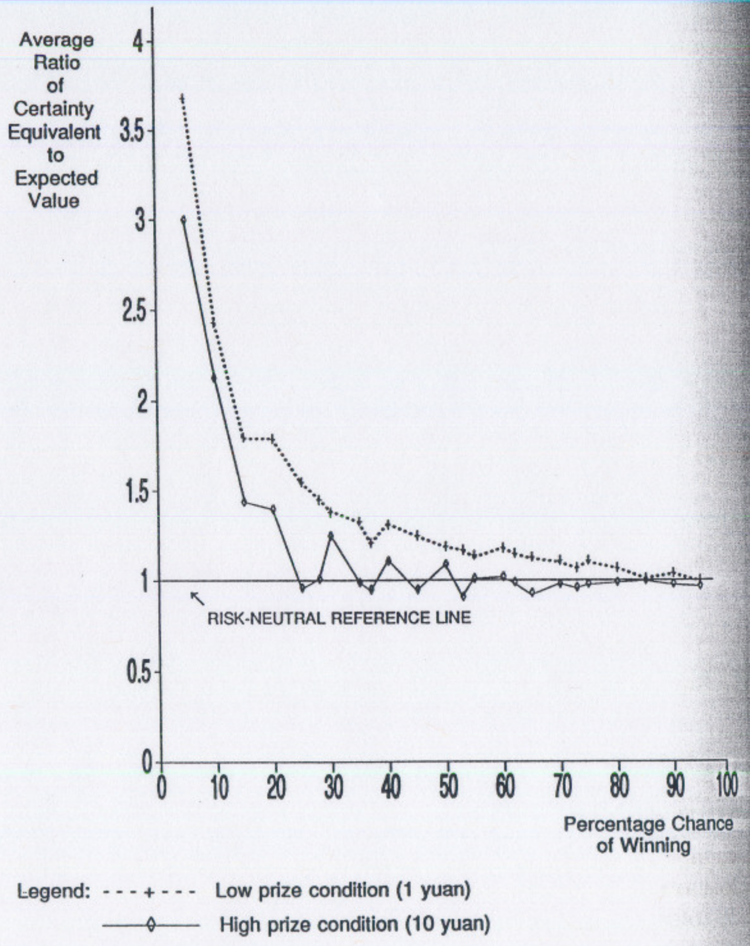

Fig. 7.2 Certainty equivalents and expected values.

the display was modified to show bids, asks, and transactions in clock time, rather than just the traditional transaction sequences, because an important goal for the experiment was to see how bids and asks adjust over time.

A good summary of your data accomplishes several goals. First, it allows you to see regularities (or irregularities) in the data that require further investigation. Graphs are a remarkably efficient means of screening for erroneous data. It is equally important to spot correct but anomalous data. For example, summary data might show that one subject in a risky choice experiment has a much lower CE ratio at low win probabilities than the other subjects. Further investigation might disclose that the low average is due to selling prices of zero. You should then check whether the prices were correctly recorded, whether the subject received proper instructions, and so forth. If the data are in fact correct, you might wish to see whether other subjects indulge in zero selling prices. The upshot might be a modified theory in which subjects with very low expected winnings and high subjective computational costs will bid zero, with implications that go to the foundations of decision theory! If you hadn't worked out the data summary, you probably wouldn't have spotted the zero bids and you would have missed the opportunity to correct your data or to extend the theory.

A second goal of qualitative data analysis is to guide subsequent quantitative analysis. For example, you may wish to analyze discrepancies between theoretical equilibrium prices and actual prices in a double-auction market. But what is the appropriate "actual price"? Is it the average transaction price in a trading period? The last transaction price? The midpoint of the bid-ask price interval? A summary graph like Figure 7.1 gives you a basis for making an appropriate choice and indicates whether other choices are likely to give different answers. Your formal statistical inferences will be more reliable if they are grounded in a good descriptive analysis.

A third goal is pedagogical. A good graphical display or set of descriptive statistics gives your reader an easily accessible overview of your data. The reader will then be encouraged to read on to your conclusions and will be in a better position to assess their credibility.

Data summaries are less important for well-known happenstance data, such as financial market data or national income accounts data compiled by government agencies. The econometrician analyzing such data probably already has an adequate perspective on the data and is aware of its main features. Her readers will want to get quickly to her contribution, perhaps a more subtle inferential statistic, and may be impatient with a lot of familiar descriptive statistics. By contrast, experimental data usually are new and in some respects unfamiliar, so a descriptive summary is essential.

Sometimes the main question addressed in an experiment can be answered directly from the summary statistics or graphs. For example, the issue in a set of recent market experiments was whether a theoretically inefficient market institution called CHQ was less efficient in practice than a theoretically more efficient institution called CH. Figure 7.3

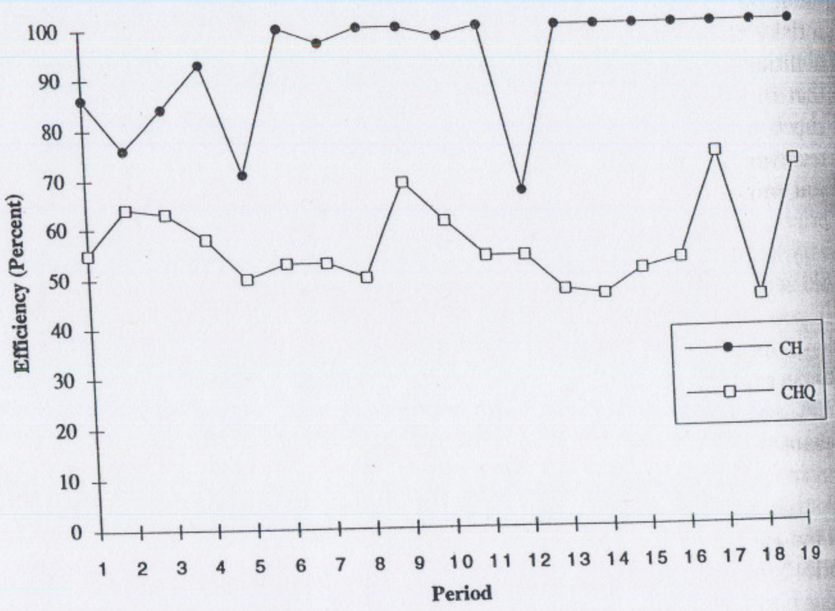

Fig. 7.3 Efficiency under the clearinghouse (CH) and quantity-only clearinghouse (CHQ) institutions. Efficiency is defined as trading profits paid as a percentage of maximum possible trading profits. The data come from all 19 CH periods and the first 19 CHQ periods of two sessions (Chq1–Ch4a and Chq2–Ch4b) reported in Friedman and Ostroy (1993).

graphs efficiency in the two sessions using both institutions. (Efficiency is defined as trading profits paid as a percentage of maximum possible trading profits. The data come from all 19 CH periods and the first 19 CHQ periods of two sessions, Chq1-Ch4a and Chq2-Ch4b, reported in Friedman and Ostroy, 1993.) The answer is obvious from the graphs – it immediately strikes your eye that efficiency is always higher in the CH markets, irrespective of the group of subjects or other nuisances. Leonard J. Savage referred to the pratice of drawing conclusions from such blindingly obvious graphs or summary statistics as the "interocular trauma test."

Is any other test really necessary? Experimental physicists usually rely on Savage's test and seldom resort to formal hypothesis testing. Some of our respected colleagues say privately that economists should follow the physicists' example. If the interocular trauma test is inconclusive, they argue, then you should rethink your experimental design or your

presentation of the data. Some other economists (anonymous referees for the most part) insist on hypothesis tests even when the Savage test seems conclusive. They argue that your clever graphical presentation may overstate the weight of the evidence and that the discipline of conducting hypothesis tests will help keep you honest.

Most practicing experimental economists, including both of us, take an intermediate position. Occasionally the Savage test should convince even the most skeptical, and then it is sufficient. More often it will not suffice. Experimental economists, unlike physicists, usually have to deal with many nuisance variables and relatively few observations, so even clever designs and large budgets can not always produce transparent results. When in doubt (or in doubt about referees) we recommend that you conduct routine hypothesis tests.

7.2 Statistical inference: Preliminaries

Suppose that your graphs and descriptive statistics do not give crystal clear answers to some of your questions, even though your experimental design and descriptive statistics are well chosen. At this point you turn to the second phase of the data analysis: formal statistical tests, or inferences. The formal tests are generally meant to provide specific answers to questions of the form "Does treatment X affect outcome Y?" For example, does the double auction market institution (treatment X = DA) increase market efficiency (outcome Y) relative to an alternative institution (treatment X = CH)? Sometimes you ask questions of the form "Is outcome Y better predicted by model M1 or by model M2?"

The most obvious way to answer the first sort of question is to compare the effects $\{y_{DA}\}$ associated with one treatment X = DA to the effects $\{y_{CH}\}$ associated an alternative treatment X = CH. If the y_{DA}'s are larger on average you might be tempted to conclude that the DA institution is more efficient. Likewise, you would be tempted to conclude that model M1 is better than M2 if on average its forecasts are more accurate. But your conclusion might be incorrect because of experimental error. The rest of this section will equip you with the conceptual tools for understanding the sources and consequences of experimental error. Later sections introduce statistical techniques for making correct inferences even when some experimental error is unavoidable.

7.2.1 Basic concepts

Statistical procedures begin with a collection of *observations*. A single observation is often called a run or experimental trial. A trial will include measurements of the treatments X and the outcomes Y. For example, a trial (or unit of observation) in a sealed-bid auction experiment might consist of the value v_t and the bid b_t of a single bidder in a given period, together with block data such as the number of bidders, the distribution parameters for v, and the auction rules.

The appropriate unit of observation is not always clear. For instance, in market experiments, is it a single transaction? A single market period? A subset of market periods? Or perhaps a single experimental session or even a whole set of sessions? The answer depends on the theoretical framework and the purpose of the experiment. For example, the market period is the natural unit when your purpose is to test theories of market equilibrium. If you were interested in the microdynamics of information acquisition, by contrast, the natural unit of observation would be transactions or even individual trader bids and asks. At the other extreme, someone interested in the asymptotics of group learning behavior would legitimately regard an entire experimental session as a single trial.

Suppose you have picked an appropriate definition of trial and now have a set of observations to analyze. The fundamental problem you now must deal with is the imperfections of your set of observations. To the extent that you get different results on replication – that is, to the extent that outcomes differ when you (or another experimenter) run the experiments again with exactly the same set of treatments – your analysis must deal with *experimental error*.

Experimental error has two sources: measurement and sampling. Measurement error is conceptually straightforward. The values in your recorded observations may not be exactly the actual values. Perhaps you misheard a bid in an oral auction, or perhaps you made a mistake in writing it down. Even more serious, you might have lost experimental control and not been aware of it at the time. For example, you might have inadvertently given role A information to a role B subject. Or in a game-theory experiment you may have transposed the intended payoff matrix on every player's screen (as did one of us recently).

Careful choice of laboratory procedures, automating data capture and transmission where possible, and building in redundancy should minimize the amount of erroneous data. You should always take a second precaution: Using your data summaries, check the raw data for large outliers and other anomalies, and check whether the anomalies are actually measurement errors. When you detect erroneous data you should throw them out before you run statistical tests, because even a few bad data points (say, due to a misplaced decimal point) can affect your results.

Sometimes failure of experimental controls produces data that still are interesting (e.g., the transposed matrix may induce a new coordination game instead of the intended coordination game) and you may want to retain it. Such reparametrization is permissible as long as your analysis recognizes the inadvertent change in experimental design (e.g., you have a randomized block but not strict factorial) and you acknowledge the problem in your write-up, perhaps in a footnote.

The rest of this chapter will presume that you have chosen effective laboratory procedures and descriptive data summaries, so that the measurement error consists mainly of minor round-off errors.

Sampling error requires a more extensive discussion. Perhaps the best way to think about it is to consider the collection of all possible trial outcomes given your treatments. Since the time of Galton's classical studies of physical characteristics in human populations, this hypothetical collection is called the *population* of outcomes. There is always some variability in the population because of uncontrolled nuisances such as subjects' attention to the task. You may prefer to think of the variability as "random fluctuations." For any given set of treatments, the variability induces some distribution for the possible outcomes. Logically enough, the induced distribution is called the *population distribution.* If you knew the population distribution, your inferential task would be trivial. For example, if the population mean for DA efficiencies were larger than the population mean for CH efficiencies, then you would correctly conclude that the DA institution is on average more efficient.

Nontrivial statistics are necessary because the population distribution can never be known precisely. Your budget and patience, however large, will allow you to run only a finite number of experiments; you can never observe outcomes of *all possible* trials. Nevertheless you do have useful information about the population distribution because you have actually run a subset of all possible trials and have recorded the outcomes. Thus your actual data constitute a finite *sample* from the population distribution. Sampling error, the second source of experimental error, arises to the extent that your sample is not representative of the underlying population. In the DA versus CH example, the mean of your DA-efficiency sample will almost always differ from the true mean of the DA population, and similarly for the CH sample. These sampling errors could be large enough to lead you to the wrong conclusion about which institution is more efficient.

7.2.2 Good samples and bad samples

You cannot expect to get a perfect sample, whose distribution exactly reproduces the population distribution. But with some care you can minimize sampling error within the bounds of your finite resources. That is, you can take steps to avoid bad samples and to get good samples.

There are two main ways of getting good samples. The first is to make the sample as close as possible to a classic *random sample,* in which each observation is independently selected from the population distribution. That is, in a random sample, each point in the population has an equal chance of being selected in each observation. The other way is to try to take a "stratified" or *balanced sample,* in which you subdivide the population into several segments and draw observations from each segment with frequency proportional to the weight of the segment in the population distribution. For example, in a voter survey (a field experiment) each interviewee could be drawn from the voter population by some random device such as throwing a dart at a printout of registered voters. This procedure could give you a truly random sample. Professional interviewers usually prefer a balanced sample, in which they segment the population by age, sex, education, location of residence, or other observable variables, and then select a proportionate number of interviewees from each segment. A balanced sample will tend to produce smaller sample errors than a random sample of the same size to the extent that outcomes differ across segments, the segments are observable, and their weights in the population are known. Otherwise, random samples are preferable.

Finding procedures that give you good (random or balanced) samples is not always easy. The general problem is that there may be unrecognized relationships among relevant variables in your experiment so that your data represent a small and atypical portion of the population rather than the population as a whole. For example, suppose an experimenter wants to measure the degree of altruism in individual subjects. If he selects subjects in the usual way, advertising the opportunity to earn "substantial cash rewards" in undergraduate economics classes and signing up volunteers, his altruism measurements probably will not be typical of the population of U.S. residents. He failed to recognize the possible relationship between the variables [attends economics class] and [responds to advertisement promising cash] and the outcome [measured altruism]. As a result, he probably collected an unbalanced, nonrandom *biased* sample.

Perhaps the most important advantage of experimental data is that it can provide better samples than happenstance data. Two examples of bad samples of happenstance data may help drive this point home.

Bad Happenstance Sample 1. A bank analyst wants to estimate his bank's profitability in its major loan categories: real estate,

commercial/industrial, and consumer. When he regresses historical bank profits on quantities (amounts outstanding in each loan category) he gets unstable coefficient estimates – the magnitude and even the sign change when he varies the beginning or ending dates of the historical data or when he switches from monthly to quarterly data. The underlying problem turns out to be that the bank's policy has been to keep tight bounds on the portfolio composition. For example from 1970 to 1985 real estate loans were not allowed to exceed 30 percent of the loan portfolio and never fell below 27 percent. The historical data therefore all come from a thin slice of the hypothetical profitability population, and as a result the separate effects of the explanatory variables (the loan categories) can't reliably be estimated from this unbalanced and nonrandom sample. Perhaps the analyst will have better luck with his statistical analysis if he can find similar banks with different portfolio policies and can construct a balanced sample from the combined data.

An econometrician would call Sample 1 a case of insufficient variation or multicollinearity. The problem need not arise from deliberate policy. For example, the historical capital/labor ratio and the factor price ratio might be almost constant in an industry, precluding good estimates of the elasticity of substitution from historical data. Since focus variables generally are controllable in the laboratory, you can avoid bad samples of this sort by choosing good experimental designs. Factorial and related designs covered in Chapter 3 ensure that the focus variables vary independently and over a sufficient range so that you can assess their effects.

Bad Happenstance Sample 2. An antitrust analyst studies the relationship between concentration and price over time in several narrowly defined industries. To her surprise she finds several industries for which periods of lower prices seem to go with periods of greater concentration (i.e., fewer competing firms). Further investigation discloses that in most of these cases both price and concentration were driven by a third variable, the price of related goods. For example, in the slide rule industry, price decreases and increasing concentration were both consequences of dramatic reductions in the price of electronic calculators.

An econometrician probably would call this an omitted-variables problem or an identification problem, and could provide a long list of related examples. The historical price data for slide rules were a biased sample of their concentration-segmented population distribution because the demand-side relationship with the electronic calculator price (or at least its impact on slide-rule quantity demanded) was not recognized. The sample suggests an incorrect inference because the price observations for high concentration were taken from the part of the population distribution associated with low demand.

Good experimental technique can prevent most problems of this sort. The experimental analog of the antitrust study would vary the focus variable (concentration) independently of the other controllable variables, including most variables which could shift demand. Randomization would neutralize the effects of other nuisance variables on the measured outcome (price). The result would be a good sample from which valid inferences could be drawn.

Despite the tremendous advantages laboratory techniques provide in creating good samples, some serious problems remain, arising particularly from learning effects and group effects. Human subjects usually learn from experience. The action a subject takes in a particular trial of an experiment may be affected by her experience in previous trials. To the extent that this sort of learning affects your measured outcomes, your sample is not random. Specifically, the trials in a single experimental session are not independent.

Group effects can also produce samples drawn disproportionately from a small subset of the population distribution. For example, in two recent double-auction market sessions with inexperienced subjects, the group of subjects in one session consistently produced more bids and fewer asks than the outwardly identical group in the other session.

In principle, the proper way to deal with these problems is to characterize the nature of sample dependence and to adjust the statistic accordingly. Beginning econometrics students learn how to deal with serially correlated time series data in just this way. Unfortunately learning and group effects have not yet been characterized with any precision, so no valid statistical correction presently is available.

Some experimentalists recently have dealt with the problem by adopting a very conservative definition of a trial – for example, count only the last (or next-to-last) period in a market session. This may be the only practical thing to do when learning effects are extreme, but we do not recommend the practice in general. The approach ignores a lot of potentially informative data, and doesn't completely cure the problem anyway–there may be group effects (or subject pool or protocol effects) that extend across sessions conducted in a given laboratory. Rassenti, Reynolds, and Smith (1988) (and some older unpublished work) deals with the problem by assuming learning effects take the form of exponential decay toward a behavioral equilibrium. We regard this approach as promising but unproven.

We have three recommendations. First, encourage your econometrically inclined colleagues to work on the problem; it probably is important in some of their favorite field data as well as in most laboratory data. Second, include appropriate caveats when you report formal statistical tests. For example, in an ABA crossover design, learning and group effects may tend to drive the observed A mean toward the observed B mean, so conventional confidence levels then would represent a lower bound on the true confidence level associated with your hypothesis test. In the bid/ask example given previously, the conventional confidence level for rejecting the null hypothesis (equal bid/ask ratios across the experiments) would represent an upper bound on the true confidence level. We recommend that you think through the uncontrolled nonrandomized nuisances in your experiment and, if you consider them significant, tell your readers the direction of probable bias in formal test statistics.

Our third recommendation is to extend your randomization scheme to different subject pools, different laboratories, and so on, whenever feasible. The folk wisdom among experimental economists is that an empirical regularity becomes credible when it is replicated with three different groups of subjects, preferably from different pools and in different laboratories. While we see no magic in the number 3, we endorse any procedures that broaden your sample of the population distribution.

7.3 Reference distributions and hypothesis tests

Hypothesis tests assess the probability that differences in observed outcomes across treatments are due to sampling error rather than due to differences in the underlying population distributions. Such an assessment requires a *reference distribution,* an empirical counterpart or proxy for the population distribution. You may construct a reference distribution directly from the samples themselves or from some external data source. Whether your source is internal or external, you may or may not decide to impose a parametric structure on the reference distribution. Your choice of reference distribution largely determines your choice of test statistic, and therefore the power and robustness of your results.

7.3.1 Internal reference distributions

The most common choice is an internal parametric distribution, usually the normal or the Student *t*. For example, suppose you wish to see whether subjects in a game theory experiment are equally likely to choose each of their two available pure strategies, $x = 0$ or 1. You can impose the parametric structure that the mean choice $\bar{x}$ is normally distributed with unknown population mean μ and known variance s^2/n, where n is the sample size and $s^2 = \Sigma_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$ is the usual variance estimate. Under the null hypothesis that the population mean is 0.5, the normalized sample mean $z = n^{1/2}(\bar{x} - 0.5)/s$ has the unit normal distribution. An observed $\bar{x} = 0.6$ from a sample of size $n = 36$ with $s = 0.2$ yields $z = 6(0.1)/0.2 = 3.0$. Tables show that the probability of drawing an observation $|z| \geq 3.0$ from the unit normal distribution is only about 0.0026 (a two-tailed test) and the probability of drawing a $z \geq 3.0$ is about 0.0013 (a one-tailed test). It is better to use the more powerful one-tailed test whenever you can specify the direction of the effect of treatment. Here you can confidently reject the hypothesis that the true population mean is 0.5 and that the observed sample mean of 0.6 was due solely to sampling error.

Of course, the test just described assumes you know the population variance. In practice, you usually only know the sample estimate s^2. The internal parametric reference distribution based on a normal population with unknown mean and unknown variance is called Student *t*, after the pseudonym adopted by the statistician William S. Gossett (1876–1937). In a *t*-test you compare the same normalized sample mean $n^{1/2}(\bar{x} - 0.5)/s$ to tabulated values for the Student *t* distribution with $\nu = n - 1$ degrees of freedom. In the example with $t = 3.0$ and $\nu = 35$, we get one- and two-tailed probabilities of about 0.0025 and 0.005. The probabilities are about twice as large as with the normal reference distribution, but they are still small enough for you to reject confidently the null hypothesis.

You can use more elaborate formulas but the same logic to test hypotheses of the form "treatment A promotes higher performance than treatment B." Assume that measured performance is normally distributed with unknown mean μ_A (μ_B) under treatment A (B) and that the unknown variance is the same under both treatments. Then the "pooled *t*" statistic

$$t_p = (\bar{x}_A - \bar{x}_B) / (s\,(1/n_A + 1/n_B)^{1/2}),$$

where the sample sizes are n_A and n_B, and the combined sample variance is s^2, has the Student *t* distribution with $\nu = n_A + n_B - 2$ degrees of freedom.

If you designed your experiment so that A and B trials occur in *n* matched pairs, you can sharpen the test. Form the matched pair differences $x_D = x_A - x_B$, and compute their mean $\bar{x}_D$ and variance $s_D{}^2$. Then form the "matched *t*" statistic, $t_m = n^{1/2}(\bar{x}_D)/s_D$. For sufficiently large values of either t_m or t_p you can confidently reject the null hypothesis that the A and B populations have the same distribution.

A numerical illustration may be in order. Recall the boys' shoes example of Section 3.3, in which we want to know whether the new sole material A wears more slowly than the old material B. The data reported in Box et al. (1978, p. 100) give sample sizes of $n_A = n_B = 10$, sample means of measured wear of $\bar{x}_A = 10.63$, $\bar{x}_B = 11.04$ (so $\bar{x}_D = -0.41$), with $s = 2.43$, and $s_D = 0.386$. Then $t_p = (10.63 - 11.04/(2.43/5^{1/2}) = -0.41/1.09 = -0.38$, while $t_m = (10^{1/2})(-0.41)/0.386 = -3.36$. Tables of the Student *t* distribution give one-sided 1 percent critical values of 2.25 for the pooled *t* ($\alpha = 0.01$, $\nu = 18$) and 2.82 for the matched pair *t* ($\alpha = 0.01$, $\nu = 9$). Since the absolute value of t_m exceeds the critical value, we conclude that the new material A wears significantly more slowly.

Why did we pose "no effect" as the null hypothesis and the effect we were looking for as the alternative hypothesis? This is the customary way to do it. Although you can find an occasional counterexample in the literature (e.g., Schotter and Braunstein, 1981; De Long and Lang, 1992), it usually is considered bad form to reach a conclusion by failing to reject the null hypothesis. Perhaps you failed to reject because the data are sparse or noisy, not because the null hypothesis really is correct. Your readers will probably find it more satisfying if you reach your conclusion by rejecting a boring null hypothesis in favor of your desired (often one-sided) alternative hypothesis, as in the example. Why use a 1 percent confidence level? Custom again. Smaller confidence levels are better, since we are talking about the probability of mistakenly rejecting a true null hypothesis. Economists often will settle for a 5 percent or even 10 percent confidence level when working with a small or noisy data set, but everyone prefers a 2 percent or 1 percent confidence level when the data are reasonably good.

Why were we able to reject the null using the matched *t* but not the pooled *t* statistic in the example? Recall that the matched-pair design, assigning materials A and B randomly (to left and right or right and left) shoe soles, is intended to eliminate experimental error due to nuisance variables. The sharp decrease in s_D relative to s, and therefore the sharp increase in t_m relative to t_p, demonstrates the success of the matched-pair design in this example.

The reference distributions discussed so far assume that the underlying populations are normally distributed. The Central Limit Theorem provides some justification for assuming that the mean of a random sample is normally distributed, even when the observations themselves are not drawn from a normally distributed population. Nevertheless, the normality assumption remains unattractive in some cases. For example, the period-by-period market efficiency data in Figure 7.3 certainly are not even approximately normal. More extreme examples occur when your equilibrium occurs at a corner, so deviations can't even be symmetric. (See Chapter 9 for an example called Bernoulli-choice experiments.) In such cases you may prefer to use a free-form (or *nonparametric*) reference distribution in testing the null hypothesis that treatments A and B yield the same population distribution of outcomes. The idea is that if the null hypothesis is true, then each assignment to A or B trials of the measured outcomes is equally likely. The reference distribution then consists of all possible assignments of the data to the treatments, and the test statistics give the probability that a difference between the A and B trials at least as extreme as observed could have come from a random assignment.

The Wilcoxon (or Mann-Whitney U) statistic is perhaps the most popular example of a nonparametric test. You (or preferably your computer programs) rank-order the data from lowest measured efficiency to highest, keeping track of whether each trial was an A or B treatment. Then you sum the ranks S for the (say) A trials. The statistic S has known mean and variance under the null hypothesis of no differential effect when there are an equal number n of observations under the A and B treatments, so the distribution of the statistic $T = \text{mean/variance}^{1/2}$ is approximately unit normal in large samples. Good statistical programs can compute the exact probabilities (confidence levels) for any T-value even in moderate-sized samples, and in samples of unequal sizes. A useful variation of this Wilcoxon test, explained on p. 226 of Conover, allows you to test the null hypothesis of equal variances instead of the usual null hypothesis of equal means.

Another popular statistic, called the binomial or signs test, uses a nonparametric reference distribution which is especially useful for matched-pair data. You (or the computer programs) count the number r of paired differences that are positive and the number w that are negative. Under the null hypothesis that positive and negative differences are equally likely, r has a binomial distribution with mean $0.5n$ and variance $n(0.5)(1 - 0.5)$, where $n = r + w$. A little algebra then shows that normalized sample mean is $z = (r - w)/(r + w)^{1/2}$. This statistic is approximately unit normal in large samples; its exact binomial distribution can be calculated precisely in small samples. (It is customary in small samples to subtract the "continuity correction" 0.5 from the numerator.) Once again, you can reject the null hypothesis of no differential effect in favor of the hypothesis that A leads to larger observations than B if z is sufficiently large.

The Wilcoxon test is computationally simple and the binomial test is even simpler. But the Wilcoxon test keeps track only of ordinal relationships and ignores quantitative sample information, and the binomial test ignores all sample information except the signs of the matched-pair differences. Ignoring information reduces power of the test. In the present era of cheap computing power it is worth considering nonparametric procedures that are computationally demanding but use all sample information. The prime example is called the *bootstrap*. To illustrate, suppose your data consists of five matched pairs (x_{Ai}, x_{Bi}), $i = 1, \ldots, 5$. Construct an internal reference distribution of hypothetical data by taking all permutations of the actual data. Thus you have $2^5 = 32$ hypothetical sets of matched pairs, one of which is the actual data. For each hypothetical data set h, compute the difference of means $\bar{x}_A^h - \bar{x}_B^h$; these thirty-two differences form the reference distribution for the actual difference $\bar{x}_A - \bar{x}_B$. The fraction of the hypothetical differences that exceed the actual difference is the confidence level with which you can reject the null hypothesis of no difference in favor of the alternative hypothesis that $\bar{x}_A > \bar{x}_B$.

You can also bootstrap unmatched data. Given n A-observations and m B-observations, there are $(n+m)!/(n!m!)$ hypothetical assignments of the $n+m$ actual observations to the two treatment levels with n assigned to A and m to B. Under the null hypothesis of no effect, the set of hypothetical A-means (B-means) defines a reference distribution for the observed A-mean (B-mean). (See Box et al., 1978, p. 97, for a numerical example.) The bootstrap reference distribution converges to the t-distribution as the sample size increases, but gives more accurate confidence levels in small samples.

7.3.2 External reference distributions

Sometimes theory prescribes a specific reference distribution. For example, you may conduct a k player game experiment where the payoff function has a unique mixed-strategy Nash equilibrium $p_1, \ldots, p_k$. Then you probably want to test the hypothesis that observed strategy frequencies $n_1, \ldots, n_k$ represent $N = n_1 + \ldots + n_k$ independent draws from the reference distribution $p_1, \ldots, p_k$ – that is, that your subjects all play the Nash-equilibrium strategy. A standard test is to compute the normalized sum of squared deviations

$$C = \sum_i \frac{(\frac{n_i}{N} - p_i)^2}{p_i} .$$

It turns out that C has the Chi-squared distribution with $k - 1$ degrees of freedom, so you locate your computed value in a standard table to determine the confidence with which you can reject the null hypothesis.

The origin of external reference distributions can be empirical rather than theoretical. Suppose, for example, you run experiments parallel to the extensive published work of Professor Jones. Using her published data (request raw data from her directly if the published data are inadequate), you can estimate the parameters of an appropriate distribution (e.g., normal or binomial) and use that fitted distribution as your reference distribution. Then go ahead and see if you can reject the usual sort of null hypothesis–for example, that the mean of your data is the same as the mean of her data (the reference mean). Alternatively, if your software permits, you can use the exact empirical distribution of her data as your reference distribution. Then you can run the usual nonparametric tests, such as the Wilcoxon and the bootstrap, to see whether you can reject the usual null hypothesis. Failure to reject the null hypothesis in this case is evidence that you successfully replicated Professor Jones's results.

7.3.3 More statistical tests

The test statistics mentioned so far – the normalized sample mean, the pooled t and matched t, the Wilcoxon T, the binomial z, and the Chi-squared statistics – are not the only ones useful for hypothesis testing. To begin with, the Chi-squared statistic is handy even in the absence of a theoretical reference distribution. For example, you may want to see whether treatments such as instructions or feedback information affect the strategy frequencies in your game-theory experiment. The standard approach is to write out a contingency table (columns defined by treatments and rows by strategies) and calculate a Chi-squared statistic analogous to C for the entire table; large values allow you to reject the null hypothesis that the treatment had no effect.

There are many other statistical tests associated with contingency tables. Perhaps the best known is Fisher's exact test. It is appropriate for contingency tables where both row totals and column totals are constrained by your design and/or by the nature of the task. See Chapter 4 of Conover (1980) for a clear exposition.

There are several general-purpose test statistics that compare an empirical distribution to a reference distribution. The Kolmogorov-Smirnoff statistic measures the maximum distance between the two cumulative distribution functions; you can reject the null hypothesis that the underlying population distributions are the same for sufficiently large values of the test statistic (Conover, ch. 6).

The tests mentioned so far deal only with a single treatment variable. Suppose your experiment features several treatment variables and you are satisfied with a (multivariate) normal reference distribution. Then you can use the classical analysis of variance (ANOVA) procedures. ANOVA allocates the variance in your data to each treatment variable and to residual variance. Appropriate variance ratios have the F-distribution (discovered by R. A. Fisher, of course) under the null hypothesis that the treatment variable has no effect. Thus, you can get ratios for each treatment variable and compare them all to tabulated critical values of the F-distribution to determine which of your treatment effects are significant. For details see any statistics text used by social scientists other than economists.

Most economists are more familiar with multiple regression than with ANOVA. Fortunately, you can get equivalent test statistics from multiple regression because ANOVA is a special case of the general linear model (see Kirk, 1982, ch. 5). The regression for two-level treatment variables is simple. Just define a 0-1 dummy variable for each treatment variable, and regress your data on a constant and the dummies. The estimated coefficient for each dummy is the mean effect of the corresponding treatment, and its t-statistic is the standard t test statistic for the null hypothesis that the treatment variable has no effect. If your design kept the treatment variables orthogonal, then these t tests are independent and the results will not be affected when you omit or include other treatment variables in the regression.

The discussions in this section focus on hypothesis testing for treatment variables. The ideas apply equally well to comparing alternative models, say models A and B. Let x_{Ai} and x_{Bi} be the forecast errors of the two models for predicting observation i. Then you can use all the matched-pair tests as well as the more general tests to try to reject the null hypothesis that the A-errors have the same distribution as the B-errors.

A final remark on statistical technique. This chapter has emphasized classical hypothesis testing and estimation because these are widely used by economists and better suited to experimental data than to happenstance data. You should also be aware that there are numerous Bayesian techniques. Roughly speaking, these techniques summarize the empirical evidence by mapping prior beliefs (before exposure to the data) into posterior beliefs (after digesting the data). Bayesian techniques generally are more consistent with decision theory and eventually may replace classical statistical techniques, but at present are not standardized for experimental (or even happenstance) data. Therefore we omit coverage, and refer the interested reader to Leamer (1978) for a general position statement and to Boylan and El-Gamal (1992) for a recent application to experimental data.

7.4 **Practical advice**

Data analysis interacts with experimental design, and you should think through both before you start conducting your experiments. Specifically,

1. Choose your laboratory protocols to reduce measurement error – automate data capture where possible, build in redundancy, and so forth. In manual experiments, have two persons record the data independently. See Section 6.11 for further suggestions.
2. Choose your treatments to produce good samples. Pay special attention to possible learning effects and group effects, since these nuisances are difficult to control or randomize. Remember that fancy statistical procedures are a poor substitute for good samples.
3. Choose experimental designs that will allow you to employ efficient statistics, such as designs that produce matched-pair data, or designs with orthogonal treatment variables.

Once you have conducted your experiments and have gathered the data, you should begin with a qualitative data analysis. We recommend that you

4. Search published literature and use your imagination to find effective graphical displays and summary statistics. Try out several possibilities before making your final choices. Popular worksheet software (Lotus, Quattro, Excel, Wingz, etc.) are well suited for this task.
5. Look for outliers and other irregularities in the data. Eliminate those due to measurement error, and think about possible causes of the correctly reported irregularities (and regularities).

If skeptical colleagues find your conclusions obvious from your qualitative analysis, then you are ready to get on to your final write-up. Usually you will run some formal statistical tests to better understand what your data have to say. If so,

6. Look for appropriate external reference distributions, arising from theory or from existing data. If external reference distributions are unavailable or insufficient, use standard parametric and nonparametric internal reference distributions.
7. Conduct the relevant hypothesis tests or equivalent parameter estimation procedures (regressions). Include a caveat if you suspect your design hasn't fully controlled for or randomized out group or learning effects.

7.5 **Application: First-price auctions**

The practice of selling an object to the highest bidder in an auction goes back to ancient times, but no satisfying theoretical analysis of this practice appeared until Vickrey (1961). His approach was to postulate what is now known as independent private values: Each bidder i knows her own value v_i and regards the unknown values of the other $n-1$ bidders as if drawn independently from some specific distribution. Vickrey then used what now is called Bayesian Nash equilibrium to predict the bids and the outcome of an auction. Assuming that traders are risk neutral and that the specific distribution is uniform on an interval $[0, \bar{v}]$, Vickrey predicted that anyone with a value of v_i would bid $b(v_i) = (n - 1)v_i /n$. This result applies to first price-sealed bid auctions (once-and-for-all bids are submitted privately and the highest bidder pays his bid price for the object) and some other outwardly different auctions such as the Dutch auction (the first bidder to stop a declining price clock gets the object at the indicated price).

After a gestation period of a decade or two, Vickrey's model spawned a large body of theoretical literature, surveyed in McAfee and McMillan (1987). Experimentalists quickly noticed that this theory had sharp predictions and important applications but was difficult to test in the field. Building on Coppinger, Smith, and Titus (1980), the study by Cox, Roberson, and Smith (1982) analyzes bidding behavior in first price and other auction institutions. The treatment variables also include the number n of bidders and the upper endpoint $\bar{v}$ on the uniform distribution of private values. For each subject, the authors separately regress the bids b_i on a constant and the values v_i, and they tabulate mean price and price variance. To compare the price data to theoretical predictions, the authors rely on a Kolmogorov-Smirnoff test; the only graph in the paper illustrates the K-S test. They also use a binomial test to compare behavior across auction institutions. The authors conclude that the first-price auction data are not consistent with the original Vickrey (1961) model, which assumes risk-neutral bidders, but generally are consistent with extensions of the model that assume uniformly risk-averse subjects.

Follow-up studies extend the environment and institutions in various ways. The most thorough report on first-price auction experiments is Cox, Smith, and Walker (1988). In one short table they summarize the outcomes of 690 auctions from 47 previous experiments. The table segments the sample into 8 subsamples according to the number of bidders and other design features (such as whether the session involved an alternative auction institution in an ABA crossover design.) For each subsample the summary statistics are the mean observed price and its deviation from the Vickrey prediction. The table also reports the t-statistic for the null hypothesis that the mean deviation is zero. The null is rejected in 7 of the 8 subsamples in favor of the alternative that price exceeds the Vickrey prediction, a result consistent with risk-averse bidding.

The authors then pursue the risk-averse bidding hypothesis by examining individual behavior. Relying on a Wilcoxon test to compare each subject's bids to the Vickrey predictions, they reject risk-neutral bidding in favor of risk averse bidding for a majority of subjects. Graphs of the points (v_t, b_t) for individual subjects suggest that subjects differ in their apparent degree of risk aversion. To pursue this possibility, the authors regress bids b_t on a constant and value v_t separately for each subject, and tabulate the estimated slope coefficients and intercepts. They also graph cumulative distribution functions for the regressions' R^2 and for F statistics across pairs of regressions. The results support the view that behavior differs significantly across subjects.

Preexisting theory did not consider heterogeneously risk-averse bidders, so the authors construct a Bayesian Nash equilibrium bidding model called CRRAM (for constant relative risk aversion model) that covers this case. They find that the existing data are generally consistent with CRRAM. Since the model was constructed to explain the existing data, the authors conduct new experiments to test the model further. CRRAM correctly predicts that tripling monetary rewards has no significant effect on the bid functions. It is less successful in predicting changes in bid functions when rewards are nonlinearly transformed. CRRAM also fails to account for nonzero intercepts in bid functions in the original data. In a final iteration of theory and experiment, the authors construct modified versions of CRRAM which allow nonzero intercepts. One version, called CRRAM*, is generally consistent with the existing data as well as with data from new experiments designed to test it.

Surely this is an impressive body of scientific research. Nevertheless it is under attack on two fronts. Skeptics can question whether the departures from Vickrey behavior really are significant and, if they are, whether alternatives other than risk aversion have received adequate consideration. Harrison (1989) forcefully argues that departures from Vickrey behavior are negligible and therefore the dominance precept is not satisfied. To make his case, Harrison presents several diagrams showing that unilateral deviations from the Vickrey bid function typically result in rather small expected losses. He points out that the deviations are highly non-normal and so he relies mainly on nonparametric statistical techniques. He finds that the (true, population) median expected loss is very likely to be less than 8 cents per bid. Other critics disagree with Harrison's emphasis on median losses and point out that even a robust statistic may not capture key features of the data (i.e., a moderate number of large losses would not be detected by the median if there are enough small losses). Some critics argue that learning explanations may improve on the risk-aversion explanations. Readers interested in the substantive issues raised by first price auction experiments should read the Kagel (1993) survey and the December 1992 *American Economic Review* interchange on Harrison (1989).

8

Reporting your results

You have thought through some important economic issue, found a way to examine it in the laboratory, designed an appropriate set of experiments, run them, and analyzed the data. You have learned a lot through the whole process, and it appears that the results may interest, even surprise others. Time to kick back and congratulate yourself on a job well done? Well, don't relax quite yet. You still have to present your results to your peers. If your write-up is sloppy or confusing, all your hard work probably will have no impact on others. If you report your results effectively, you may help people change how they think about the issue. You already have had the personal satisfaction of learning something new. Now by effectively communicating this learning to others, you can amplify the social benefit of your work as well as your personal satisfaction.

This chapter offers suggestions on how to report the results of your experiments effectively. We emphasize the preparation of articles for academic journals, but most of the suggestions apply equally well to seminar presentations, consulting reports, or book chapters. The first section discusses the scope of research you should try to cover in a single paper. Next we present customary ways of organizing the paper, and offer advice on polishing your prose, tables, and figures. The rest of the chapter discusses current standards for documenting your work and offers advice on how to schedule various stages of your project. We illustrate many of our points in a discussion of asset-market experiments.

8.1 Coverage

Every essayist, whether an economist, or journalist (or physicist for that matter) must decide *what material* to cover and at *what depth*

to cover it. Coverage decisions can be particularly difficult for experimental economists. Usually you will get some puzzling results in your initial laboratory sessions, so you conduct follow-up sessions. Often the new results create as many puzzles as they solve, so you conduct more follow-up sessions, creating new puzzles, and so on. The process eventually terminates, either because you resolve all the important puzzles or (more likely) because you run out of time, money, or patience. At this point you may have far more material than you can fit into a single paper, but the scope of this material is probably too narrow for a publishable book. Somehow you will have to select a subset of your material.

In choosing which data to report you must balance two conflicting objectives. First, to keep your readers' attention and to aid their retention, you want to focus on a single issue or a small set of closely related issues. Therefore you want to select only the most directly relevant data. Second, you want to present an accurate and complete picture of your results. In particular, you want to avoid selection biases.

Roth (1990), taking a cue from Leamer (1983), warns that experimentalists too are susceptible to selection biases in reporting their results. He argues forcefully in favor of treating the entire set of trials in an investigation as a single experiment. If the designation of "experiment" were reserved for various subsets of trials, he argues, investigators might be tempted to report selectively from the trials they have conducted, with dysfunctional consequences for the discipline as a whole. However, Roth acknowledges the other side to the argument by quoting the example of Robert Millikan and Felix Ehrenhaft from a report by the National Academy of Science's Committee on the Conduct of Science (1989). Ehrenhaft reported all his data and concluded, incorrectly, that there is no lower limit on the magnitude of electrical charge found in nature. Millikan, on the other hand, used only what he regarded as his "best" data sets to demonstrate the unitary charge of electron, and went on to win the Nobel Prize for this landmark discovery.

How should you resolve the data-selection dilemma? We believe that within your budget and time constraints you should vary treatments and replicate sufficiently to obtain a reasonably broad base of valid data, and you should analyze all of it until you understand its main characteristics. Then you should select the most relevant portion of the data for closer analysis, after satisfying yourself that your selection does not distort the conclusions. In your written report you should briefly but carefully *describe your selection process* and then devote most of your report to analyzing the data selected. That way your readers can judge the relevance of your data for themselves, and know where to go for additional evidence. Our advice admittedly places a heavy burden on you, the experimentalist, but we think the burden is justified because the scientific validity of your results is at stake.

The decision regarding depth of coverage also must balance conflicting needs. First again, you want to be brief and not tax your readers' patience with dispensable details. But second, you want to be sufficiently complete so readers understand what you have done and how you reached your conclusions. Many of your readers probably are not as familiar with your procedures as they are with standard econometric procedures for field data. Consequently, they may misinterpret what you did if you omit too many details.

With some extra work, you can resolve this conflict satisfactorily. In the text of your paper, try to convey the main features of your procedures and omit most of the details. But in an appendix, write up your procedures in sufficient detail that any competent experimentalist could fully replicate your work, and make the appendix available on request. In doing so, you will assist your fellow experimentalists, depersonalize the empirical basis of economics, and strengthen its scientific foundations. To drive the point home, we reprint the *Econometrica* guidelines in Appendix IV. These guidelines should generally be followed even if you have no intention of submitting your work to that journal.

8.2 Organization

Your experimental paper should be organized generally in the same manner as other empirical economics papers. In recent decades, empirical papers in economics usually have the following organizational plan:

Part A Introduction. Statement of issues, background information, literature survey, overview of the paper and results.
Part B Relevant theory. A brief summary often suffices.
Part C Data and results.
Part D Conclusions and discussion.

Experimentalists face some expositional issues that other empirical economists usually can ignore. If you present theory before describing your laboratory environment, you are left to defend the gaps between the two. You may prefer to describe your laboratory environment, institutions, and treatments first, before specifying the theoretical models that may be relevant to understanding the outcomes of such economies. This is especially useful if the relevant theory is poorly developed. Presentation of data and results also requires careful exposition because typically your data are new and in some respects unfamiliar to most of your readers.

Experimental economists generally deal with these expositional problems by modifying the basic organizational plan as follows.

Part A Introduction. Statement of issues, background information, literature survey (may go elsewhere), overview of the paper and results.
Part B1 Laboratory procedures. Basic environment and institutions, treatments, design, subject pool, etc.
Part B2 Relevant theory. Can precede B1 if relevance is clear from introduction. May conclude with a list of testable hypotheses.
Part C1 Descriptive data analysis. Graphs and summary statistics.
Part C2 Inferential data analysis. Hypothesis tests or the like. May be omitted if conclusions are evident in the descriptive data analysis.
Part D Conclusions and discussion.
Appendices Instructions to subjects, raw data, mathematical derivations, procedural and statistical details, etc. To be published if the editors desire, otherwise available on request.

This outline is for pedagogical purposes only. It is best to think about our outline and to look at the organization of good published articles that are relevant to your work. Then choose a tentative organization and modify it in response to colleagues' comments that make sense to you.

8.3 Prose, tables, and figures

For reasons we do not fully understand, wordsmithing standards seem higher in economics than in most other experimental disciplines such as psychology and biology, and most economists spend a lot of time polishing their prose. Unless you don't care about publication, or unless you are a gifted writer, you also will devote a large fraction of your research time to prose polishing. Remember that if better writing makes your work accessible to even 10 percent more readers, the return is well worth the investment. You should expect to rewrite your paper several times before you are done with it. It may help to ask yourself the following questions as you work on your prose.

Did I leave out any information my readers need to understand this sentence or result?
Have I repeated myself too often on this point?
Is there a way to rearrange the paragraphs or sentences to make the material easier to absorb?
Can I recast this sentence to make its meaning clearer on first reading? Did I slow the reader down by making gratuitous backward or forward references (e.g., "See Section 8.4 below")?
Is there a more apt or vivid way to make this point?

Good writing is an art. It does not come naturally to most economists (ourselves included), but we all improve with practice. You can increase your rate of improvement by reading Strunk and White (1979), McCloskey (1985, 1987), and Hamermesh (1992), and by taking their advice to heart.

Many readers will skim your article, pausing to look more closely at diagrams, graphs, and tables. Even careful readers usually depend heavily on figures and tables. Therefore the success of your paper depends disproportionately on the quality of your figures and tables, and you can get a high payoff from polishing them so they are easy to understand. As you polish, ask yourself the same kind of questions as for your prose. For example, do lines 5 and 6 of this table convey any useful information? Would a separate diagram help clarify this fundamental point? Do I have too many lines in this graph?

The *Journal of Finance* and a few other academic journals require that each table and figure be completely self-contained, suitable for reproduction in a textbook without your surrounding prose. In our view this standard is a bit extreme, but the general idea is a good one. Ask yourself: Will my readers remember the meaning of this acronym used as a column head? When in doubt, make the column heading self-explanatory or define it in a caption or note. And so forth. Good published work on related issues is the best source of ideas for improving your tables and figures. You may find Tufte (1983, 1990) useful as general references.

8.4 Documentation and replicability

Philosophers of science assign a central role to replicability. More specifically, in the opening paragraphs of the *New Palgrave Dictionary* entry on experimental economics, Smith (1987) explains why progress in our discipline depends on experimentalists being able to replicate one anothers' work. As an experimental economist, you have the responsibility of documenting your work so that it is replicable. Given your documentation and other necessary resources such as access to subjects or special software, another competent experimentalist should be able to conduct an experiment that you would regard as essentially the same as your own. Further, she should be able to process your raw data in the same way you did.

To meet this replicability standard, four types of documentation are necessary:

Subjects Maintain printed or electronic copies of instructions to subjects. Also, maintain records of how, when, and where you recruited and trained subjects. Your institution probably also requires you to maintain records of cash payments to subjects.
Laboratory environments Maintain copies of software and special materials, and descriptions (at least) of hardware you used, in sufficient detail that your laboratory environments could be recreated.
Raw data Keep electronic or hard copies of all your valid data. Include records of time and circumstance, such as a lab log.
Data processing Keep records of your specific procedures, such as the SAS (a popular statistical software) procedures used to produce Table 3 of your paper.

When you have finished your project, you should consider sending your data to a public archive. Some funding agencies, such as the National Science Foundation, require this. Many use the U.S. national archive of social science data maintained by the Inter-university Consortium for Political and Social Science Research (ICPSR). The mailing address is PO Box 1248, Ann Arbor, Michigan 48106-1248.

8.5 Project management

Unless you have previous experience, you probably feel a bit uncertain about how to combine planning, experimentation, data analysis, oral and written presentations, and documentation. You probably will begin and end these tasks more or less in the order listed, but there will be considerable overlap. We offer our advice on project management in the form of answers to several questions that may be on your mind.

When should I begin presenting my results? As soon as you have a reasonably broad set of valid data (i.e., without important glitches), you should begin to analyze it, and when you obtain an interesting result you should think about presenting it. Initial oral presentations usually are best made to an informal and friendly audience at the time you are finishing the first complete draft of your paper. Don't wait until you have highly polished results because then you would miss the opportunity to act on your colleagues' good suggestions.

Should I write up my results in one long paper or several shorter papers? Both of us tend to err on the side of putting too much material into a single paper, but we've certainly seen the opposite error as well. Remember that the scope of a paper is defined by the issues addressed, not the number of experiments. Basically, it is a judgment call. If you are unsure, ask your colleagues for advice.

When should I submit a paper for publication? Journal standards for experimental economics are the same for other kinds of empirical economics. Read Hamermesh (1992) and consult trusted colleagues if you are unsure whether your paper is ready for submission.

When should I make my documentation available to other experimentalists? The current custom is to offer all documentation except raw data on request as soon as you begin to circulate a draft or working paper version of your results. There is no consensus as yet on raw data; some experimentalists have delayed sending it for as long as two years from the time of initial publication of results. Others honor requests for raw data before publishing anything. You incurred the costs of producing the data so you deserve the right of first access. On the other hand, the full social benefits will be realized only when the data are available for cross validation, new tests by other investigators, and student training. We hope that it becomes customary to release data upon acceptance for publication or within a year of completion of the main experiments, whichever comes first.

8.6 **Application: Asset-market experiments**

Field data are exceptionally plentiful and accurate for asset markets. Every day there is a new mountain of precise price data for stocks, bonds, commodity futures, options, and foreign-currency markets. Despite their impressive mass and precision, the field data have some weaknesses. Trading volume data are reasonably good, but accurate allocation data are much harder to obtain. More important, traders' preferences, endowments, and information are not observable in field settings. Hence you can't directly measure allocational efficiency or the fundamental value (i.e., the value incorporating and aggregating all current information) for an asset market in the field. You can measure price volatility for field assets, but you can't determine how much of it

is efficient response to new information and how much of it is excessive and inefficient.

Laboratory asset market data have complementary strengths and weaknesses. Budgetary considerations dictate that only a few traders will participate in laboratory markets over relatively short periods of time. However, traders' preferences and information can be controlled, so you can measure efficiency directly. If you are interested in the effects of the trading institutions, you can systematically vary them in the laboratory. Experimental studies of asset markets were initiated to examine the abilities of markets to disseminate information and to allocate resources efficiently when the initial distribution of information is asymmetric. We shall describe only the main features of a few studies here. For a detailed survey, see Sunder (1993).

Plott and Sunder (1982) initially designed their experiment in 1980 to learn how large a fraction of traders must have information in order for the market to behave as if all traders are informed. The authors expected the results to show that, as the number of traders who have information at the outset increases, the allocative efficiency of the market will rise. This sort of quantitative link between initial information dissemination and market efficiency cannot be confirmed from field data because the researcher cannot know the information conditions of the individual traders.

Plott and Sunder (1982) made important abstractions and borrowed from the prior experimental studies in creating their laboratory model of the stock market for the purpose of testing the efficiency hypothesis. First, stocks have indefinite lives and pay periodic dividends whose amounts are uncertain. They abstracted away from indefinite lives to a single dividend because multiperiod lives were not critical to the principle of information dissemination in markets. Second, exploration of the issues of information efficiency needed uncertainty of payments, and they borrowed the design of uncertainty in their first market session from Plott and Wilde's (1982) experiment on professional diagnosis versus self-diagnosis. When this information structure proved to be too complicated, they simplified it in the subsequent market sessions. Third, an experimental model of the stock market had to permit each participant to be a buyer as well as a seller. This feature was borrowed from Forsythe, Palfrey, and Plott (1982). Each trader was given an initial endowment of two assets and a large working capital loan. The working capital loan enabled each trader to buy and sell freely within a trading period, though the net short sale within a period had been restricted to the initial endowment of two assets in order to limit the risk of subjects' bankruptcy. Fourth, the per unit dividends were specified so as to hold the rational-expectations equilibrium price to a constant level within each period. Fifth, dividends were varied across the three classes of traders in order to generate gains from trading and to enable a measure of allocative efficiency of the market to be defined and examined. Finally, information about the realized state of the world that determined the dividends was withheld from some traders in order to examine if these traders are able to learn the information through the market process itself.

Thus the focus variable in Plott and Sunder's (1982) experiment is information (i.e., prior notification of the realized state) with three levels: none, insiders (e.g., two of the four traders of each type are notified), and all. Nuisance treatment variables include the state probabilities and the state-contingent valuation schedules, and whether or not the number and identities of insiders are announced. Basically the design is randomized block, each block consisting of two to nine trading periods. The results supported the rational-expectations (RE) model, as prices and allocations converged to efficient levels and insiders' excess profits became insignificant.

While Plott and Sunder reported the results of all five market sessions, they also selectively used information from their early sessions to guide their exploration. Their results can be used to illustrate the critical and controversial nature of the issues discussed in Section 8.1. Only one out of a total of nine private-information periods of the first two market sessions betray any hint of information dissemination. Using the statistical averages, the null hypothesis of no-dissemination would not have been rejected. Yet, the behavior of market in period 9 of market 2 suggested that, under appropriate conditions, such dissemination might occur. The authors then conducted a third market session with experienced traders that yielded firm evidence in favor of information dissemination. Millikan's use of his "best data" can be an excellent example to follow if you apprise your reader of all the facts of the case.

Clear evidence of market efficiency from the third market session led the authors to seek replication in a fourth session with a fresh set of subjects. Having replicated, they wrote the first draft of the paper and presented the results at two workshops. Comments received at the workshops led to a fifth market session in which the number of states of the world was increased to three. Design, conduct, and presentation of the experiment took only six weeks, much less than the authors' other work.

Do the striking efficiency results stand up in more difficult environments? Having observed dissemination of information from the informed to the uninformed, Plott and Sunder (1988) designed an experiment to examine if, and under what conditions, the markets can

perform the more difficult task of aggregating diverse information in possession of individual traders. Can markets behave as if everybody has all the information? They took the three-state design of the fifth session of their 1982 paper and altered the information structure. If state *X* was realized, half the traders were told that the state is "Not *Y*" while the other half were told that it was "Not *Z*." Would the market behave as if every trader knows for sure that the state is *X*? Results of their initial sessions revealed the answer to be negative, and shifted the focus of research to finding market environments in which such aggregation can occur. The subsequent sessions revealed that information is aggregated in markets that fulfill either of the two conditions: (1) homogenous preferences (same dividend distribution for all traders) or (2) trading a set of securities that span the state space. In further work, Forsythe and Lundholm (1990) found that even in incomplete markets with heterogenous preferences, additional trading experience can lead to information aggregation.

Unlike their 1982 paper, market sessions for Plott and Sunder (1988) were conducted over a span of three years at geographically dispersed locations. The first market session was found to aggregate information only because, it was later discovered, one subject was inadvertently given information she should not have had. This session was excluded from the published work. The working versions of both papers included complete raw data appendixes which were later analyzed in published articles by other authors.

Copeland and Friedman (1987) report the first computerized asset-market experiments. (See Williams, 1980, and Anderson et al., 1989, for evidence that computerized asset markets are more difficult than oral.) Their environments had several dimensions of additional complexity including news (i.e., information regarding the realized state arriving during the trading period), and possibly heterogeneous states. To cope with the large number of potentially important nuisances they employed a 2^4 half-factorial design with the fourth variable confounded with the three-way interaction of the other variables. In this and later work, the authors found that the rational-expectations model continues to outperform alternative simple models in most dimensions, although there are some interesting anomalies. Two follow-up papers by the same authors examine the interaction of an information market with the asset market, and examine an empirically oriented model of partial information aggregation. After several rejections and numerous revisions, the papers eventually were published in 1991 and 1992.

Smith, Suchanek, and Williams (1988) draw quite different conclusions from a different environment examined in dozens of experimental sessions over several years. They report frequent large bubbles – episodes where the asset price rises far above the fundamental value for an extended period of time, usually ending in a sudden price crash to or below the fundamental value. The environment differs from most previous asset-market studies in at least two respects: They generally have only one trader type (so there are no induced gains from trade), and they use long-lived assets with little stationary repetition. Despite some useful follow-up work by Porter and Smith (1990) that systematically tests several hypotheses regarding bubble formation, it is not yet clear which design differences are responsible for the inefficient prices. Follow-up work continues in several laboratories around the United States.

9

The Emergence of experimental economics

Why has the experimental tradition been so late to emerge in economics? In Chapter 1 we argued that a discipline becomes experimental when innovators develop techniques for conducting relevant experiments. However, development of experimental technology is only a part of the story and raises as many questions as it answers. Why were innovators able to develop new techniques in the 1960s and 70s and not before? Why did mainstream economists begin to acknowledge the relevance of laboratory experiments in the 1980s and not even later? To answer such questions we must look at the development of the economics discipline as a whole.

In this chapter we offer a brief historical account of the emergence of an experimental tradition in economics, and our own tentative explanation of its timing. We are not historians and do not try to be complete and definitive; our goals are more modest. Now that you are familiar with the techniques of experimental economics, you should understand how they arose and how they relate to other experimental traditions in the social sciences. Our historical account may provide useful perspectives. You may also find the story of some interest in its own right.

We begin with some ideas about the evolution of scientific thought, mostly drawn from Kuhn (1970) and Lakatos (1978), and apply these ideas to economic theory. The historical narrative in the next several sections is based on Smith (1991) as well as on personal conversations and correspondence with Charles Plott and several of the other people involved. We trace the development of experimental economics up to early 1980s when it found increasing acceptance into mainstream economics. After a quick geographical sketch of activity in experimental economics in the early 1990s, we discuss the divergence of the discipline