# 19

# Sample Surveys

*"Data! data! data!" he cried impatiently. "I can't make bricks without clay."*
*—Sherlock Holmes*[1]

## 1. INTRODUCTION

An investigator usually wants to generalize about a class of individuals. This class is called the *population*. For example, in forecasting the results of a presidential election in the U.S., one relevant population consists of all eligible voters. Studying the whole population is usually impractical. Only part of it can be examined, and this part is called the *sample*. Investigators will make generalizations from the part to the whole. In more technical language, they make *inferences* from the sample to the population.[2]

Usually, there are some numerical facts about the population which the investigators want to know. Such numerical facts are called *parameters*. In forecasting a presidential election in the U.S., two relevant parameters are

- the average age of all eligible voters,
- the percentage of all eligible voters who are currently registered to vote.

Ordinarily, parameters like these cannot be determined exactly, but can only be estimated from a sample. Then a major issue is accuracy. How close are the estimates going to be?

Parameters are estimated by *statistics*, or numbers which can be computed from a sample. For instance, with a sample of 10,000 Americans, an investigator could calculate the following two statistics:

- the average age of the eligible voters in the sample,
- the percentage of the eligible voters in the sample who are currently registered to vote.

Statistics are what investigators know; parameters are what they want to know.

Estimating parameters from the sample is justified when the sample represents the population. This is impossible to check just by looking at the sample. The reason: to see whether the sample is like the population in the ways that matter, investigators would have to know the facts about the population that they are trying to estimate—a vicious circle. Instead, one has to look at how the sample was chosen. Some methods tend to do badly. Others are likely to give representative samples.

The two main lessons of this chapter:

- the method of choosing the sample matters a lot;
- the best methods involve the planned introduction of chance.

Similar issues come up when assigning subjects to treatment or control in experiments: see part I.

## 2. THE *LITERARY DIGEST* POLL

In 1936, Franklin Delano Roosevelt was completing his first term of office as president of the U.S. It was an election year, and the Republican candidate was Governor Alfred Landon of Kansas. The country was struggling to recover from the Great Depression. There were still nine million unemployed: real income had dropped by one-third in the period 1929–1933 and was just beginning to turn upward. But Landon was campaigning on a program of economy in government, and Roosevelt was defensive about his deficit financing.[3]

*Landon.* The spenders must go.

*Roosevelt.* We had to balance the budget of the American people before we could balance the budget of the national government. That makes common sense, doesn't it?

The Nazis were rearming Germany, and the Civil War in Spain was moving to its hopeless climax. These issues dominated the headlines in the *New York Times*, but were ignored by both candidates.

*Landon.* We must mind our own business.

Most observers thought Roosevelt would be an easy winner. Not so the *Literary Digest* magazine, which predicted an overwhelming victory for Landon, with Roosevelt getting only 43% of the popular vote. This prediction was based on the largest number of people ever replying to a poll—about 2.4 million individuals. It was backed by the enormous prestige of the *Digest*, which had called the winner in every presidential election since 1916. However, Roosevelt won the 1936 election by a landslide—62% to 38%. (The *Digest* went bankrupt soon after.)

The magnitude of the *Digest*'s error is staggering. It is the largest ever made by a major poll. Where did it come from? The number of replies was more than big enough. In fact, George Gallup was just setting up his survey organization.[4] Using his own methods, he drew a sample of 3,000 people and predicted what the *Digest* predictions were going to be—well in advance of their publication—with an error of only one percentage point. Using another sample of about 50,000 people, he correctly forecast the Roosevelt victory, although his prediction of Roosevelt's share of the vote was off by quite a bit. Gallup forecast 56% for Roosevelt; the actual percentage was 62%, so the error was 62% − 56% = 6 percentage points. (Survey organizations use "percentage points" as the units for the difference between actual and predicted percents.) The results are summarized in table 1.

Table 1. The election of 1936.

| | *Roosevelt's percentage* |
|---|---|
| The election result | 62 |
| The *Digest* prediction of the election result | 43 |
| Gallup's prediction of the *Digest* prediction | 44 |
| Gallup's prediction of the election result | 56 |

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.
Source: George Gallup, *The Sophisticated Poll-Watcher's Guide* (1972).

To find out where the *Digest* went wrong, you have to ask how they picked their sample. A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross section of the public. A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*. The *Digest*'s procedure was to mail questionnaires to 10 million people. The names and addresses of these 10 million people came from sources like telephone books and club membership lists. That tended to screen out the poor, who were unlikely to belong to clubs or have telephones. (At the time, for example, only one household in four had a telephone.) So there was a very strong bias against the poor in the *Digest*'s sampling procedure. Prior to 1936, this bias may not have affected the predictions very much, because rich and poor voted along similar lines. But in 1936, the political split followed economic lines more closely. The poor voted overwhelmingly for Roosevelt, the rich were for Landon. One reason for the magnitude of the *Digest*'s error was selection bias.

> When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.

The *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey

organization still has to get their opinions. This is harder than it looks. If a large number of those selected for the sample do not in fact respond to the questionnaire or the interview, *non-response bias* is likely.

The non-respondents differ from the respondents in one obvious way: they did not respond. Experience shows they tend to differ in other important ways as well.[5] For example, the *Digest* made a special survey in 1936, with questionnaires mailed to every third registered voter in Chicago. About 20% responded, and of those who responded over half favored Landon. But in the election Chicago went for Roosevelt, by a two-to-one margin.

> Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.

In the main *Digest* poll, only 2.4 million people bothered to reply, out of the 10 million who got the questionnaire. These 2.4 million respondents do not even represent the 10 million people who were polled, let alone the population of all voters. The *Digest* poll was spoiled both by selection bias and non-response bias.[6]

Special surveys have been carried out to measure the difference between respondents and non-respondents. It turns out that lower-income and upper-income people tend not to respond to questionnaires, so the middle class is over-represented among respondents. For these reasons, modern survey organizations prefer to use personal interviews rather than mailed questionnaires. A typical response rate for personal interviews is 65%, compared to 25% for mailed questionnaires.[7] However, the problem of non-response bias still remains, even with personal interviews. Those who are not at home when the interviewer calls may be quite different from those who are at home, with respect to working hours, family ties, social background, and therefore with respect to attitudes. Good survey organizations keep this problem in mind, and have ingenious methods for dealing with it (section 6).

> Some samples are really bad. To find out whether a sample is any good, ask how it was chosen. Was there selection bias? non-response bias? You may not be able to answer these questions just by looking at the data.

In the 1936 election, how did Gallup predict the *Digest* predictions? He just chose 3,000 people at random from the same lists the *Digest* was going to use, and mailed them all a postcard asking how they planned to vote. He knew that a random sample was likely to be quite representative, as will be explained in the next two chapters.

## 3. THE YEAR THE POLLS ELECTED DEWEY

Thomas Dewey rose to fame as a crusading D.A. in New York City, and went on to capture the governor's mansion in Albany. In 1948 he was the Republican candidate for president, challenging the incumbent Harry Truman. Truman began political life as a protégé of Boss Pendergast in Kansas City. After being elected to the Senate, Truman became FDR's vice president, succeeding to the presidency when Roosevelt died. Truman was one of the most effective presidents of the 20th century, as well as one of the most colorful. He kept a sign on his desk, "The buck stops here." Another of his favorite aphorisms became part of America's political vocabulary: "If you can't stand the heat, stay out of the kitchen." But Truman was the underdog in 1948, for it was a troubled time. World War II had barely ended, and the uneasy half-peace of the Cold War had just begun. There was disquiet at home, and complicated involvement abroad.

Three major polls covered the election campaign: Crossley, for the Hearst newspapers; Gallup, syndicated in about 100 independent newspapers across the country; and Roper, for *Fortune* magazine. By fall, all three had declared Dewey the winner, with a lead of around 5 percentage points. Gallup's prediction was based on 50,000 interviews; and Roper's on 15,000. As the *Scranton Tribune* put it,

DEWEY AS GOOD AS ELECTED,
STATISTICS CONVINCE ROPER

The statistics didn't convince the American public. On Election Day, Truman scored an upset victory with just under 50% of the popular vote. Dewey got just over 45% (table 2).

Table 2. The election of 1948.

| | *The predictions* | | | |
|---|---|---|---|---|
| *The candidates* | *Crossley* | *Gallup* | *Roper* | *The results* |
| Truman | 45 | 44 | 38 | 50 |
| Dewey | 50 | 50 | 53 | 45 |
| Thurmond | 2 | 2 | 5 | 3 |
| Wallace | 3 | 4 | 4 | 2 |

Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

To find out what went wrong for the polls, it is necessary to find out how they chose their samples.[8] The method they all used is called *quota sampling*. With this procedure, each interviewer was assigned a fixed quota of subjects to interview. The numbers falling into certain categories (like residence, sex, age, race, and economic status) were also fixed. In other respects, the interviewers were free to select anybody they liked. For instance, a Gallup Poll interviewer in St. Louis was required to interview 13 subjects, of whom:[9]

- exactly 6 were to live in the suburbs, and 7 in the central city,
- exactly 7 were to be men, and 6 women.

Of the 7 men (and there were similar quotas for the women):

- exactly 3 were to be under forty years old, and 4 over forty,
- exactly 1 was to be black, and 6 white.

The monthly rentals to be paid by the 6 white men were specified also:

- 1 was to pay $44.01 or more;
- 3 were to pay $18.01 to $44.00;
- 2 were to pay $18.00 or less.

Remember, these are 1948 prices!

From a common-sense point of view, quota sampling looks good. It seems to guarantee that the sample will be like the voting population with respect to all the important characteristics that affect voting behavior. (Distributions of residence, sex, age, race, and rent can be estimated quite closely from Census data.) But the 1948 experience shows this procedure worked very badly. We are now going to see why.

The survey organizations want a sample which faithfully represents the nation's political opinions. However, no quotas can be set on Republican or Democratic votes. The distribution of political opinion is precisely what the survey organizations do not know and are trying to find out. The quotas for the other variables are an indirect effort to make the sample reflect the nation's politics. Fortunately or unfortunately, there are many factors which influence voting behavior besides the ones the survey organizations control for. There are rich white men in the suburbs who vote Democratic, and poor black women in the central cities who vote Republican. As a result, survey organizations may hand-pick a sample which is a perfect cross section of the nation on all the demographic variables, but find the sample voting one way while the nation goes the other. This possibility must have seemed quite theoretical—before 1948.

The next argument against quota sampling is the most important. It involves a crucial feature of the method, which is easy to miss the first time through. Within the assigned quotas, the interviewers are free to choose anybody they like. That leaves a lot of room for human choice. And human choice is always subject to bias. In 1948, the interviewers chose too many Republicans. On the whole, Republicans are wealthier and better educated than Democrats. They are more likely to own telephones, have permanent addresses, and live on nicer blocks. Within each demographic group, Republicans are marginally easier to interview. If you were an interviewer, you would probably end up with too many Republicans.

The interviewers chose too many Republicans in every presidential election from 1936 through 1948, as shown by the Gallup Poll results in table 3. Prior to 1948, the Democratic lead was so great that it swamped the Republican bias in the polls. The Democratic lead was much slimmer in 1948, and the Republican bias in quota sampling had real impact.

Table 3. The Republican bias in the Gallup Poll, 1936–1948.

| *Year* | *Gallup's prediction of Republican vote* | *Actual Republican vote* | *Error in favor of the Republicans* |
|---|---|---|---|
| 1936 | 44 | 38 | 6 |
| 1940 | 48 | 45 | 3 |
| 1944 | 48 | 46 | 2 |
| 1948 | 50 | 45 | 5 |

Note: Percentages are of the majority-party vote, except in 1948.
Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

> In quota sampling, the sample is hand-picked to resemble the population with respect to some key characteristics. The method seems reasonable, but does not work very well. The reason is unintentional bias on the part of the interviewers.

The quotas in quota sampling are sensible enough, although they do not guarantee success—far from it. But the method of filling the quotas, free choice by the interviewers, is disastrous.[10] The alternative is to use objective and impartial chance mechanisms to select the sample. That will be the topic of the next section.

## 4. USING CHANCE IN SURVEY WORK

Even in 1948, a few survey organizations used *probability methods* to draw their samples. Now, many organizations do. What is a probability method for drawing a sample? To get started, imagine carrying out a survey of 100 voters in a small town with a population of 1,000 eligible voters. Then, it is feasible to list all the eligible voters, write the name of each one on a ticket, put all 1,000 tickets in a box, and draw 100 tickets at random. Since there is no point interviewing the same person twice, the draws are made without replacement. In other words, the box is shaken to mix up the tickets. One is drawn out at random and set aside. That leaves 999 in the box. The box is shaken again, a second ticket is drawn out and set aside. The process is repeated until 100 tickets have been drawn. The people whose tickets have been drawn form the sample.

This process is called *simple random sampling*: tickets have simply been drawn at random without replacement. At each draw, every ticket in the box has an equal chance to be chosen. The interviewers have no discretion at all in whom they interview, and the procedure is impartial—everybody has the same chance to get into the sample. Consequently, the law of averages guarantees that the percentage of Democrats in the sample is likely to be close to the percentage in the population.

*Simple random sampling* means drawing at random without replacement.

What happens in a more realistic setting, when the Gallup Poll tries to predict a presidential election? A natural idea is to take a nationwide simple random sample of a few thousand eligible voters. However, this isn't as easy to do as it sounds. Drawing names at random, in the statistical sense, is hard work. It is not at all the same as choosing people haphazardly.

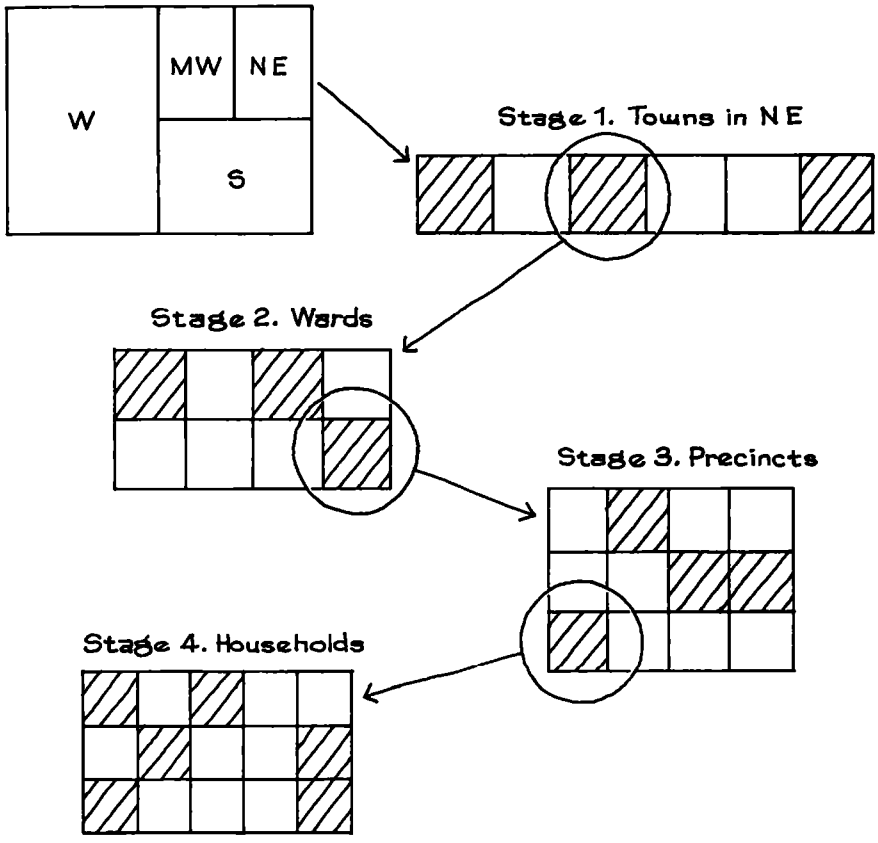
To begin drawing eligible voters at random, you would need a list of all of them—well over 200 million names. There is no such list.[11] Even if there were, drawing a few thousand names at random from 200 million is a job in itself. (Remember, on each draw every name in the box has to have an equal chance of being selected.) And even if you could draw a simple random sample, the people would be scattered all over the map. It would be prohibitively expensive to send interviewers around to find them all.

It just is not practical to take a simple random sample. Consequently, most survey organizations use a probability method called *multistage cluster sampling*. The name is complicated, and so are the details. But the idea is straightforward. It will be described in the context of the Gallup pre-election surveys during the period from 1952 through 1984; these surveys were all done using just about the same procedure. The Gallup Poll makes a separate study in each of the four geographical regions of the United States Northeast, South, Midwest, and West (figure 1). Within each region, they group together all the population centers of similar sizes. One such grouping might be all towns in the Northeast with a population between 50 and 250 thousand. Then, a random sample of these towns is selected. Interviewers are stationed in the selected towns, and no interviews are conducted in the other towns of that group. Other groupings are handled the same way. This completes the first stage of sampling.[12]

For election purposes, each town is divided up into *wards*, and the wards are subdivided into *precincts*. At the second stage of sampling, some wards are selected—at random—from each town chosen in the stage before. At the third stage, some precincts are drawn at random from each of the previously selected wards. At the fourth stage, households are drawn at random from each selected precinct.[13] Finally, some members of the selected households are interviewed. Even here, no discretion is allowed. For instance, Gallup Poll interviewers are instructed to "speak to the youngest man 18 or older at home, or if no man is at home, the oldest woman 18 or older."[14]

This design offers many of the advantages of quota sampling. For instance, it is set up so the distribution of the sample by residence is the same as the distribution for the nation. But each stage in the selection procedure uses an objective and impartial chance mechanism to select the sample units. This completely eliminates the worst feature of quota sampling: selection bias on the part of the interviewer.

Figure 1. Multistage cluster sampling.




Simple random sampling is the basic probability method. Other methods can be quite complicated. But all probability methods for sampling have two important features:

- the interviewers have no discretion at all as to whom they interview;
- there is a definite procedure for selecting the sample, and it involves the planned use of chance.

As a result, with a probability method it is possible to compute the chance that any particular individuals in the population will get into the sample.[15]

Quota sampling is not a probability method. It fails both tests. The interviewers have a lot of discretion in choosing subjects. And chance only enters in the most unplanned and haphazard way. What kinds of people does the interviewer like to approach? Who is going to be walking down a particular street at a particular time of day? No survey organization can put numbers on these kinds of chances.

## 5. HOW WELL DO PROBABILITY METHODS WORK?

Since 1948, the Gallup Poll and many other major polls have used probability methods to choose their samples. The Gallup Poll record in post-1948 presidential elections is shown in table 4. There are three points to notice. (i) The sample size has gone down sharply. The Gallup Poll used a sample of size about 50,000 in 1948; they now use samples less than a tenth of that size. (ii) There is no longer any consistent trend favoring either Republicans or Democrats. (iii) The accuracy has gone up appreciably.

From 1936 to 1948, the errors were around 5%. Since then, they are quite a bit smaller. (In 1992, the error went back up to 6%; the reason will be discussed on p. 346.) Using probability methods to select the sample, the Gallup Poll has been able to predict the elections with startling accuracy, sampling less than 5 persons in 100,000—which proves the value of probability methods in sampling.

Table 4. The Gallup Poll record in presidential elections after 1948.

| *Year* | *Sample size* | *Winning candidate* | *Gallup Poll prediction* | *Election result* | *Error* |
|---|---|---|---|---|---|
| 1952 | 5,385 | Eisenhower | 51% | 55.1% | 4.1% |
| 1956 | 8,144 | Eisenhower | 59.5% | 57.4% | 2.1% |
| 1960 | 8,015 | Kennedy | 51% | 49.7% | 1.3% |
| 1964 | 6,625 | Johnson | 64% | 61.1% | 2.9% |
| 1968 | 4,414 | Nixon | 43% | 43.4% | 0.4 of 1% |
| 1972 | 3,689 | Nixon | 62% | 60.7% | 1.3% |
| 1976 | 3,439 | Carter | 48% | 50.1% | 2.1% |
| 1980 | 3,500 | Reagan | 47% | 50.7% | 3.7% |
| 1984 | 3,456 | Reagan | 59% | 58.8% | 0.2 of 1% |
| 1988 | 4,089 | Bush | 56% | 53.4% | 2.6% |
| 1992 | 2,019 | Clinton | 49% | 43.0% | 6.0% |
| 1996 | 2,895 | Clinton | 52% | 49.2% | 2.8% |
| 2000 | 3,571 | Bush | 48% | 47.9% | 0.1 of 1% |
| 2004 | 2,014 | Bush | 49% | 50.6% | 1.6% |

Note: The percentages are of the popular vote. The error is the absolute difference "predicted – actual."
Source: The Gallup Poll (American Institute of Public Opinion) for predictions; *Statistical Abstract*, 2006, Table 384 for actuals.

Why do probability methods work so well? At first, it may seem that judgment is needed to choose the sample. For instance, quota sampling guarantees that the percentage of men in the sample will be equal to the percentage of men in the population. With probability sampling, we can only say that the percentage of men in the sample is likely to be close to the percentage in the population: certainty is reduced to likelihood. But judgment and choice usually show bias, while chance is impartial. That is why probability methods work better than judgment.

> To minimize bias, an impartial and objective probability method should be used to choose the sample.

## 6. A CLOSER LOOK AT THE GALLUP POLL

Some degree of bias is almost inevitable even when probability methods are used to select the sample, due to the many practical difficulties that survey organizations must overcome. The discussion here is organized around the questionnaire used by the Gallup Poll in the presidential election of 1984. See figures 2 and 3.

Figure 2. The Gallup Poll ballot, 1984. The interviewers use secret ballots, to minimize the number of undecided respondents.

| DEMOCRATIC | REPUBLICAN |
|---|---|
| ☐ MONDALE<br>& Ferraro | ☐ REAGAN<br>& Bush |

" I'D SAY I'M ABOUT FORTY-TWO PERCENT FOR REAGAN, THIRTY-NINE PERCENT FOR MONDALE, AND NINETEEN PERCENT UNDECIDED."

*The nonvoters.* In a typical presidential election, between one-third and one-half of the eligible voters fail to vote. The job of the Gallup Poll is to predict what the voters will do; the non-voters are irrelevant and should be screened out of the sample as far as possible. That is not so easy. There is a stigma to non-voting, and many respondents say they will vote even if they know better. The problem of screening out non-voters is handled by questions 1–6, and some later questions too: this is a hot topic for pollsters. Question 3, for instance, asks where the respondent would go to vote (figure 3, p. 345). Respondents who know the answer are more likely to vote. Question 13 (p. 347) asks whether the respondent voted in the last election, and is phrased to make a negative answer easy to give—compensating for the stigma attached to non-voting. Respondents who voted last time are more likely to vote this time.

This battery of questions is used to decide whether the respondent is likely to vote; the election forecasts are based only on that part of the sample judged likely to vote. It is a matter of record who actually votes in each election. Post-election studies by the Gallup organization show that their judgments as to who will vote are reasonably accurate. The studies also show that screening out likely non-voters increases the accuracy of the election forecasts, because preferences of likely voters differ from preferences of likely non-voters.[16]

*The undecided.* Some percentage of the subjects being interviewed are undecided how they will vote. Question 7, which asks for the preferences, is designed to keep the percentage as small as possible. To begin with, it asks how the respondent would vote the day of the interview, rather than Election Day. Subjects who cannot decide are asked to indicate "the candidates toward whom you lean as of today." A final device is the paper ballot (figure 2, p. 343). Instead of naming their preferences out loud, the respondents just mark the ballot and drop it into a box carried by the interviewer.[17]

These techniques have been found to minimize the percentage of undecided. But there are still some left, and if they are thought likely to vote, the Gallup Poll has to guess how. Some information about political attitudes is available from questions 12–14 (p. 347). This information might be used to predict how the undecided respondents are going to vote, but it is difficult to say how well the predictions work.

*Response bias.* The answers given by respondents are influenced to some extent by the phrasing of the questions, and even the tone or attitude of the interviewer. This kind of distortion is called *response bias.* There was a striking example in the 1948 election survey: changing the order of the candidates' names was found to change the response by 5%, the advantage being with the candidate who was named first. To control response bias, all interviewers use the same questionnaire, and the interview procedure is standardized as far as possible. The ballot technique was found to reduce the effect of the political attitudes of the interviewer on the responses of the subjects.

*Non-response bias.* Even with personal interviews, many subjects are missed. Since they tend to be different from the subjects available for the interview, a non-response bias is created. To some extent, this bias can be adjusted out, by giving more weight to the subjects who were available but hard to get. This information is obtained by question 20 (p. 347), which asks whether the subject was at home on the previous days. This is done quite subtly, as you can tell by reading the question.

Figure 3. The Gallup Poll questionnaire for the 1984 election. Courtesy of the Gallup Poll News Service.

SURVEY: A1813

DATE: October 25, 1984

No publication, reproduction, dissemination or other use of this questionnaire or any replies thereto, written or oral, is authorized by The Gallup Organization, Inc. Violators of this notice will be prosecuted to the fullest extent of the law.

U.S. & Worldwide

Sponsored by leading Newspapers, Corporations and Agencies.

# The Gallup Survey

Copyright 1977 The Gallup Organization, Inc. Princeton, New Jersey 08540

SUGGESTED INTRODUCTION: I'm taking a GALLUP SURVEY. I'd like YOUR opinion on some topics of interest.

Time started: _____

Time ended: _____

Length: _____

1. How much thought have you given to the coming November elections—quite a lot, or only a little?
   - 1 ☐ Quite a lot
   - 2 ☐ Some — (volunteered)
   - 3 ☐ Little
   - y ☐ None

2. Have you ever voted in this precinct or district?
   - 1 ☐ Yes
   - 2 ☐ No
   - y ☐ Don't know

3. Where do people who live in this neighborhood go to vote?
   - 1 ☐ Specify: _____
   - y ☐ Don't know

4a. Are you NOW registered so that you can vote in the election this November?
   - 1 ☐ Yes — (GO TO Q. 5)
   - 2 ☐ No
   - 3 ☐ Don't have to register — (GO TO Q. 5)
   - y ☐ Don't know

4b. Do you plan to register so that you can vote in the November election?
   - 1 ☐ Yes
   - 2 ☐ No
   - 3 ☐ Other: _____

5. Generally speaking, how much interest would you say you have in politics — a great deal, a fair amount, only a little, or no interest at all?
   - 1 ☐ Great deal
   - 2 ☐ Fair amount
   - 3 ☐ Little
   - y ☐ None

6. How often would you say you vote — always, nearly always, part of the time, or seldom?
   - 1 ☐ Always
   - 2 ☐ Nearly always
   - 3 ☐ Part of the time
   - 4 ☐ Seldom
   - 5 ☐ Other: _____
   - y ☐ Never vote

7. Suppose you were voting TODAY for president and vice president of the United States. Here is a Gallup Poll secret ballot listing the candidates for these offices. (TEAR OFF ATTACHED BALLOT AND HAND TO RESPONDENT.) Will you please MARK that secret ballot for the candidates you favor today — and then drop the folded ballot into the box.

   INTERVIEWER: IF RESPONDENT HANDS BACK BALLOT AND SAYS HE HASN'T MADE UP HIS MIND OR REFUSES TO MARK IT SAY:

   Well, would you please mark the ballot for the candidates toward whom you lean as of today?

   IF RESPONDENT STILL CAN'T DECIDE OR REFUSES TO MARK THE BALLOT, PLEASE WRITE THAT ON THE BALLOT AND BE SURE TO DROP IT IN THE BOX.

8. Right now, how strongly do you feel about your choice—very strongly, fairly strongly, or not strongly at all?
   - 1 ☐ Very strongly
   - 2 ☐ Fairly strongly
   - 3 ☐ Not strongly at all
   - 4 ☐ Didn't make choice
   - y ☐ Don't know

9a. Do you, yourself, plan to vote in the election this November, or not?
   - 1 ☐ Yes
   - 2 ☐ No
   - y ☐ Don't know
   - (GO TO Q. 10a)

9b. How certain are you that you will vote—ABSOLUTELY certain, FAIRLY certain, or NOT certain?
   - 1 ☐ Absolutely
   - 2 ☐ Fairly
   - y ☐ Not certain

10a. If the elections for Congress were being held TODAY, which party would you like to see win in this Congressional district, the Democratic Party or the Republican Party?
   - 1 ☐ Democratic
   - 2 ☐ Republican
   - 3 ☐ Other
   - (GO TO Q. 11)
   - y ☐ Undecided, refused

10b. As of today, do you lean more to the Democratic Party or more to the Republican Party?
   - 1 ☐ Democratic
   - 2 ☐ Republican
   - 3 ☐ Other:
   - 4 ☐ Undecided
   - y ☐ Refused

11. Here is a picture of a ladder. (HAND RESPONDENT CARD 1.) Suppose we say the top of the ladder (POINT) marked 10 represents a person who definitely will vote in the election this November, and the bottom of the ladder (POINT) marked zero represents a person who definitely will not vote in the election. How far up or down the ladder would you place yourself? (INTERVIEWER: CIRCLE NUMBER.)

   10 9 8 7 6 5 4 3 2 1 0

   y ☐ Don't know

*Check data.* The Gallup sample usually includes proportionately too many people with higher education. In a detailed analysis, less weight is put on the responses of those subjects (question 16). Other demographic data can be used in a similar way. This weighting technique is called "ratio estimation." Do not confuse ratio estimation with quota sampling. Ratio estimation is an objective, arithmetic technique applied to the sample after it is chosen, to compensate for various small biases in the sampling procedure. Quota sampling is a method for choosing the sample. It has a large, subjective component—when the interviewer chooses the subjects—and introduces large biases.

*Interviewer control.* In large-scale survey work, there is always the problem of making sure interviewers follow instructions. Some redundancy is built into the questionnaire, so the answers can be checked for consistency: inconsistencies suggest the interviewer may not be doing the job properly. A small percentage of the subjects are reinterviewed by administrative staff, as a further check on the quality of the work.

*Talk is cheap.* It is a little risky to predict what people will do on Election Day from what they tell the interviewer they are going to do. People may be unwilling to reveal their true preferences. Even if they do, they may change their minds later. Words and deeds are often different.

*The 1992 election.* In 1992, there was a fairly large percentage of undecided respondents, and Gallup allocated all of them to Clinton. That turned out to be a bad idea. Many of the undecided seem in the end to have voted for Perot, explaining Gallup's large error for the 1992 election (table 4, p. 342). Predicted and actual votes for Clinton, Bush, and Perot are shown below.

| | *Clinton* | *Bush* | *Perot* |
|---|---|---|---|
| Gallup | 49% | 37% | 14% |
| Actual | 43.0% | 37.4% | 18.9% |

## 7. TELEPHONE SURVEYS

Many surveys are now conducted by telephone. The savings in costs are dramatic, and—if the work is up to scratch—the results are good. The Gallup Poll changed over to the telephone in 1988, with 200 interviewers covering the whole country in a few days, from offices in Atlanta, Austin, Lincoln, Minneapolis, and Omaha.

How do they pick the sample? In 1988, the Gallup Poll used a multistage cluster sample based on area codes, "exchanges," and "banks:"

| *Area code* | *Exchange* | *Bank* | *Digits* |
|---|---|---|---|
| 415 | 767 | 26 | 76 |

In 1992, they switched to a simpler design. There are 4 time zones in the U.S. The Gallup Poll divided each zone into 3 types of areas, according to population density (heavy, medium, light). That gives $4 \times 3 = 12$ strata. For example, one stratum consisted of heavily populated areas in the Eastern time zone; another consisted of lightly populated areas on Pacific time. Within each stratum, the Gallup Poll just drew a simple random sample of telephone numbers, using the computer to

Figure 3. The Gallup Poll questionnaire for the 1984 election, continued. Courtesy of the Gallup Poll News Service.

NOW, HERE ARE A FEW QUESTIONS SO THAT MY OFFICE CAN KEEP TRACK OF THE CROSS-SECTION OF PEOPLE I'VE TALKED TO:

12. In politics, as of TODAY, do you consider yourself a Republican, Democrat, or Independent?
    1 ☐ Republican
    2 ☐ Democrat
    3 ☐ Independent
    4 ☐ Other: _____

13. In the election in November 1980 — when Carter ran against Reagan and Anderson — did things come up which kept you from voting, or did you happen to vote? For whom?
    1 ☐ Carter
    2 ☐ Reagan
    3 ☐ Anderson
    4 ☐ Other
    5 ☐ Voted, don't remember for whom
    6 ☐ No, didn't vote
    y ☐ Don't remember if voted

14. Are you, or is your (husband/wife) a member of a labor union?
    1 ☐ Yes, respondent is
    2 ☐ Yes, spouse is
    3 ☐ Yes, both are
    y ☐ No, neither is

15. (HAND RESPONDENT CARD 2) Please tell me which of the categories on this card MOST NEARLY describes the kind of work the chief wage earner in your immediate family does. Just call off the number, please. (INTERVIEWER: IF THE CHIEF WAGE EARNER IS UNEMPLOYED, ASK WHAT TYPE OF WORK HE/SHE WOULD DO IF EMPLOYED.)
    1 ☐ 11 ☐
    2 ☐ 12 ☐
    3 ☐ 13 ☐
    4 ☐ 14 ☐
    5 ☐ 15 ☐
    6 ☐ 16 ☐ Other: _____
    7 ☐ 17 ☐ Can't say
    8 ☐
    9 ☐
    10 ☐

16. What was the last grade or class you COMPLETED in school?
    1 ☐ None or Grades 1–4
    2 ☐ Grades 5, 6, 7
    3 ☐ Grade 8
    4 ☐ High school, incomplete (Grades 9–11)
    5 ☐ High school, graduated (Grade 12)
    6 ☐ Technical, trade, or business
    7 ☐ College, university, incomplete
    8 ☐ College, university, graduated

17. What is your religious preference — Protestant, Roman Catholic, Jewish, or an Orthodox church such as the Greek or Russian Orthodox Church?
    1 ☐ Protestant
    2 ☐ Roman Catholic
    3 ☐ Jewish
    4 ☐ Orthodox Church
    5 ☐ Other: _____
    y ☐ None

18. How many persons 18 years and over are there now living in this household, including yourself? Include lodgers, servants, or other employees living in the household. (CIRCLE NUMBER)
    1 2 3 4 5 6 7 8 9 or more

19. (HAND RESPONDENT CARD 3) From what nationality group or groups are you mainly descended? Just call off the number please.
    1 ☐ 11 ☐
    2 ☐ 12 ☐
    3 ☐ 13 ☐
    4 ☐ 14 ☐ Don't know (VOLUNTEERED) or refused
    5 ☐
    6 ☐
    7 ☐
    8 ☐
    9 ☐
    10 ☐

20a. We are interested in finding out how often people are at home to watch TV or listen to the radio. Would you mind telling me whether or not you happened to be at home yesterday (last night/last Saturday) at this particular time?

(INTERVIEWER: SEE INTERVIEWER'S BULLETIN FOR HANDLING THIS QUESTION.)
    1 ☐ Yes, at home
    2 ☐ No, not at home

20b. How about the day (night/Saturday) before at this time?
    1 ☐ Yes, at home
    2 ☐ No, not at home

20c. And how about the day (night/Saturday) before at this time? That was _____.
    1 ☐ Yes, at home
    2 ☐ No, not at home

21. And what is your age?
    RECORD AGE: _____

22. CHECK WHETHER:
    1 ☐ White man
    2 ☐ White woman
    3 ☐ Black man
    4 ☐ Black woman
    5 ☐ Other man (SPECIFY) _____
    6 ☐ Other woman (SPECIFY) _____

So that my office can check my work in this interview if it wants to, may I have your name, address, and telephone number please?

NAME: _____

ADDRESS: _____
House No. or RFD Route, St. or Rd., Apt. No.)

CITY: _____ STATE: _____ ZIP _____

TELEPHONE: Area Code _____ Phone No. _____ y ☐ No tel.

PLACE INTERVIEWER BADGE NUMBER HERE

I hereby attest that this is a true and honest interview.

_____
(Interviewer's Signature)

Date of interview: _____

Time interview ended: _____

exclude businesses by checking the yellow pages. Choosing telephone numbers at random is called RDD, for *random digit dialing*.[18]

People who do not have phones must be different from the rest of us, and that does cause a bias in telephone surveys. The effect is small, because these days nearly everybody has a phone. On the other hand, about one-third of residential telephones are unlisted. Rich people and poor people are more likely to have unlisted numbers, so the telephone book tilts toward the middle class. Sampling from directories would create a real bias, but random digit dialing gets around this difficulty. In 2005, survey organizations were just beginning to work on the questions raised by cell phones. What about dropped calls? Who pays for air time? What to do with people who have land lines and cell phones?

Non-respondents create problems, as usual. So the Gallup Poll does most of its interviewing on evenings and weekends, when people are more likely to be at home. If there is no answer, the interviewer will call back up to 3 times.[19] (Some designs have up to 15 call-backs; that is better, but more expensive.) For many purposes, results are comparable to those from face-to-face interviews, and the cost is about one-third as much. That is why survey organizations are using the telephone.

## 8. CHANCE ERROR AND BIAS

The previous sections indicated the practical difficulties faced by real survey organizations. People are not at home, or they do not reveal their true preferences, or they change their minds. However, even if all these difficulties are assumed away, the sample is still likely to be off—due to chance error.

To focus the issue, imagine a box with a very large number of tickets, some marked 1 and the others marked 0. That is the population. A survey organization is hired to estimate the percentage of 1's in the box. That is the parameter. The organization draws 1,000 tickets at random without replacement. That is the sample. There is no problem about response—the tickets are all there in the box. Drawing them at random eliminates selection bias. And the tickets do not change back and forth between 0 and 1. As a result, the percentage of 1's in the sample is going to be a good estimate for the percentage of 1's in the box. But the estimate is still likely to be a bit off, because the sample is only part of the population. Since the sample is chosen at random, the amount off is governed by chance:

$$\text{percentage of 1's in sample} = \text{percentage of 1's in box} + \text{chance error.}$$

Now there are some questions to ask about chance errors—

- How big are they likely to be?
- How do they depend on the size of the sample? the size of the population?
- How big does the sample have to be in order to keep the chance errors under control?

These questions will be answered in the next two chapters.

In more complicated situations, the equation has to take bias into account:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Chance error is often called "sampling error:" the "error" comes from the fact that the sample is only part of the whole. Similarly, bias is called "non-sampling error"—the error from other sources, like non-response. Bias is often a more serious problem than chance error, but methods for assessing bias are not well developed. Usually, "bias" means prejudice. However, statistics is a dry subject. For a statistician, bias just means any kind of systematic error in an estimate. "Non-sampling error" is a more neutral term, and may be better for that reason.

## Exercise Set A

1. A survey is carried out at a university to estimate the percentage of undergraduates living at home during the current term. What is the population? the parameter?

2. The registrar keeps an alphabetical list of all undergraduates, with their current addresses. Suppose there are 10,000 undergraduates in the current term. Someone proposes to choose a number at random from 1 to 100, count that far down the list, taking that name and every 100th name after it for the sample.

   (a) Is this a probability method?
   (b) Is it the same as simple random sampling?
   (c) Is there selection bias in this method of drawing a sample?

3. The monthly Gallup Poll opinion survey is based on a sample of about 1,500 persons, "scientifically chosen as a representative cross section of the American public." The Gallup Poll thinks the sample is representative mainly because—

   (i) it resembles the population with respect to such characteristics as race, sex, age, income, and education

   or

   (ii) it was chosen using a probability method.

4. In the Netherlands, all men take a military pre-induction exam at age 18. The exam includes an intelligence test known as "Raven's progressive matrices," and includes questions about demographic variables like family size. A study was done in 1968, relating the test scores of 18-year-old men to the number of their brothers and sisters.[20] The records of all the exams taken in 1968 were used.

   (a) What is the population? the sample?
   (b) Is there any sampling error? Explain briefly.

5. Polls often conduct pre-election surveys by telephone. Could this bias the results? How? What if the sample is drawn from the telephone book?

6. About 1930, a survey was conducted in New York on the attitude of former black slaves towards their owners and conditions of servitude.[21] Some of the interviewers were black, some white. Would you expect the two groups of interviewers to get similar results? Give your reasons.

7. One study on slavery estimated that "11.9% of slaves were skilled craftsmen." This estimate turns out to be based on the records of thirty plantations in Plaquemines Parish, Louisiana.[22] Is it trustworthy? Explain briefly.

8. In one study, the Educational Testing Service needed a representative sample of college students.[23] To draw the sample, they first divided up the population of all colleges and universities into relatively homogeneous groups. (One group consisted of all public universities with 25,000 or more students; another group consisted of all private four-year colleges with 1,000 or fewer students; and so on.) Then they used their judgment to choose one representative school from each group. That created a sample of schools. Each school in the sample was then asked to pick a sample of students. Was this a good way to get a representative sample of students? Answer yes or no, and explain briefly.

9. A study was done on the prevalence of chest diseases in a Welsh coal mining town; 600 volunteers had chest X-rays done.[24] At the time, the two main chest diseases in the town were pneumoconiosis (scarring of the lung tissue due to inhalation of dust) and tuberculosis. The data were analyzed by the order in which the volunteers presented themselves. The percentage with tuberculosis among the first 200 subjects to appear for the examination was probably _____ the percentage among the last 200. Fill in the blank, using one of the phrases

   (i) about the same as (ii) quite a bit different from

   Explain your reasoning.

10. Television advertising sales are strongly influenced by the Nielsen ratings. In its annual report, the Nielsen organization does not describe how it takes samples. The report does say:[25]

    > Nielsen, today as in the past, is dedicated to using the newest, most reliable, and thoroughly tested research technologies. This is a commitment to those we serve through the television, cable, and advertising communities....
    >
    > The Nielsen data in this booklet are estimates of the audiences and other characteristics of television usage as derived from Nielsen Television Index and Nielsen Station Index measurements. The use of mathematical terms herein should not be regarded as a representation by Nielsen that such measurements are exact to precise mathematical values....

    Comment briefly.

11. The *San Francisco Examiner* ran a story headlined—

    > 3 IN 10 BIOLOGY TEACHERS BACK BIBLICAL CREATIONISM
    >
    > *Arlington, Texas.* Thirty percent of high school biology teachers polled believe in the biblical creation and 19 percent incorrectly think that humans and dinosaurs lived at the same time, according to a nationwide survey published Saturday.
    >
    > "We're doing something very, very, very wrong in biology education," said Dana Dunn, one of two sociologists at the University of Texas, Arlington.
    >
    > Dunn and Raymond Eve sent questionnaires to 20,000 high school biology teachers selected at random from a list provided by the National Science Teachers Association and received 200 responses....

    The newspaper got it wrong. Dunn and Eve did not send out 20,000 question-

naires: they chose 400 teachers at random from the National Science Teachers association list, sent questionnaires to these 400 people, and received 200 replies.[26] Why do these corrections matter?

12. In any survey, a fair number of people who are in the original sample cannot be contacted by the survey organization, or are contacted but refuse to answer questions. A high non-response rate is a serious problem for survey organizations. True or false, and explain: this problem is serious because the investigators have to spend more time and money getting additional people to bring the sample back up to its planned size.

*The answers to these exercises are on pp. A78–79.*

## 9. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. A survey organization is planning to do an opinion survey of 2,500 people of voting age in the U.S. True or false, and explain: the organization will choose people to interview by taking a simple random sample.

2. Two surveys are conducted to measure the effect of an advertising campaign for a certain brand of detergent.[27] In the first survey, interviewers ask housewives whether they use that brand of detergent. In the second, the interviewers ask to see what detergent is being used. Would you expect the two surveys to reach similar conclusions? Give your reasons.

3. One study on slavery estimated that a slave had only a 2% chance of being sold into the interstate trade each year. This estimate turns out to be based on auction records in Anne Arundel County, Maryland.[28] Is it trustworthy? Explain briefly.

4. In one study, it was necessary to draw a representative sample of Japanese-Americans resident in San Francisco.[29] The procedure was as follows. After consultation with representative figures in the Japanese community, the four most representative blocks in the Japanese area of the city were chosen. All persons resident in those four blocks were taken for the sample. However, a comparison with Census data shows that the sample did not include a high-enough proportion of Japanese with college degrees. How can this be explained?

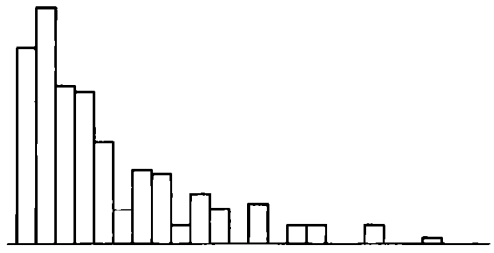
5. (Hypothetical.) A survey is carried out by the finance department to determine the distribution of household size in a certain city. They draw a simple random sample of 1,000 households. After several visits, the interviewers find people at home in only 653 of the sample households. Rather than face such a high non-response rate, the department draws a second batch of households, and uses the first 347 completed interviews in the second batch to bring the sample up to its planned strength of 1,000 households. The department counts 3,087 people in these 1,000 households, and estimates the average household size in the city to be about 3.1 persons. Is this estimate likely to be too low, too high, or about right? Why?

6. "Ecstasy" was a popular drug in the 1990s. It produced a sense of euphoria derisively called the "yuppie high." One investigator made a careful sample survey to estimate the prevalence of drug use at Stanford University. Two assistants were stationed on the main campus plaza and instructed to interview all students who passed through at specified times. As it turned out, 39% of 369 students interviewed said they had used Ecstasy at least once.[30] Does the investigator's procedure give a probability sample of Stanford students? Answer yes or no, and explain.

7. A coin is tossed 1,000 times. There are two options:
   (i) To win $1 if the number of heads is between 490 and 510.
   (ii) To win $1 if the percentage of heads is between 48% and 52%.

   Which option is better? Or are they the same? Explain.

8. Can you tell whether the figure below is a probability histogram or a histogram for data? If so, which is it and why? If you can't tell, why not?



9. One hospital has 218 live births during the month of January.[31] Another has 536. Which is likelier to have 55% or more male births? Or is it equally likely? Explain. (There is about a 52% chance for a live-born infant to be male.)

10. A coin will be tossed 100 times. You get to pick 11 numbers. If the number of heads turns out to equal one of your 11 numbers, you win a dollar. Which 11 numbers should you pick, and what is your chance (approximately) of winning? Explain.

11. A sorcerer has hidden a Porsche in one of an infinite row of boxes

The sorcerer will let you drive away with the car if you can find it. But you are only allowed to look in 11 boxes. He agrees to give you a hint, by tossing a coin 100 times and counting the number of heads. He will not tell you this number, or the number of the box in which he hid the car. But he will tell you the sum of the two numbers.

(a) If the sum is 65, which 11 boxes would you look in?
(b) As in (a), except replace 65 by 95.
(c) What is the general rule?
(d) Following this rule, how likely are you to get the Porsche?

12. The *San Francisco Chronicle* reported on a survey of top high-school students in the U.S. According to the survey,

> Cheating is pervasive. Nearly 80 percent admitted some dishonesty, such as copying someone's homework or cheating on an exam. The survey was sent last spring to 5,000 of the nearly 700,000 high achievers included in the 1993 edition of *Who's Who Among American High School Students*. The results were based on the 1,957 completed surveys that were returned. "The survey does not pretend to be representative of all teenagers," said *Who's Who* spokesman Andrew Weinstein. "Students are listed in *Who's Who* if they are nominated by their teachers or guidance counselors. Ninety-eight percent of them go on to college."

(a) Why isn't the survey "representative of all teenagers"?
(b) Is the survey representative "of the nearly 700,000 high achievers included in the 1993 edition of *Who's Who Among American High School Students*"? Answer yes or no, and explain briefly.

## 10. SUMMARY

1. A *sample* is part of a *population*.

2. A *parameter* is a numerical fact about a population. Usually a parameter cannot be determined exactly, but can only be estimated.

3. A *statistic* can be computed from a sample, and used to estimate a parameter. A statistic is what the investigator knows. A parameter is what the investigator wants to know.

4. When estimating a parameter, one major issue is accuracy: how close is the estimate going to be?

5. Some methods for choosing samples are likely to produce accurate estimates. Others are spoiled by *selection bias* or *non-response bias*. When thinking about a sample survey, ask yourself:

- What is the population? the parameter?
- How was the sample chosen?
- What was the response rate?

6. Large samples offer no protection against bias.

7. In *quota sampling*, the sample is hand picked by the interviewers to resemble the population in some key ways. This method seems logical, but often

gives bad results. The reason: unintentional bias on the part of the interviewers, when they choose subjects to interview.

8. *Probability methods* for sampling use an objective chance process to pick the sample, and leave no discretion to the interviewer. The hallmark of a probability method: the investigator can compute the chance that any particular individuals in the population will be selected for the sample. Probability methods guard against bias, because blind chance is impartial.

9. One probability method is *simple random sampling*. This means drawing subjects at random without replacement.

10. Even when using probability methods, bias may come in. Then the estimate differs from the parameter, due to bias and chance error:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Chance error is also called "sampling error," and bias is "non-sampling error."

# 20

# Chance Errors in Sampling

*To all the ladies present and some of those absent.*
—THE TOAST USUALLY PROPOSED BY JERZY NEYMAN

## 1. INTRODUCTION

Sample surveys involve chance error. This chapter will explain how to find the likely size of the chance error in a percentage, for simple random samples from a population whose composition is known. That mainly depends on the size of the sample, not the size of the population. First, an example. A health study is based on a representative cross section of 6,672 Americans age 18 to 79. A sociologist now wishes to interview these people. She does not have the resources to do them all, in fact she only has enough money to sample 100 of them. To avoid bias, she is going to draw the sample at random. In the imaginary dialogue which follows, she is discussing the problem with her statistician.[1]

*Soc.* I guess I have to write all the 6,672 names on separate tickets, put them in a box, and draw out 100 tickets at random. It sounds like a lot of work.

*Stat.* We have the files on the computer, code-numbered from 1 to 6,672. So you could just draw 100 numbers at random in that range. Your sample would be the people with those code numbers.

*Soc.* Yes, but then I still have to write the numbers from 1 to 6,672 on the tickets. You haven't saved me much time.

*Stat.* That isn't what I had in mind. With a large box, it's hard to mix the tickets properly. If you don't, most of the draws probably come from the tickets you put in last. That could be a serious bias.

*Soc.* What do you suggest?

*Stat.* The computer has a random number generator. It picks a number at random from 1 to 6,672. The person with that code number goes into the sample. Then it picks a second code number at random, different from the first. That's the second person to go into the sample. The computer keeps going until it gets 100 people. Instead of trying to mix the tickets yourself, let the random numbers do the mixing. Besides, the computer saves all that writing.

*Soc.* OK. But if we use the computer, will my sample be representative?

*Stat.* What do you have in mind?

*Soc.* Well, there were 3,091 men and 3,581 women in the original survey: 46% were men. I want my sample to have 46% men. Besides that, I want them to have the right age distribution. Then there's income and education to think about. Of course, what I really want is a group whose attitudes to health care are typical.

*Stat.* Let's not get into attitudes right now. First things first. I drew a sample to show you. Look at table 1. The first person chosen by the computer was female, so was the second. But the third was male. And so on. Altogether, you got 51 men. That's pretty close.

Table 1. One hundred people were chosen at random and classified by sex. Fifty-one were men (M), and 49 were women (F). In the population, the percentages were 46% and 54%.

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | F | M | F | M | M | F | M | M | M | M | F | M | M | M | M | F | M | F | F |
| F | M | M | F | M | F | F | M | F | F | M | M | F | F | F | M | F | M | F | M |
| F | M | F | F | M | M | F | M | M | F | M | F | M | F | M | M | M | F | F | F |
| F | M | M | M | F | M | F | M | M | F | M | M | M | M | F | F | F | M | F | M |
| F | M | F | M | M | M | F | F | F | F | M | M | F | M | M | F | F | F | F | F |

*Soc.* But there should only be 46 men. There must be something wrong with the computer.

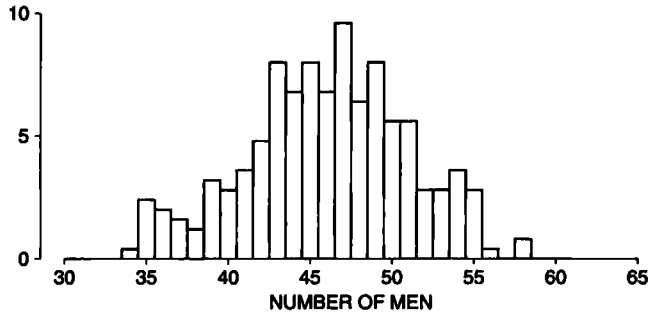
*Stat.* No, not really. Remember, the people in the sample are drawn at random. Just by the luck of the draw, you could get too many men—or too few. I had the computer take a lot of samples for you, 250 in all (table 2). The number of men ranged from a low of 34 to a high of 58. Only 17 samples out of the lot had exactly 46 men. There's a histogram (figure 1).

*Soc.* What stops the numbers from being 46?

*Stat.* Chance variability. Remember the Kerrich experiment I told you about the other day?

*Soc.* Yes, but that was about coin tossing, not sampling.

Figure 1. Histogram for the number of men in samples of size 100.




*Stat.* Well, there isn't much difference between coin tossing and sampling. Each time you toss the coin, you either get a head or a tail, and the number of heads either goes up by one or stays the same. The chances are 50–50 each time. It's the same with sampling. Each time the computer chooses a person for the sample, it either gets a man or a woman, so the number of men either goes up by one or stays the same. The chances are just about 46 to 54 each time—taking 100 tickets out of the box can't change the proportions in the box very much.

*Soc.* What's the point?

*Stat.* The chance variability in sampling is just like the chance variability in coin tossing.

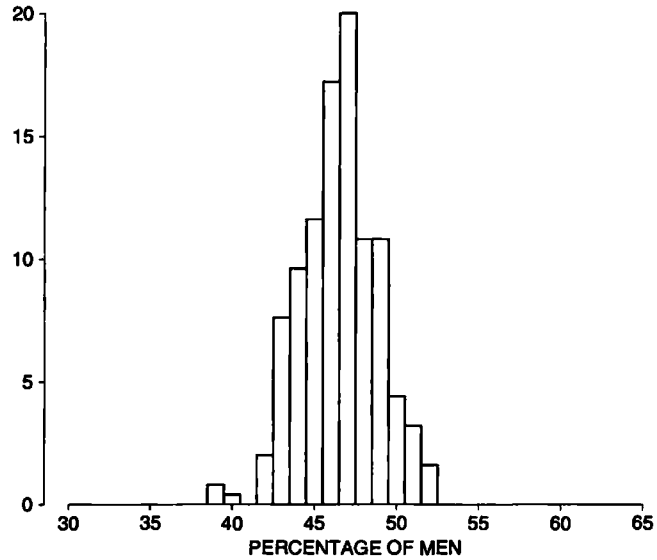
*Soc.* Hmmm. What happens if we increase the size of the sample? Won't it come out more like the population?

*Stat.* Right. For instance, suppose we increase the sample size by a factor of four, to 400. I got the computer to draw another 250 samples, this time with 400 people in each sample. With some of these samples, the percentage of men is below 46%, with others it is above. The low is 39%, the high is 54%.

Table 2. Two hundred fifty random samples were drawn from the respondents to a health study, of whom 46% were men. The sample size was 100. The number of men in each sample is shown below.

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 51 | 40 | 49 | 34 | 36 | 43 | 42 | 45 | 48 | 47 | 51 | 47 | 50 | 54 | 39 | 42 | 47 | 43 | 46 | 46 | 51 | 43 | 53 | 43 | 51 |
| 42 | 49 | 46 | 44 | 55 | 36 | 49 | 44 | 43 | 45 | 42 | 42 | 45 | 43 | 55 | 53 | 49 | 46 | 45 | 42 | 48 | 44 | 43 | 41 | 44 |
| 47 | 54 | 54 | 39 | 39 | 52 | 43 | 36 | 39 | 43 | 43 | 46 | 47 | 44 | 55 | 50 | 53 | 55 | 45 | 43 | 47 | 40 | 47 | 40 | 51 |
| 43 | 56 | 40 | 40 | 49 | 47 | 45 | 49 | 41 | 43 | 45 | 54 | 49 | 50 | 44 | 46 | 48 | 52 | 45 | 47 | 50 | 53 | 46 | 44 | 47 |
| 47 | 46 | 54 | 42 | 44 | 47 | 47 | 36 | 52 | 50 | 51 | 48 | 46 | 45 | 54 | 48 | 46 | 41 | 49 | 37 | 49 | 45 | 50 | 43 | 54 |
| 39 | 55 | 38 | 49 | 44 | 43 | 47 | 51 | 46 | 51 | 49 | 42 | 50 | 48 | 52 | 54 | 47 | 51 | 49 | 44 | 37 | 43 | 41 | 48 | 39 |
| 50 | 41 | 48 | 47 | 50 | 48 | 46 | 37 | 41 | 55 | 43 | 48 | 44 | 40 | 50 | 58 | 47 | 47 | 48 | 45 | 52 | 35 | 45 | 41 | 35 |
| 38 | 44 | 50 | 44 | 35 | 48 | 49 | 35 | 41 | 37 | 46 | 49 | 42 | 53 | 47 | 48 | 36 | 51 | 45 | 43 | 52 | 46 | 49 | 51 | 44 |
| 51 | 51 | 39 | 45 | 44 | 40 | 50 | 50 | 46 | 50 | 49 | 47 | 45 | 49 | 39 | 44 | 48 | 42 | 47 | 38 | 53 | 47 | 48 | 51 | 49 |
| 45 | 42 | 46 | 49 | 45 | 45 | 42 | 45 | 53 | 54 | 47 | 43 | 41 | 49 | 48 | 35 | 55 | 58 | 35 | 47 | 52 | 43 | 45 | 44 | 46 |

Figure 2. Histogram for the percentages of men in samples of size 400. There are 250 samples, drawn at random from the respondents to the health study.




There's a histogram (figure 2). You can compare it with the histogram for samples of size 100. Multiplying the sample size by four cuts the likely size of the chance error in the percentage by a factor of two.

*Soc.* Can you get more specific about this chance error?

*Stat.* Let me write an equation:

percentage in sample = percentage in population + chance error.

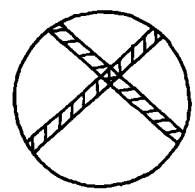
Of course, the chance error will be different from sample to sample—remember the variability in table 2.

*Soc.* So if I let you draw one sample for me, with this random-number business, can you say how big my chance error will be?

*Stat.* Not exactly, but I can tell you its likely size. If you let me make a box model, I can compute the standard error, and then....

*Soc.* Wait. There's one point I missed earlier. How can you have 250 different samples with 100 people each? I mean, $250 \times 100 = 25{,}000$, and we only started with 6,672 people.

*Stat.* The samples are all different, but they have some people in common. Look at the sketch. The inside of the circle is like the 6,672 people, and each shaded strip is like a sample:



The strips are different, but they overlap. Actually, we only scratched the surface with our sampling. The number of different samples of size 100 is over $10^{200}$. That's 1 followed by two hundred 0's. Some physicists don't even think there are that many elementary particles in the whole universe.

## 2. THE EXPECTED VALUE AND STANDARD ERROR

The sociologist of the previous section was thinking about taking a sample of size 100 from a population of 6,672 subjects in a health study. She knew that the percentage of men in the sample would be somewhere around the percentage of men in the population.

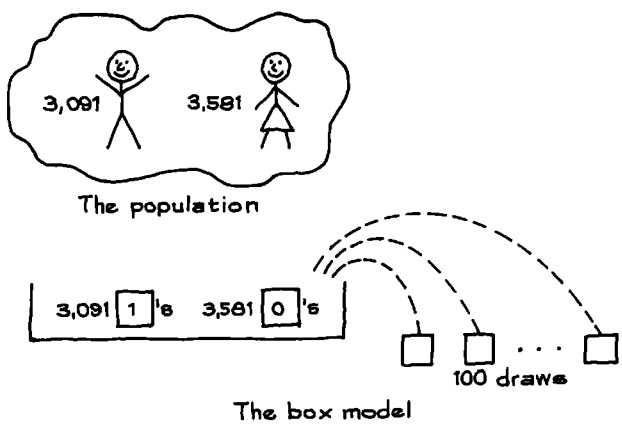
> With a simple random sample, the expected value for the sample percentage equals the population percentage.

However, the sample percentage will not be exactly equal to its expected value—it will be off by a chance error. How big is this error likely to be? The answer is given by the standard error. For the sociologist's problem, the standard error is 5 percentage points. In other words, the sociologist should expect the percentage of men in her sample to be off the percentage in the population by 5 percentage points or so. The method for calculating such standard errors will now be presented. The idea: (i) find the SE for the number of men in the sample; then (ii) convert to percent, relative to the size of the sample. The size of the sample just means the number of sample people—100, in this case.

To compute an SE, you need a box model. The sociologist took a sample of size 100 from a population consisting of 3,091 men and 3,581 women. She classified the people in the sample by sex and counted the men. So there should be only 1's and 0's in the box (section 5 of chapter 17). The number of men in the sample is like the sum of 100 draws from the box

$$\left|3{,}091\ \boxed{1}\text{'s}\ \ 3{,}581\ \boxed{0}\text{'s}\right|.$$

She used a simple random sample, so the tickets must be drawn without replacement. This completes the box model.



The fraction of 1's in the box is 0.46. Therefore, the SD of the box is $\sqrt{0.46 \times 0.54} \approx 0.50$. The SE for the sum of 100 draws is $\sqrt{100} \times 0.5 = 5$. The sum of 100 draws from the box will be around 46, give or take 5 or so. In other words, the number of men in the sociologist's sample of 100 is likely to be around 46, give or take 5 or so. The SE for the number of men is 5. Now 46 out of 100 is 46%, and 5 out of 100 is 5%. Therefore, the percentage of men in the sample is likely to be around 46%, give or take 5% or so. This 5% is the SE for the percentage of men in the sample.

> To compute the SE for a percentage, first get the SE for the corresponding number; then convert to percent, relative to the size of the sample. As a cold mathematical formula,
>
> $$\text{SE for percentage} = \frac{\text{SE for number}}{\text{size of sample}} \times 100\%.$$

What happens as the sample gets bigger? For instance, if the sociologist took a sample of size 400, the SE for the number of men in the sample would be

$$\sqrt{400} \times 0.5 = 10.$$

Now 10 represents 2.5% of 400, the size of the sample. The SE for the percentage of men in a sample of 400 would be 2.5%. Multiplying the size of the sample by 4 divided the SE for the percentage by $\sqrt{4} = 2$.

> Multiplying the size of a sample by some factor divides the SE for a percentage not by the whole factor—but by its square root.

The formulas are exact when drawing with replacement. And they are good approximations for draws made without replacement, provided the number of draws is small relative to the number of tickets in the box. For example, take the sociologist's SE. No matter which 100 tickets are drawn, among the tickets left in the box, the percentage of 1's will be very close to 46%. So, as far as the chances are concerned, there isn't much difference between drawing with or without replacement. More about this in section 4.

This section showed how the SE for a percentage can be obtained from the SE for the corresponding number. But these two SEs behave quite differently. When the sample size goes up, the SE for the number goes up—and the SE for the percentage goes down. That is because the SE for the number goes up slowly relative to the sample size (pp. 276, 303):

- The SE for the sample number goes up like the square root of the sample size.
- The SE for the sample percentage goes down like the square root of the sample size.

## Exercise Set A

1. A town has 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters. A box model is used to work out the expected value and the SE for the percentage of Democrats in the sample. Match each phrase on list A with a phrase or a number on list B. (Items on list B may be used more than once, or not all.)

| *List A* | *List B* |
|---|---|
| population | number of 1's among the draws |
| population percentage | percentage of 1's among the draws |
| sample | 40% |
| sample size | box |
| sample number | draws |
| sample percentage | 1,000 |
| denominator for sample percentage | 12,000 |

2. A university has 25,000 students, of whom 10,000 are older than 25. The registrar draws a simple random sample of 400 students.
   (a) Find the expected value and SE for the number of students in the sample who are older than 25.
   (b) Find the expected value and SE for the percentage of students in the sample who are older than 25.
   (c) The percentage of students in the sample who are older than 25 will be around _____, give or take _____ or so.

3. A coin will be tossed 10,000 times. Match the SE with the formula. (One formula will be left over.)

| *SE for the . . .* | *Formula* |
|---|---|
| percentage of heads | $\sqrt{10{,}000} \times 50\%$ |
| number of heads | $\dfrac{50}{10{,}000} \times 100\%$ |
| | $\sqrt{10{,}000} \times 0.5$ |

4. Five hundred draws are made at random with replacement from $\boxed{0\ 0\ 0\ 1}$. True or false, and explain:
   (a) The number of 1's among the draws is exactly equal to the sum of the draws.
   (b) The expected value for the percentage of 1's among the draws is exactly equal to 25%.

5. The box $\boxed{0\ 0\ 0\ 1\ 2}$ has an average of 0.6, and the SD is 0.8. True or false: the SE for the percentage of 1's in 400 draws can be found as follows—

$$\text{SE for number of 1's} = \sqrt{400} \times 0.8 = 16$$

$$\text{SE for percent of 1's} = \frac{16}{400} \times 100\% = 4\%$$

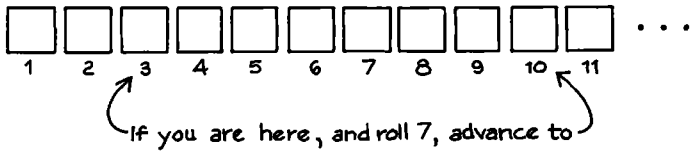
Explain briefly.

6. Nine hundred draws are made at random with replacement from a box which has 1 red marble and 9 blue ones. The SE for the percentage of red marbles in the sample is 1%. A sample percentage which is 1 SE above its expected value equals _____.

$$10\% + 1\% \qquad 1.01 \times 10\%$$

Choose one option, and explain briefly.

7. Someone plays a dice game 100 times. On each play, he rolls a pair of dice, and then advances his token along the line by a number of squares equal to the total number of spots thrown. (See the diagram.) About how far does he move? Give or take how much?




8. According to Sherlock Holmes,

> While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will be up to, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.[2]

The statistician doesn't quite say that. What is Sherlock Holmes forgetting?

*The answers to these exercises are on pp. A79–80.*

*Technical note.* When drawing at random with replacement from a 0–1 box, the SE for the number of 1's among the draws is

$$\sqrt{\text{no. of draws}} \times \text{SD of box.}$$

So the SE for the percentage of 1's among the draws is

$$(\sqrt{\text{no. of draws}} \times \text{SD of box/no. of draws}) \times 100\%.$$

By algebra, this simplifies to $(\text{SD of box}/\sqrt{\text{no. of draws}}) \times 100\%$. In many books, this would be written $(\sqrt{pq}/\sqrt{n}) \times 100\%$, where $p$ is the fraction of 1's in the box, $q$ is the fraction of 0's, and $n$ is the number of draws.

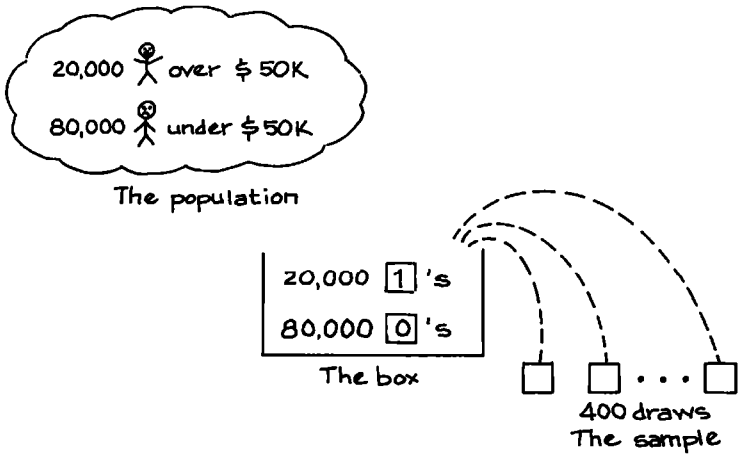
## 3. USING THE NORMAL CURVE

This section will review the expected value and SE for a sample percentage, and use the normal curve to compute chances.

*Example 1.* In a certain town, the telephone company has 100,000 subscribers. It plans to take a simple random sample of 400 of them as part of a market research study. According to Census data, 20% of the company's subscribers earn over $50,000 a year. The percentage of persons in the sample with incomes over $50,000 a year will be around _____, give or take _____ or so.

*Solution.* The first step is to make a box model. Taking a sample of 400 subscribers is like drawing 400 tickets at random from a box of 100,000 tickets. There is one ticket in the box for each person in the population, and one draw for each person in the sample. The drawing is done at random without replacement.

The problem involves classifying the people in the sample according to whether their incomes are more than $50,000 a year or not, and then counting the ones whose incomes are above that level. So each ticket in the box should be marked 1 or 0. The people earning more than $50,000 get 1's and the others get 0's. It is given that 20% of the subscribers earn more than $50,000 a year, so 20,000 of the tickets in the box are marked 1. The other 80,000 are marked 0. The sample is like 400 draws from the box. And the number of people in the sample who earn more than $50,000 a year is like the sum of the draws. That completes the first step, setting up the box model.




Now you have to work on the sum of the draws from the 0–1 box. The expected value for the sum is $400 \times 0.2 = 80$. To compute the standard error, you need the SD of the box. This is $\sqrt{0.2 \times 0.8} = 0.4$. There are 400 draws, so the SE for the sum is $\sqrt{400} \times 0.4 = 8$. The sum will be around 80, give or take 8 or so. In other words, the number of people in the sample earning more than $50,000 a year will be around 80, give or take 8 or so.

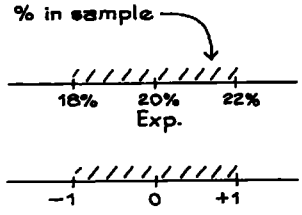
However, the question is about percent. You convert to percent relative to the size of the sample: 80 out of 400 is 20%, and 8 out of 400 is 2%. The expected value for the sample percentage is 20%, and the SE is 2%. That completes the solution: the percentage of high earners in the sample will be around 20%, give or take 2% or so. (It may be unfortunate, but statisticians use the %-sign as an abbreviation both for "percent" and for "percentage point.")

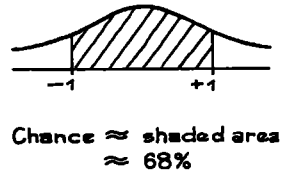
Of course, the expected value for the sample percent is pretty easy to figure, without the detour through the sample number. When drawing at random from a box of 0's and 1's, the expected value for the percentage of 1's among the draws equals the percentage of 1's in the box (p. 359).

> When drawing at random from a box of 0's and 1's, the percentage of 1's among the draws is likely to be around _____, give or take _____ or so. The expected value for the percentage of 1's among the draws fills in the first blank. The SE for the percentage of 1's among the draws fills in the second blank.

*Example 2.* (Continues example 1.) Estimate the chance that between 18% and 22% of the persons in the sample earn more than $50,000 a year.

*Solution.* The expected value for the sample percentage is 20%, and the SE is 2%. Now convert to standard units:




This completes the solution.

Here, the normal curve was used to figure chances. Why is that legitimate? There is a probability histogram for the number of high earners in the sample (figure 3). Areas in this histogram represent chances. For instance, the area between 80 and 90 represents the chance of drawing a sample which has between 80 and 90 high earners. As discussed in chapter 18, this probability histogram follows the normal curve (top panel of figure 3). Conversion to percent is only a change of scale, so the probability histogram for the sample percentage (bottom panel) looks just like the top histogram—and follows the curve too. In example 2, the curve was used on the probability histogram for the sample percentage, not on a histogram for data.

Examples 1 and 2 are about qualitative data. The incomes start out as quantitative data—numbers. However, the problems involve classifying and counting. Each person is classified as earning more than $50,000 a year, or less. Then the high earners are counted. In other words, the data are treated as qualitative: each income either has or doesn't have the quality of being more than $50,000 a year.

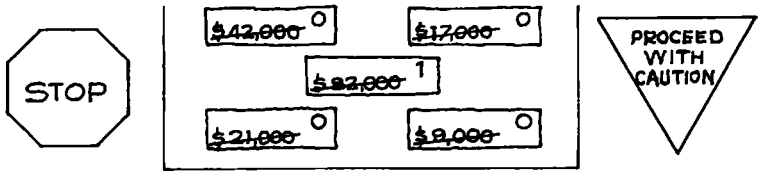
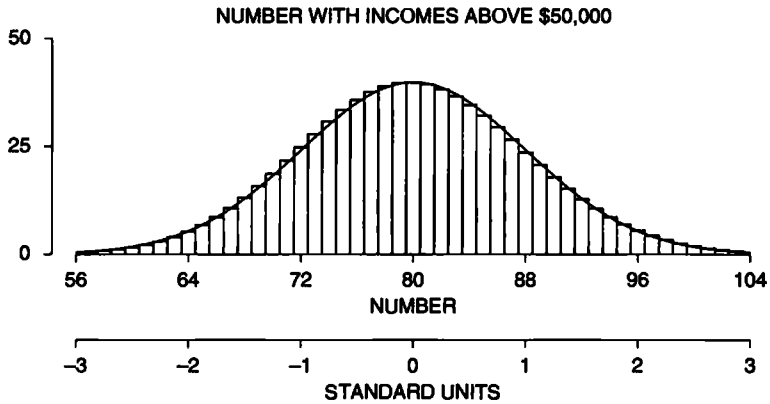
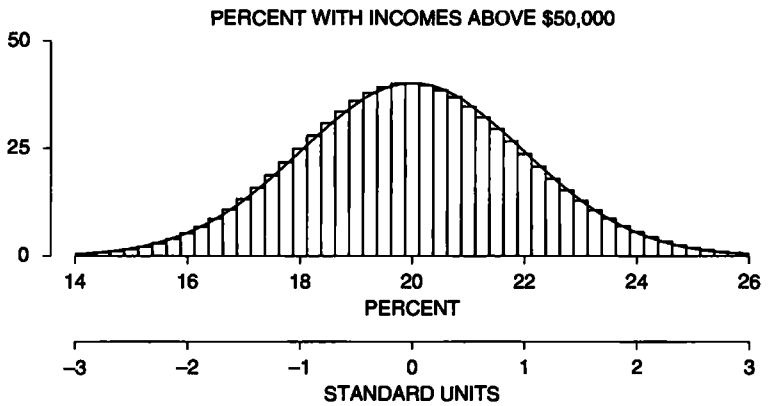


Figure 3. The top panel shows the probability histogram for the number of sample persons with incomes over $50,000. The bottom panel shows the probability histogram for the percentage of sample persons with incomes over $50,000. In standard units, the two histograms are exactly the same.[3] (Four hundred persons are chosen at random from a population of 100,000.)




When do you change to a 0–1 box? To answer this question, think about the arithmetic being done on the sample values. The arithmetic might involve:

- adding up the sample values, to get an average;

or

- classifying and counting, to get a percent.

If the problem is about classifying and counting, put 0's and 1's in the box (section 5 of chapter 17).

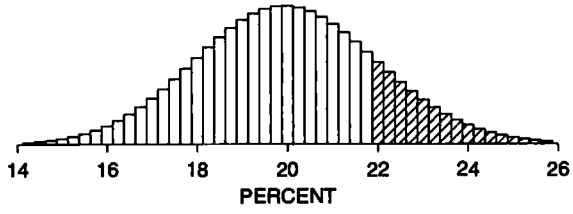
## Exercise Set B

1. You are drawing at random from a large box of red and blue marbles. Fill in the blanks.
   (a) The expected value for the percentage of reds in the _____ equals the percentage of reds in the _____. Options: sample, population
   (b) As the number of draws goes up, the SE for the _____ of reds in the sample goes up but the SE for the _____ of reds goes down. Options: number, percentage

2. In a certain town, there are 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters.
   (a) The expected value for the percentage of Democrats in the sample is _____. The SE for the percentage of Democrats in the sample is _____.
   (b) The percentage of Democrats in the sample is likely to be around _____, give or take _____ or so.
   (c) Find the chance that between 39% and 41% of the registered voters in the sample are Democrats.

3. According to the Census, a certain town has a population of 100,000 people age 18 and over. Of them, 60% are married, 10% have incomes over $75,000 a year, and 20% have college degrees.[4] As part of a pre-election survey, a simple random sample of 1,600 people will be drawn from this population.
   (a) To find the chance that 58% or less of the people in the sample are married, a box model is needed. Should the number of tickets in the box be 1,600, or 100,000? Explain. Then find the chance.
   (b) To find the chance that 11% or more of the people in the sample have incomes over $75,000 a year, a box model is needed. Should each ticket in the box show the person's income? Explain. Then find the chance.
   (c) Find the chance that between 19% and 21% of the people in the sample have a college degree.

4. The figure below is the probability histogram for the percent of sample persons with incomes above $50,000 (example 1, and bottom panel of figure 3). The shaded area represents _____. Fill in the blank with a phrase.




5. (a) In the top panel of figure 3, the area of the rectangle over 88 represents what?
   (b) In the bottom panel of figure 3, the area of the rectangle over 22% represents what?
   (c) The rectangles in parts (a) and (b) have equal areas. Is that a coincidence?

*The answers to these exercises are on pp. A80–81.*

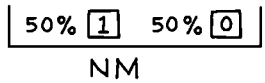
## 4. THE CORRECTION FACTOR

It is just after Labor Day, 2004. The presidential campaign (Bush versus Kerry) is in full swing, and the focus is on the Southwest. Pollsters are trying to predict the results. There are about 1.5 million eligible voters in New Mexico, and about 15 million in the state of Texas. Suppose one polling organization takes a simple random sample of 2,500 voters in New Mexico, in order to estimate the percentage of voters in that state who are Democratic. Another polling organiza tion takes a simple random sample of 2,500 voters from Texas. Both polls use exactly the same techniques. Both estimates are likely to be a bit off, by chance error. For which poll is the chance error likely to be smaller?

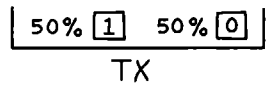
The New Mexico poll is sampling one voter out of 600, while the Texas poll is sampling one voter out of 6,000. It does seem that the New Mexico poll should be more accurate than the Texas poll. However, this is one of the places where intuition comes into head-on conflict with statistical theory, and it is intuition which has to give way. In fact, the accuracy expected from the New Mexico poll is just about the same as the accuracy to be expected from the Texas poll.

> When estimating percentages, it is the absolute size of the sample which determines accuracy, not the size relative to the population. This is true if the sample is only a small part of the population, which is the usual case.[5]

A box model will help in focusing the issue. We'll need two boxes, NM and TX. Box NM represents New Mexico, box TX represents Texas. Box NM has 1,500,000 tickets, one for each voter. The tickets corresponding to Democrats are marked 1, the others are marked 0. To keep life simple, we make the percentage of 1's in the box equal to 50%. We hire a polling organization to take a simple random sample from box NM, without telling them what is in the box. (Remember, taking a simple random sample means drawing at random without replacement.) The job of the polling organization is to estimate the percentage of 1's in the box. Naturally, they use the percentage of 1's in their sample.




NM




TX

Now for Box TX. This represents Texas, so it has 15,000,000 tickets. Again, we mark 1 on half the tickets in the box, and 0 on the others. Another polling organization is hired to take a simple random sample of 2,500 tickets from box TX, without knowing the composition of the box. This organization too will estimate the percentage of 1's in the box by the percentage in the sample, and will be off by a chance error.

Box NM and box TX have been set up with the same percentage composition, and the two samples are the same size. Intuition would insist that the organization sampling from box NM will have a much smaller chance error, because

box NM is so much smaller. But statistical theory shows that the likely size of the chance error is just about the same for the two polls.

The issue has now been stated sharply. How does statistical theory justify itself? To begin with, suppose the samples were drawn with replacement. Then it wouldn't matter at all which box was used. There would be a 50–50 chance to get a 0 or a 1 on each draw, and the size of the box would be completely irrelevant. Box NM and box TX have the same SD of 0.5, so both polling organizations would have the same SE for the number of 1's among the draws:

$$\sqrt{2{,}500} \times 0.5 = 25.$$

As a result, they would both have the same SE for the percentage of 1's among the draws:

$$\frac{25}{2{,}500} \times 100\% = 1\%.$$

If they drew at random with replacement, both organizations would be off by about 1 percentage point or so.

In fact, the draws are made without replacement. However, the number of draws is just a tiny fraction of the number of tickets in the box. Taking the draws without replacement barely changes the composition of the box. On each draw, the chance of getting a 1 must still be very close to 50%, and similarly for 0. As far as the chances are concerned, there is almost no difference between drawing with or without replacement.

In essence, that is why the size of the population has almost nothing to do with the accuracy of estimates. Still, there is a shade of difference between drawing with and without replacement. When drawing without replacement, the box does get a bit smaller, reducing the variability slightly. So the SE for drawing without replacement is a little less than the SE for drawing with replacement. There is a mathematical formula that says how much smaller:

$$\begin{array}{c}\text{SE when drawing}\\ \text{WITHOUT replacement}\end{array} = \begin{array}{c}\text{correction}\\ \text{factor}\end{array} \times \begin{array}{c}\text{SE when drawing}\\ \text{WITH replacement}\end{array}$$

The correction factor itself is somewhat complicated:

$$\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - \text{one}}}$$

When the number of tickets in the box is large relative to the number of draws,

Table 3. The correction factor; the number of draws is fixed at 2,500.

| *Number of tickets in the box* | *Correction factor (to five decimals)* |
|---|---|
| 5,000 | 0.70718 |
| 10,000 | 0.86607 |
| 100,000 | 0.98743 |
| 500,000 | 0.99750 |
| 1,500,000 | 0.99917 |
| 15,000,000 | 0.99992 |

the correction factor is nearly 1 and can be ignored (table 3, p. 368). Then it is the absolute size of the sample which determines accuracy, through the SE for drawing with replacement. The size of the population does not really matter. On the other hand, if the sample is a substantial fraction of the population, the correction factor must be used.

In our box model, the percentage of 1's was the same for both boxes. In reality, the percentage of Democrats will be different for the two states. However, even quite a large difference will generally not matter very much. In the 2004 presidential election, for example, 50% of the voters in New Mexico chose the Republican candidate (Bush), compared to 61% in Texas.[6] But the SDs for the two states are almost the same:

| 50% [1] 50% [0] | | 61% [1] 39% [0] |
|---|---|---|
| NM | | TX |
| $SD = \sqrt{.50 \times .50} = .50$ | | $SD = \sqrt{.61 \times .39} \approx .49$ |

A sample of size 2,500 will do as well in Texas as in New Mexico, although Texas is 10 times larger. The Texan in the cartoon is just wrong.

"A BIG STATE NEEDS A BIG SAMPLE, PARDNER."

A non-mathematical analogy may help. Suppose you took a drop of liquid from a bottle, for chemical analysis. If the liquid is well mixed, the chemical composition of the drop should reflect the composition of the whole bottle, and it really wouldn't matter if the bottle was a test tube or a gallon jug. The chemist doesn't care whether the drop is 1% or 1/100 of 1% of the solution.

The analogy is precise. There is one ticket in the box for each molecule in the bottle. If the liquid is well mixed, the drop is like a random sample. The number of molecules in the drop corresponds to the number of tickets drawn. This number—the sample size—is so large that chance error in the percentages is negligible.

## Exercise Set C

1. One public opinion poll uses a simple random sample of size 1,500 drawn from a town with a population of 25,000. Another poll uses a simple random sample of size 1,500 from a town with a population of 250,000. The polls are trying to estimate the percentage of voters who favor single-payer health insurance. Other things being equal:
   (i) the first poll is likely to be quite a bit more accurate than the second.
   (ii) the second poll is likely to be quite a bit more accurate than the first.
   (iii) there is not likely to be much difference in accuracy between the two polls.

2. You have hired a polling organization to take a simple random sample from a box of 100,000 tickets, and estimate the percentage of 1's in the box. Unknown to them, the box contains 50% 0's and 50% 1's. How far off should you expect them to be:
   (a) if they draw 2,500 tickets?
   (b) if they draw 25,000 tickets?
   (c) if they draw 100,000 tickets?

3. A survey organization wants to take a simple random sample in order to estimate the percentage of people who have seen a certain television program. To keep the costs down, they want to take as small a sample as possible. But their client will only tolerate chance errors of 1 percentage point or so in the estimate. Should they use a sample of size 100, 2,500, or 10,000? You may assume the population to be very large; past experience suggests the population percentage will be in the range 20%–40%.

4. One hundred draws are made at random with replacement from each of the following boxes. The SE for the percentage of 1's among the draws is smallest for box _____ and largest for box _____. Or is the SE the same for all three boxes?
   (A) $\boxed{\boxed{0}\,\boxed{1}}$ (B) $\boxed{10\ \boxed{0}\text{'s}\ 10\ \boxed{1}\text{'s}}$ (C) $\boxed{1{,}000\ \boxed{0}\text{'s}\ 1{,}000\ \boxed{1}\text{'s}}$

5. A box contains 2 red marbles and 8 blue ones. Four marbles are drawn at random. Find the SE for the percentage of red marbles drawn, when the draws are made
   (a) with replacement. (b) without replacement.

*The answers to these exercises are on p. A81.*

## 5. THE GALLUP POLL

The Gallup Poll predicts the vote with good accuracy, by sampling several thousand eligible voters out of 200 million. How is this possible? The previous section focused on simple random sampling, but the conclusions hold for most probability methods of drawing samples, including the one used by the Gallup

Poll: the likely size of the chance error in sample percentages depends mainly on the absolute size of the sample, and hardly at all on the size of the population. The huge number of eligible voters makes it hard work to draw the sample, but does not affect the standard error.

Is 2,500 a big enough sample? The square root law provides a benchmark. For example, with 2,500 tosses of a coin, the standard error for the percentage of heads is only 1%. Similarly, with a sample of 2,500 voters, the likely size of the chance error is only a percentage point or so. That is good enough unless the election is very close, like Bush versus Gore in 2000. The Electoral College would be a major complication: the Gallup Poll only predicts the popular vote.

## 6. REVIEW EXERCISES

*Review exercises may also cover material from previous chapters.*

1. Complete the following table for the coin-tossing game.

| | *Number of heads* | | *Percent of heads* | |
|---|---|---|---|---|
| *Number of tosses* | *Expected value* | *SE* | *Expected value* | *SE* |
| 100 | 50 | 5 | 50% | 5% |
| 2,500 | | | | 1% |
| 10,000 | | | | |
| 1,000,000 | | | | |

2. A die is rolled one thousand times. The percentage of aces (⚀) should be around _____, give or take _____ or so.
   (a) The first step in solving this problem is
      (i) computing the SD of the box.
      (ii) computing the average of the box.
      (iii) setting up the box model.
   Choose one option and explain.
   (b) Now solve the problem.

3. A group of 50,000 tax forms has an average gross income of $37,000, with an SD of $20,000. Furthermore, 20% of the forms have a gross income over $50,000. A group of 900 forms is chosen at random for audit. To estimate the chance that between 19% and 21% of the forms chosen for audit have gross incomes over $50,000, a box model is needed.
   (a) Should the number of tickets in the box be 900 or 50,000?
   (b) Each ticket in the box shows

   a zero or a one  a gross income

   (c) True or false: the SD of the box is $20,000.
   (d) True or false: the number of draws is 900.
   (e) Find the chance (approximately) that between 19% and 21% of the forms chosen for audit have gross incomes over $50,000.

(f) With the information given, can you find the chance (approximately) that between 9% and 11% of the forms chosen for audit have gross incomes over $75,000? Either find the chance, or explain why you need more information.

4. As in exercise 3, except it is desired to find the chance (approximately) that the total gross income of the audited forms is over $33,000,000. Work parts (a) through (d); then find the chance or explain why you need more information.

5. (Hypothetical.) On the average, hotel guests who take elevators weigh about 150 pounds with an SD of about 35 pounds. An engineer is designing a large elevator for a convention hotel, to lift 50 such people. If she designs it to lift 4 tons, the chance it will be overloaded by a random group of 50 people is about _____. Explain briefly.

6. The Census Bureau is planning to take a sample amounting to 1/10 of 1% of the population in each state in order to estimate the percentage of the population in that state earning over $100,000 a year. Other things being equal:

   (i) The accuracy to be expected in California (population 35 million) is about the same as the accuracy to be expected in Nevada (population 2 million).
   (ii) The accuracy to be expected in California is quite a bit higher than in Nevada.
   (iii) The accuracy to be expected in California is quite a bit lower than in Nevada.

   Explain.

7. Five hundred draws are made at random from the box

   | 60,000 $\boxed{0}$'s 20,000 $\boxed{1}$'s |

   True or false, and explain:

   (a) The expected value for the percentage of 1's among the draws is exactly 25%.
   (b) The expected value for the percentage of 1's among the draws is around 25%, give or take 2% or so.
   (c) The percentage of 1's among the draws will be around 25%, give or take 2% or so.
   (d) The percentage of 1's among the draws will be exactly 25%.
   (e) The percentage of 1's in the box is exactly 25%.
   (f) The percentage of 1's in the box is around 25%, give or take 2% or so.

8. In a certain town, there are 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters. There is about a 50–50 chance that the percentage of Democrats in the sample will be bigger than _____. Fill in the blank, and explain.

9. Six hundred draws will be made at random with replacement from the box $\boxed{0}\boxed{0}\boxed{1}$. The number of 1's among the draws will be around _____ give or take _____ or so.

10. A coin is tossed 2,000 times. Someone wishes to compute the SE for the number of heads among the tosses as $\sqrt{2{,}000} \times 0.5 \approx 22$. Is this the right SE? Answer yes or no, and explain briefly.

11. A university has 25,000 students, of whom 17,000 are undergraduates. The housing office takes a simple random sample of 500 students; 357 out of the 500 are undergraduates. Fill in the blanks.

    (a) For the number of undergraduates in the sample, the observed value is _____ but the expected value is _____.

    (b) For the percentage of undergraduates in the sample, the observed value is _____ but the expected value is _____.

12. There are 50,000 households in a certain city. The average number of persons age 16 and over living in each household is known to be 2.38; the SD is 1.87. A survey organization plans to take a simple random sample of 400 households, and interview all persons age 16 and over living in the sample households. The total number of interviews will be around _____, give or take _____ or so. Explain briefly.

## 7. SUMMARY

1. The sample is only part of the population, so the percentage composition of the sample usually differs by some amount from the percentage composition of the whole population.

2. For probability samples, the likely size of the chance error (the amount off) is given by the standard error.

3. To figure the SE, a box model is needed. When the problem involves classifying and counting, or taking percents, there should only be 0's and 1's in the box. Change the box, if necessary.

4. When drawing at random from a 0–1 box, the expected value for the percentage of 1's in the sample equals the percentage of 1's in the box. To find the SE for the percentage, first get the SE for the corresponding number, then convert to percent. The formula:

$$\text{SE for percentage} = \frac{\text{SE for number}}{\text{size of sample}} \times 100\%.$$

5. When the sample is only a small part of the population, the number of individuals in the population has almost no influence on the accuracy of the sample percentage. It is the absolute size of the sample (that is, the number of individuals in the sample) which matters, not the size relative to the population.

6. The square root law is exact when draws are made with replacement. When the draws are made without replacement, the formula gives a good approximation—provided the number of tickets in the box is large relative to the number of draws.

7. When drawing without replacement, to get the exact SE you have to multiply by the correction factor:

$$\sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - \text{one}}}$$

When the number of tickets in the box is large relative to the number of draws, the correction factor is nearly one.

# 21

# The Accuracy of Percentages

*In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment, and a very easy one, but people do not practise it much.... Most people, if you describe a train of events to them, will tell you what the result would be. They can put those events together in their minds, and argue from them that something will come to pass. There are few people, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk of reasoning backward....*

—*Sherlock Holmes*[1]

## 1. INTRODUCTION

The previous chapter reasoned from the box to the draws. Draws were made at random from a box whose composition was known, and a typical problem was finding the chance that the percentage of 1's among the draws would be in a given interval. As Sherlock Holmes points out, it is often very useful to turn this reasoning around, going instead from the draws to the box. A statistician would call this *inference* from the sample to the population. Inference is the topic of this chapter.

For example, suppose a survey organization wants to know the percentage of Democrats in a certain district. They might estimate it by taking a simple random sample. Naturally, the percentage of Democrats in the sample would be used to estimate the percentage of Democrats in the district—an example of reasoning backward from the draws to the box. Because the sample was chosen at random,

it is possible to say how accurate the estimate is likely to be, just from the size and composition of the sample. This chapter will explain how.

The technique is one of the key ideas in statistical theory. It will be presented in the polling context. A political candidate wants to enter a primary in a district with 100,000 eligible voters, but only if he has a good chance of winning. He hires a survey organization, which takes a simple random sample of 2,500 voters. In the sample, 1,328 favor the candidate, so the percentage is

$$\frac{1{,}328}{2{,}500} \times 100\% \approx 53\%.$$

The candidate is discussing this result with his pollster.

*Politician.* I win.

*Pollster.* Not so fast. You want to know the percentage you'd get among all the voters in the district. We only have it in the sample.

*Politician.* But with a good sample, it's bound to be the same.

*Pollster.* Not true. It's what I said before. The percentage you get in the sample is different from what you'd get in the whole district. The difference is what we call chance error.

*Politician.* Could the sample be off by as much as three percentage points? If so, I lose.

*Pollster.* Actually, we can be about 95% confident that we're right to within two percentage points. It looks good.

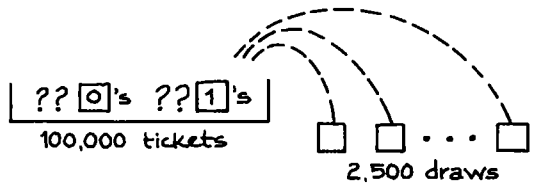
"I'M BEHIND YOU 100 PERCENT, PLUS OR MINUS 3 PERCENT OR SO."

*Politician.* What gives you the size of the chance error?

*Pollster.* The standard error. Remember, we talked about that the other day. As I was telling you....

*Politician.* Sorry, I'm expecting a phone call now.

The politician has arrived at the crucial question to ask when considering survey data: how far wrong is the estimate likely to be? As the pollster wanted to say, the likely size of the chance error is given by the standard error. To figure that, a box model is needed. There should be one ticket in the box for each voter, making 100,000 tickets in all. Each ticket should be marked 1 or 0, where 1 means a vote for the candidate, 0 a vote against him. There are 2,500 draws made at random from the box. The data are like the draws, and the number of voters in the sample who favor the candidate is like the sum of the draws. This completes the model.




To get the SE for the sum, the survey organization needs the SD of the box. This is

$$\sqrt{(\text{fraction of 1's}) \times (\text{fraction of 0's})}.$$

At this point, the pollsters seem to be stuck. They don't know how each ticket in the box should be marked. They don't even know the fraction of 1's in the box. That parameter represents the fraction of voters in the district who favor their candidate, which is exactly what they were hired to find out. (Hence the question marks in the box.)

Survey organizations lift themselves over this sort of obstacle by their own bootstraps.[2] They substitute the fractions observed in the sample for the unknown fractions in the box. In the example, 1,328 people out of the sample of 2,500 favored the candidate. So $1{,}328/2{,}500 \approx 0.53$ of the sample favored him, and the other 0.47 were opposed. The estimate is that about 0.53 of the 100,000 tickets in the box are marked 1, the other 0.47 being marked 0.

On this basis, the SD of the box is estimated as $\sqrt{0.53 \times 0.47} \approx 0.50$. The SE for the number of voters in the sample who favor the candidate is estimated as $\sqrt{2{,}500} \times 0.50 = 25$. The 25 measures the likely size of the chance error in the 1,328. Now 25 people out of 2,500 (the size of the sample) is 1%. The SE for the percentage of voters in the sample favoring the candidate is estimated as 1 percentage point. This completes the bootstrap procedure for estimating the standard error.

As far as the candidate is concerned, this calculation shows that his pollster's estimate of 53% is only likely to be off by 1 percentage point or so. It is very

unlikely to be off by as much as 3 percentage points—that's 3 SEs. He is well on the safe side of 50%, and he should enter the primary.

> *The bootstrap.* When sampling from a 0–1 box whose composition is unknown, the SD of the box can be estimated by substituting the fractions of 0's and 1's in the sample for the unknown fractions in the box. The estimate is good when the sample is reasonably large.

The bootstrap procedure may seem crude. But even with moderate-sized samples, the fraction of 1's among the draws is likely to be quite close to the fraction in the box. Similarly for the 0's. If survey organizations use their sample fractions in the formula for the SD of the box, they are not likely to be far wrong in estimating the SE.

One point is worth more discussion. The expected value for the number of 1's among the draws (translation—the expected number of sample voters who favor the candidate) is

$$2{,}500 \times \text{fraction of 1's in the box.}$$

This is unknown, because the fractions of 1's in the box is unknown. The SE of 25 says about how far the 1,328 is from its expected value. In statistical terminology, the 1,328 is an observed value; the contrast is with the unknown expected value. (Observed values are discussed on p. 292.)

*Example 1.* In fall 2005, a city university had 25,000 registered students. To estimate the percentage who were living at home, a simple random sample of 400 students was drawn. It turned out that 317 of them were living at home. Estimate the percentage of students at the university who were living at home in fall 2005. Attach a standard error to the estimate.

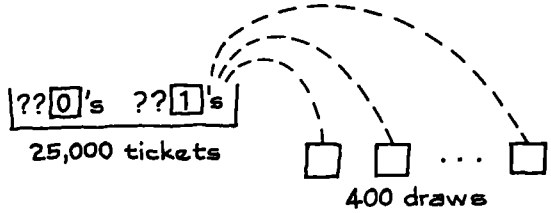
*Solution.* The sample percentage is

$$\frac{317}{400} \times 100\% \approx 79\%$$

That is the estimate for the population percentage.

For the standard error, a box model is needed. There are 25,000 tickets in the box, one for each student in the population. There are 400 draws from the box, one for each student in the sample. This problem involves classifying and counting, so each ticket in the box should be marked 1 or 0. We are counting students who were living at home. The tickets corresponding to these students should be marked 1; the others, 0. There are 400 draws made at random from the box. The data are like the draws, and the number of students in the sample who were living at home is like the sum of the draws. That completes the model. (See the sketch at the top of the next page.)

The fraction of 1's in the box is a parameter. It represents the fraction of all the students at this university who were living at home in fall 2005. It is unknown, but can be estimated as 0.79—the fraction observed in the sample. Similarly, the




fraction of 0's in the box is estimated as 0.21. So the SD of the box is estimated by the bootstrap method as $\sqrt{0.79 \times 0.21} \approx 0.41$. The SE for the number of students in the sample who were living at home is estimated as $\sqrt{400} \times 0.41 \approx 8$. The 8 gives the likely size of the chance error in the 317. Now convert to percent, relative to the size of the sample:

$$\frac{8}{400} \times 100\% = 2\%$$

The SE for the sample percentage is estimated as 2%. Let's summarize. In the sample, 79% of the students were living at home. The 79% is off the mark by 2 percentage points or so. That is what the SE tells us.

The discussion in this section focused on simple random sampling, where the mathematics is easiest. In practice, survey organizations use much more complicated designs. Even so, with probability methods it is generally possible to say how big the chance errors are likely to be—one of the great advantages of probability methods for drawing samples.

## Exercise Set A

1. Fill in the blanks, and explain.
   (a) In example 1 on p. 378, the 317 is the _____ value for the number of students in the sample who were living at home. Options:
   (i) expected (ii) observed
   (b) The SD of the box is _____ 0.41. Options:
   (i) exactly equal to (ii) estimated from the data as
   (c) The SE for the number of students in the sample who were living at home is _____ 8. Options: (i) exactly equal to (ii) estimated from the data as

2. In a certain city, there are 100,000 persons age 18 to 24. A simple random sample of 500 such persons is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in that city who are currently enrolled in college.[3] Put a give-or-take number on the estimate.
   (a) The first step in solving this problem is:
   (i) finding the SD of the box.
   (ii) finding the average of the box.
   (iii) writing down the box model.
   Choose one option, and explain.
   (b) Now solve the problem.

3. In a simple random sample of 100 graduates from a certain college, 48 were earning $50,000 a year or more. Estimate the percentage of all graduates of that college earning $50,000 a year or more.[4] Put a give-or-take number on the estimate.

4. A simple random sample of size 400 was taken from the population of all manufacturing establishments in a certain state: 11 establishments in the sample had 100 employees or more. Estimate the percentage of manufacturing establishments with 100 employees or more.[5] Attach a standard error to the estimate.

5. In the same state, a simple random sample of size 400 was taken from the population of all persons employed by manufacturing establishments: 187 people in the sample worked for establishments with 100 employees or more. Estimate the percentage of people who worked for establishments with 100 employees or more. Attach a standard error to the estimate.

6. Is the difference between the percentages in exercises 4 and 5 due to chance error?

*The next two exercises are designed to illustrate the bootstrap method for estimating the SD of the box.*

7. Suppose there is a box of 100,000 tickets, each marked 0 or 1. Suppose that in fact, 20% of the tickets in the box are 1's. Calculate the standard error for the percentage of 1's in 400 draws from the box.

8. Three different people take simple random samples of size 400 from the box in exercise 7, without knowing its contents. The number of 1's in the first sample is 72. In the second, it is 84. In the third, it is 98. Each person estimates the SE by the bootstrap method.
   (a) The first person estimates the percentage of 1's in the box as _____, and figures this estimate is likely to be off by _____ or so.
   (b) The second person estimates the percentage of 1's in the box as _____, and figures this estimate is likely to be off by _____ or so.
   (c) The third person estimates the percentage of 1's in the box as _____, and figures this estimate is likely to be off by _____ or so.

9. In a certain town, there are 25,000 people aged 18 and over. To estimate the percentage of them who watched a certain TV show, a statistician chooses a simple random sample of size 1,000. As it turns out, 308 of the sample people did see the show. Complete the following table; the first 3 lines refer to the sample percentage who saw the show. (N/A = not applicable.)

| | *Known to be* | *Estimated from the data as* |
|---|---|---|
| Observed value | 30.8% | N/A |
| Expected value | N/A | 30.8% |
| SE | | |
| SD of box | | |
| Number of draws | | |

*The answers to these exercises are on pp. A81–82.*

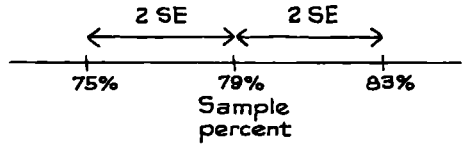
## 2. CONFIDENCE INTERVALS

In the example of the previous section, 79% of the students in the sample were living at home: the sample percentage was 79%. How far can the population percentage be from 79%? (Remember, "population percentage" means the percentage of all students at the university who were living at home.) The standard error was estimated as 2%, suggesting a chance error of around 2% in size. So the population percentage could easily be 77%. This would mean a chance error of 2%:

| sample percentage | = | population percentage | + | chance error |
|---|---|---|---|---|
| 79% | = | 77% | + | 2% |

The population percentage could also be 76%, corresponding to a chance error of 3%. This is getting unlikely, because 3% represents 1.5 SEs. The population percentage could even be as small as 75%, but this is still more unlikely; 4% represents 2 SEs. Of course, the population percentage could be on the other side of the sample percentage, corresponding to negative chance errors. For instance, the population percentage could be 83%. Then the estimate is low by 4%: the chance error is −4%, which is −2 SEs.

With chance errors, there is no sharp dividing line between the possible and the impossible. Errors larger in size than 2 SEs do occur—infrequently. What happens with a cutoff at 2 SEs? Take the interval from 2 SEs below the sample percentage to 2 SEs above:




This is a *confidence interval* for the population percentage, with a *confidence level* of about 95%. You can be about 95% confident that the population percentage is caught inside the interval from 75% to 83%.

What if you want a different confidence level? Anything except 100% is possible, by going the right number of SEs in either direction from the sample percentage. For instance:

- The interval "sample percentage ± 1 SE" is a 68%-confidence interval for the population percentage.
- The interval "sample percentage ± 2 SEs" is a 95%-confidence interval for the population percentage.
- The interval "sample percentage ± 3 SEs" is a 99.7%-confidence interval for the population percentage.

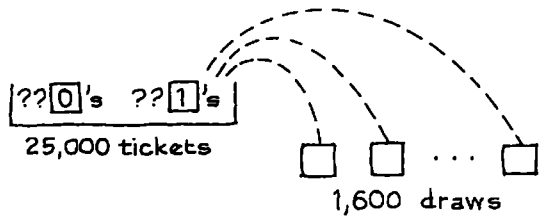
However, even 10 SEs may not give 100% confidence, because there is the remote possibility of very large chance errors. There are no definite limits to the normal curve: no matter how large a finite interval you choose, the normal curve has some area outside that interval.[6]

*Example 2.* A simple random sample of 1,600 persons is taken to estimate the percentage of Democrats among the 25,000 eligible voters in a certain town. It turns out that 917 people in the sample are Democrats. Find a 95%-confidence interval for the percentage of Democrats among all 25,000 eligible voters.

*Solution.* The percentage of Democrats in the sample is

$$\frac{917}{1{,}600} \times 100\% \approx 57.3\%.$$

The estimate: about 57.3% of the eligible voters in the town are Democrats. For the standard error, a box model is needed. There is one ticket in the box for each eligible voter in the town, making 25,000 tickets in all. There are 1,600 draws, corresponding to the sample size of 1,600. This problem involves classifying (Democrat or not) and counting, so each ticket is marked 1 or 0. It is Democrats that are being counted. So the tickets corresponding to Democrats are marked 1, the others are marked 0. There are 1,600 draws made at random from the box. The data are like the draws, and the number of Democrats in the sample is like the sum of the draws. That completes the model.




The fraction of 1's in the box (translation—the fraction of Democrats among the 25,000 eligible voters) is unknown, but can be estimated by 0.573, the fraction of Democrats in the sample. Similarly, the fraction of 0's in the box is estimated as 0.427. So the SD of the box is estimated by the bootstrap method as $\sqrt{0.573 \times 0.427} \approx 0.5$. The SE for the number of Democrats in the sample is estimated as $\sqrt{1{,}600} \times 0.5 = 20$. The 20 gives the likely size of the chance error in the 917. Now convert to percent, relative to the size of the sample:

$$\frac{20}{1{,}600} \times 100\% = 1.25\%.$$

The SE for the percentage of Democrats in the sample is 1.25%. The percentage of Democrats in the sample is likely to be off the percentage of Democrats in the population, by 1.25 percentage points or so. A 95%-confidence interval for the percentage of Democrats among all 25,000 eligible voters is

$$57.3\% \pm 2 \times 1.25\%.$$

That is the answer. We can be about 95% confident that between 54.8% and 59.8% of the eligible voters in this town are Democrats.

Confidence levels are often quoted as being "about" so much. There are two reasons. (i) The standard errors have been estimated from the data. (ii) The nor-

mal approximation has been used. If the normal approximation does not apply, neither do the methods of this chapter. There is no hard-and-fast rule for deciding. The best way to proceed is to imagine that the population has the same percentage composition as the sample. Then try to decide whether the normal approximation would work for the sum of the draws from the box. For instance, a sample percentage near 0% or 100% suggests that the box is lopsided, so a large number of draws will be needed before the normal approximation takes over (section 5 of chapter 18). On the other hand, if the sample percentage is near 50%, the normal approximation should be satisfactory when there are only a hundred draws or so.

## Exercise Set B

1. Fill in the blanks, and explain.
   (a) In example 2 on p. 382, the 917 is the _____ value for the number of Democrats in the sample. Options: (i) expected (ii) observed
   (b) The SD of the box is _____ $\sqrt{0.573 \times 0.427}$. Options:
   (i) exactly equal to (ii) estimated from the data as
   (c) The SE for the number of Democrats in the sample is _____ 20. Options:
   (i) exactly equal to (ii) estimated from the data as

2. Refer back to exercise 2 on p. 379.
   (a) Find a 95%-confidence interval for the percentage of persons age 18 to 24 in the city who are currently enrolled in college.
   (b) Repeat, for a confidence level of 99.7%.
   (c) Repeat, for a confidence level of 99.7%, supposing the size of the sample was 2,000, of whom 776 were currently enrolled in college.

3. A box contains 1 red marble and 99 blues; 100 marbles are drawn at random with replacement.
   (a) Find the expected number of red marbles among the draws, and the SE.
   (b) What is the chance of drawing fewer than 0 red marbles?
   (c) Use the normal curve to estimate this chance.
   (d) Does the probability histogram for the number of red marbles among the draws look like the normal curve?

4. A box contains 10,000 marbles, of which some are red and the others blue. To estimate the percentage of red marbles in the box, 100 are drawn at random without replacement. Among the draws, 1 turns out to be red. The percentage of red marbles in the box is estimated as 1%, with an SE of 1%. True or false: a 95%-confidence interval for the percentage of red marbles in the box is 1% ± 2%. Explain.

*The answers to these exercises are on pp. A82–83.*

## 3. INTERPRETING A CONFIDENCE INTERVAL

In example 1 on p. 378, a simple random sample was taken to estimate the percentage of students registered at a university in fall 2005 who were living at home. An approximate 95%-confidence interval for this percentage ran from 75%

to 83%, because

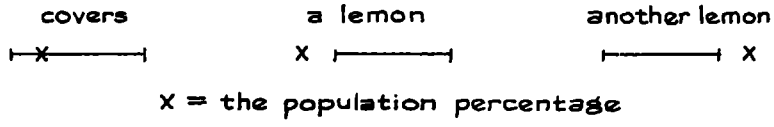
$$\text{sample percentage} \pm 2\,\text{SE} = 75\% \text{ to } 83\%.$$

It seems more natural to say "There is a 95% chance that the population percentage is between 75% and 83%." But there is a problem here. In the frequency theory, a chance represents the percentage of the time that something will happen. No matter how many times you take stock of all the students registered at that university in the fall of 2005, the percentage who were living at home back then will not change. Either this percentage was between 75% and 83%, or not. There really is no way to define the chance that the parameter will be in the interval from 75% to 83%. That is why statisticians have to turn the problem around slightly.[7] They realize that the chances are in the sampling procedure, not in the parameter. And they use the new word "confidence" to remind you of this.

| The chances are in the sampling procedure, not in the parameter. |
|---|

The confidence level of 95% says something about the sampling procedure, and we are going to see what that is. The first point to notice: the confidence interval depends on the sample. If the sample had come out differently, the confidence interval would have been different. With some samples, the interval "sample percentage ± 2 SE" traps the population percentage. (The word statisticians use is *cover*.) But with other samples, the interval fails to cover. It's like buying a used car. Sometimes you get a lemon—a confidence interval which doesn't cover the parameter.

Three confidence intervals




The confidence level of 95% can now be interpreted. For about 95% of all samples, the interval

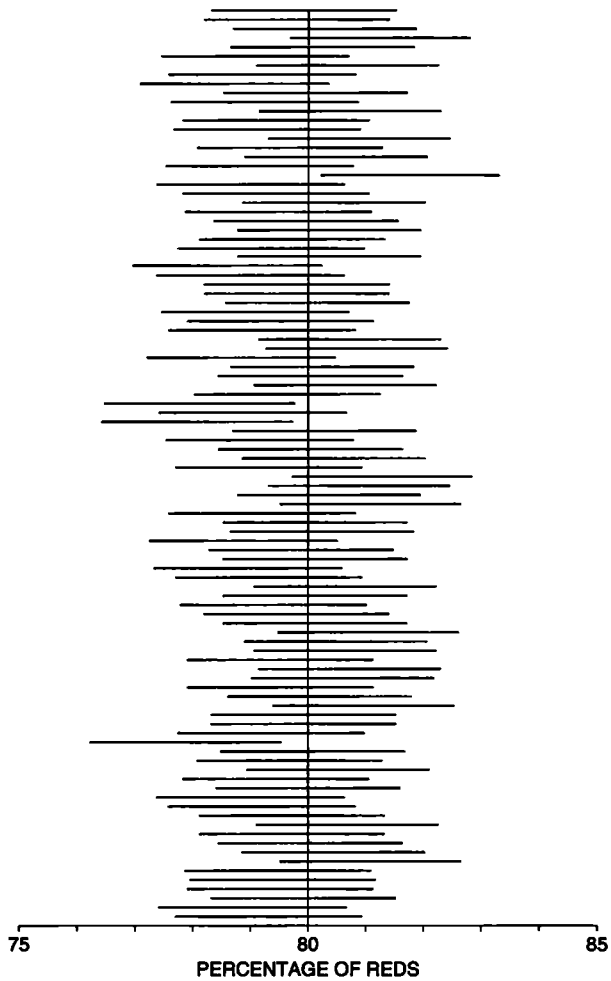
$$\text{sample percentage} \pm 2\,\text{SE}$$

covers the population percentage, and for the other 5% it does not. Of course, investigators usually cannot tell whether their particular interval covers the population percentage, because they do not know that parameter. But they are using a procedure that works 95% of the time: take a simple random sample, and go 2 SEs either way from the sample percentage. It is as if their interval was drawn at random from a box of intervals, where 95% cover the parameter and only 5% are lemons. This beats second-hand cars.

> A confidence interval is used when estimating an unknown parameter from sample data. The interval gives a range for the parameter, and a confidence level that the range covers the true value.

Confidence levels are a bit difficult, because they involve thinking not only about the actual sample but about other samples that could have been drawn. The interpretation is illustrated in figure 1. A hundred survey organizations are hired to estimate the percentage of red marbles in a large box. Unknown to the pollsters,

Figure 1. Interpreting confidence intervals. The 95%-confidence interval is shown for 100 different samples. The interval changes from sample to sample. For about 95% of the samples, the interval covers the population percentage, marked by a vertical line.[8]



this percentage is 80%. Each organization takes a simple random sample of 2,500 marbles, and computes a 95%-confidence interval for the percentage of reds in the box, using the formula

$$\text{percentage of reds in sample} \pm 2\,\text{SE}.$$

The percentage of reds is different from sample to sample, and so is the estimated standard error. As a result, the intervals have different centers and lengths. Some of the intervals cover the percentage of red marbles in the box, others fail. About 95% of them should cover the percentage, which is marked by a vertical line. In fact, 96 out of 100 do. Of course, this is only a computer simulation, designed to illustrate the theory. In practice, an investigator would have only one sample, and would not know the parameter.

Probabilities are used when you reason forward, from the box to the draws; confidence levels are used when reasoning backward, from the draws to the box (see the chapter opening quote). There is a lot to think about here, but keep the main idea of the chapter in mind.

> A sample percentage will be off the population percentage, due to chance error. The SE tells you the likely size of the amount off.

Confidence levels were introduced to make this idea more quantitative.

## Exercise Set C

1. Probabilities are used when reasoning from the _____ to the _____; confidence levels are used when reasoning from the _____ to the _____. Options:
   box draws

2. (a) The chance error is in the _____ value. Options: observed, expected
   (b) The confidence interval is for the _____ percentage. Options:
   sample population

3. Refer to exercises 7 and 8 on p. 380. Compute a 95%-confidence interval for the percentage of 1's in the box, using the data obtained by the person in exercise 8(a). Repeat for the other two people. Which of the three intervals cover the population percentage, that is, the percentage of 1's in the box? Which do not? (Remember, the three people in exercise 8 do not know the contents of the box; but you do, from exercise 7.)

4. A box contains a large number of red and blue marbles; the proportion of red marbles is known to be 50%. A simple random sample of 100 marbles is drawn from the box. Say whether each of the following statements is true or false, and explain briefly.
   (a) The percentage of red marbles in the sample has an expected value of 50%, and an SE of 5%.
   (b) The 5% measures the likely size of the chance error in the 50%.

(c) The percentage of reds in the sample will be around 50%, give or take 5% or so.
(d) An approximate 95%-confidence interval for the percentage of reds in the sample is 40% to 60%.
(e) There is about a 95% chance that the percentage of reds in the sample will be in the range from 40% to 60%.

5. A box contains a large number of red and blue marbles, but the proportions are unknown; 100 marbles are drawn at random, and 53 turn out to be red. Say whether each of the following statements is true or false, and explain briefly.
   (a) The percentage of red marbles in the box can be estimated as 53%; the SE is 5%.
   (b) The 5% measures the likely size of the chance error in the 53%.
   (c) The 53% is likely to be off the percentage of red marbles in the box, by 5% or so.
   (d) A 95%-confidence interval for the percentage of red marbles in the box is 43% to 63%.
   (e) A 95%-confidence interval for the percentage of red marbles in the sample is 43% to 63%.

6. A simple random sample of 1,000 persons is taken to estimate the percentage of Democrats in a large population. It turns out that 543 of the people in the sample are Democrats. True or false, and explain:
   (a) The sample percentage is $(543/1{,}000) \times 100\% = 54.3\%$; the SE for the sample percentage is 1.6%.
   (b) $54.3\% \pm 3.2\%$ is a 95%-confidence interval for the population percentage.
   (c) $54.3\% \pm 3.2\%$ is a 95%-confidence interval for the sample percentage.
   (d) There is about a 95% chance for the percentage of Democrats in the population to be in the range $54.3\% \pm 3.2\%$.

7. (Continues exercise 6; hard.) True or false, and explain: If another survey organization takes a simple random sample of 1,000 persons, there is about a 95% chance that the percentage of Democrats in their sample will be in the range $54.3\% \pm 3.2\%$.

8. At a large university, 54.3% of the students are female and 45.7% are male. A simple random sample of 1,000 persons is drawn from this population. The SE for the sample percentage of females is figured as 1.6%. True or false: There is about a 95% chance for the percentage of females in the sample to be in the range $54.3\% \pm 3.2\%$. Explain.

*The answers to these exercises are on pp. A83–84.*

## 4. *CAVEAT EMPTOR*

The methods of this chapter were developed for simple random samples. They may not apply to other kinds of samples. Many survey organizations use fairly complicated probability methods to draw their samples (section 4 of chapter 19). As a result, they have to use more complicated methods for estimating their standard errors. Some survey organizations do not bother to use probability methods at all. Watch out for them.

> Warning. The formulas for simple random samples may not apply to other kinds of samples.

Here is the reason. Logically, the procedures in this chapter all come out of the square root law (section 2 of chapter 17). When the size of the sample is small relative to the size of the population, taking a simple random sample is just about the same as drawing at random with replacement from a box—the basic situation to which the square root law applies. The phrase "at random" is used here in its technical sense: at each stage, every ticket in the box has to have an equal chance to be chosen. If the sample is not taken at random, the square root law does not apply, and may give silly answers.[9]

People often think that a statistical formula will somehow check itself while it is being used, to make sure that it applies. Nothing could be further from the truth. In statistics, as in old-fashioned capitalism, the responsibility is on the consumer.

Caveat emptor

Let the buyer beware

$$\bar{x} \pm z_\alpha \times s/\sqrt{n}$$

## Exercise Set D

1. A psychologist is teaching a class with an enrollment of 100. He administers a test of passivity to these students and finds that 20 of them score over 50. The conclusion: approximately 20% of all students would score over 50 on this test. Recognizing that this estimate may be off a bit, he estimates the likely size of the error as follows:

$$\text{SE for number} = \sqrt{100} \times \sqrt{0.2 \times 0.8} = 4$$

$$\text{SE for percent} = (4/100) \times 100\% = 4\%$$

   What does statistical theory say?

2. A small undergraduate college has 1,000 students, evenly distributed among the four classes: freshman, sophomore, junior, and senior. In order to estimate the percentage of students who have ever smoked marijuana, a sample is taken by the following procedure: 25 students are selected at random without replacement from each of the four classes. As it turns out, 35 out of the 100 sample students admit to having smoked. So, it is estimated that 35% out of the 1,000 students at the college would admit to having smoked. A standard error is attached to this estimate, by the following procedure:

$$\text{SE for number} = \sqrt{100} \times \sqrt{0.35 \times 0.65} \approx 5$$

$$\text{SE for percent} = (5/100) \times 100\% = 5\%$$

What does statistical theory say?

*The answers to these exercises are on p. A84.*

## 5. THE GALLUP POLL

The Gallup Poll does not use a simple random sample (section 4 of chapter 19). As a result, they do not estimate their standard errors using the method of this chapter. However, it is interesting to compare their samples to simple random samples of the same size. For instance, in 1952 they predicted a 51% vote for Eisenhower, based on a sample of 5,385 people. With a simple random sample,

$$\text{SE for number} = \sqrt{5{,}385} \times \sqrt{0.51 \times 0.49} \approx 37$$

$$\text{SE for percent} = \frac{37}{5{,}385} \times 100\% \approx 0.7 \text{ of } 1\%.$$

In fact, Eisenhower got 54.9% of the vote in that election. The Gallup Poll estimate was off by 3.9 percentage points. This is nearly 6 times the SE for a simple random sample. Table 1 shows the comparison for every presidential election from 1952 to 2004.

Table 1. Comparing the Gallup Poll with a simple random sample. The errors of prediction are on the whole quite a bit bigger than those to be expected from a simple random sample of the same size.

| *Year* | *Sample size* | *SE for simple random sample* | *Actual error* |
|---|---|---|---|
| 1952 | 5,385 | 0.7 of 1% | 3.9% |
| 1956 | 8,144 | 0.5 of 1% | 2.1% |
| 1960 | 8,015 | 0.6 of 1% | 1.3% |
| 1964 | 6,625 | 0.6 of 1% | 2.9% |
| 1968 | 4,414 | 0.7 of 1% | 0.4 of 1% |
| 1972 | 3,689 | 0.8 of 1% | 1.8% |
| 1976 | 3,439 | 0.9 of 1% | 2.0% |
| 1980 | 3,500 | 0.8 of 1% | 3.5% |
| 1984 | 3,456 | 0.8 of 1% | 0.5 of 1% |
| 1988 | 4,089 | 0.8 of 1% | 2.9% |
| 1992 | 2,019 | 1.1% | 6.1% |
| 1996 | 2,895 | 0.9% | 2.8% |
| 2000 | 3,571 | 0.8 of 1% | 0.2% |
| 2004 | 2,014 | 1.1% | 1.6% |

Source: See table 4 in chapter 19.

In 11 elections out of 14, the error was considerably larger than the SE for a simple random sample. One reason is that predictions are based only on part of the sample, namely, those people judged likely to vote (section 6 of chapter 19).

This eliminates about half the sample. Table 2 compares the errors made by the Gallup Poll with SEs computed for simple random samples whose size equals the number of likely voters. The simple random sample formula is still not doing a good job at predicting the size of the errors.

Why not? Well, the Gallup Poll is not drawing tickets at random from a box—although the telephone samples used from 1992 onwards come closer to simple random sampling than designs used before that (pp. 340–341, 346). Three other issues should be mentioned: (i) the process used to screen out the non-voters may break down at times; (ii) some voters may still not have decided how to vote when they are interviewed; (iii) voters may change their minds between the last pre-election poll and election day, especially in close contests. In a volatile, three-way contest like the 1992 election, such problems take their toll (p. 346).

Table 2. The accuracy of the Gallup Poll compared to that of a simple random sample whose size equals the number of likely voters in the Gallup Poll sample.

| *Year* | *Number of likely voters* | *SE for simple random sample* | *Actual error* |
|---|---|---|---|
| 1952 | 3,350 | 0.9 of 1% | 3.9% |
| 1956 | 4,950 | 0.7 of 1% | 2.1% |
| 1960 | 5,100 | 0.7 of 1% | 1.3% |
| 1964 | 4,100 | 0.8 of 1% | 2.9% |
| 1968 | 2,700 | 1.0% | 0.4 of 1% |
| 1972 | 2,100 | 1.1% | 1.8% |
| 1976 | 2,000 | 1.1% | 2.0% |
| 1980 | 2,000 | 1.1% | 3.5% |
| 1984 | 2,000 | 1.1% | 0.5 of 1% |
| 1988 | 2,600 | 1.0% | 2.9% |
| 1992 | 1,600 | 1.2% | 6.1% |
| 1996 | 1,100 | 1.5% | 2.8% |
| 2000 | 2,400 | 1.0% | 0.2% |
| 2004 | 1,600 | 1.2% | 1.6% |

Note: The number of likely voters is rounded.
Source: The Gallup Poll (American Institute of Public Opinion).

## Exercise Set E

1. A Gallup Poll pre-election survey based on a sample of 1,000 people estimates a 65% vote for the Democratic candidate in a certain election. True or false, and explain: the likely size of the chance error in this estimate can be figured as follows—

$$\sqrt{1{,}000} \times \sqrt{0.65 \times 0.35} \approx 15, \qquad \frac{15}{1{,}000} \times 100\% = 1.5\%$$

2. One thousand tickets are drawn at random without replacement from a large box, and 651 of the draws show a 1. The fraction of 1's in the box is estimated as 65%. True or false, and explain: the likely size of the chance error in this estimate can be figured as follows—

$$\sqrt{1{,}000} \times \sqrt{0.65 \times 0.35} \approx 15, \qquad \frac{15}{1{,}000} \times 100\% = 1.5\%$$

3. The following article appeared on the *New York Times* Op Ed page of August 27, 1988, headlined MAYBE BUSH HAS ALREADY WON.

   The presidential campaign, only now formally set to begin, is in fact virtually finished. Despite the Niagara of news stories about how the candidates are touting their running mates, haggling over debates and sniping at each other, the die is just about cast.

   A significant indicator is the Gallup Poll, which this week shows Vice President Bush ahead of Gov. Michael S. Dukakis by 4 percentage points. In the half century since George Gallup began his electoral opinion surveys in Presidential years, his "trial heats" in the last week or so of September have foretold with notable accuracy the outcome on election day.

   The late James A. Farley, the Democrats' peerless tactician of 50 years ago, always argued that voters made up their minds by Labor Day.... It is now established, moreover, that when traditional nonvoters—the object of get-out-the-vote efforts—are persuaded to vote, they too cast their ballots in the same proportion as the rest of the electorate.... Significant changes in the percentages from September to November are due only to altered voter enthusiasm....

   (a) How does the article explain differences in voter opinion between September and November?
   (b) What else could explain a difference between Gallup Poll results in late September and election results in early November?
   (c) A difference of several percentage points between Gallup Poll results in late September and election results in early November is: very unlikely, unlikely but possible, quite possible. Choose one option, and explain.

*The answers to these exercises are on p. A84.*

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. A survey organization draws a simple random sample of 1,000 registered voters in a certain town. In the sample, 32% approve of the Mayor. The organization estimates that 32% of all 50,000 registered voters in the town approve of the Mayor. How to figure the SE? The organization realizes that the number in the sample who approve _____ 1,000 draws _____ box _____. Fill in each blank (33 words or less). Then work out the SE.

2. The Residential Energy Consumption Survey found in 2001 that 47% of American households had internet access.[10] A market survey organization repeated this study in a certain town with 25,000 households, using a simple random sample of 500 households: 239 of the sample households had internet access.

   (a) The percentage of households in the town with internet access is estimated as _____; this estimate is likely to be off by _____ or so.
   (b) If possible, find a 95%-confidence interval for the percentage of all 25,000 households with internet access. If this is not possible, explain why not.

3. Of the 500 sample households in the previous exercise, 7 had three or more large-screen TVs.
   (a) The percentage of households in the town with three or more large-screen TVs is estimated as _____; this estimate is likely to be off by _____ or so.
   (b) If possible, find a 95%-confidence interval for the percentage of all 25,000 households with three or more large-screen TVs. If this is not possible, explain why not.

4. (This continues exercise 3.) Among the sample households, 121 had no car, 172 had one car, and 207 had two or more cars. Estimate the percentage of households in the town with one or more cars; attach a standard error to the estimate. If this is not possible, explain why not.

5. The National Assessment of Educational Progress administers standardized achievement tests to nationwide samples of 17-year-olds in school. One year, the tests covered history and literature. You may assume that a simple random sample of size 6,000 was taken. Only 36.1% of the students in the sample knew that Chaucer wrote *The Canterbury Tales*, but 95.2% knew that Edison invented the light bulb.[11]
   (a) If possible, find a 95%-confidence interval for the percentage of all 17-year-olds in school who knew that Chaucer wrote *The Canterbury Tales*. If this is not possible, why not?
   (b) If possible, find a 95%-confidence interval for the percentage of all 17-year-olds in school who knew that Edison invented the light bulb. If this is not possible, why not?

6. True or false: with a well-designed sample survey, the sample percentage is very likely to equal the population percentage. Explain.

7. (Hypothetical.) One year, there were 252 trading days on the New York Stock Exchange, and IBM common stock went up on 131 of them: $131/252 \approx 52\%$. A statistician attaches a standard error to this percentage as follows:

$$\text{SE for number} = \sqrt{252} \times \sqrt{0.52 \times 0.48} \approx 8$$

$$\text{SE for percent} = \frac{8}{252} \times 100\% \approx 3\%$$

   Is this the right SE? Answer yes or no, and explain.

8. A simple random sample of 3,500 people age 18 or over is taken in a large town to estimate the percentage of people (age 18 and over in that town) who read newspapers. It turns out that 2,487 people in the sample are newspaper readers.[12] The population percentage is estimated as

$$\frac{2{,}487}{3{,}500} \times 100\% \approx 71\%$$

   The standard error is estimated as 0.8 of 1%, because

$$\sqrt{3{,}500} \times \sqrt{0.71 \times 0.29} \approx 27, \qquad \frac{27}{3{,}500} \times 100\% \approx 0.8 \text{ of } 1\%$$

(a) Is 0.8 of 1% the right SE? Answer yes or no, and explain.
(b) 71% ± 1.6% is a _____ for the _____. Fill in the blanks and explain.

9. (Hypothetical.) A bank wants to estimate the amount of change people carry. They take a simple random sample of 100 people, and find that on the average, people in the sample carry 73¢ in change. They figure the standard error is 4¢, because

$$\sqrt{100} \times \sqrt{0.73 \times 0.27} \approx 4, \quad 4/100 = .04$$

Are they right? Answer yes or no, and explain.

10. In Keno, there are 80 balls numbered from 1 to 80, and 20 are drawn at random. If you play a double-number, you win if both numbers are chosen. This bet pays 11 to 1, and you have very close to a 6% chance of winning.[13] If you play 100 times and stake $1 on a double-number each time, your net gain will be around _____, give or take _____ or so.

11. One hundred draws will be made at random without replacement from a large box of numbered tickets. There are two options:

(i) To win $1 if the sum of the draws is bigger than 710.
(ii) To win $1 if the average of the draws is bigger than 7.1.

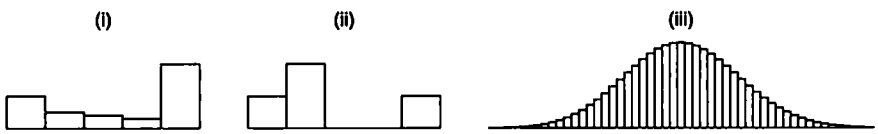
Which is better? Or are they the same? Explain.

12. A monthly opinion survey is based on a sample of 1,500 persons, "scientifically chosen as a representative cross section of the American public." The press release warns that the estimates are subject to chance error, but guarantees that they are "reliable to within two percentage points." The word "reliable" is ambiguous. According to statistical theory, the guarantee should be interpreted as follows:

(i) In virtually all these surveys, the estimates will be within two percentage points of the parameters.
(ii) In most such surveys, the estimates will be within two percentage points of the parameters, but in some definite percentage of the time larger errors are expected.

Explain.

13. One hundred draws are made at random with replacement from the box $\boxed{1}\boxed{2}\boxed{2}\boxed{5}$. One of the graphs below is a histogram for the numbers drawn. Another is the probability histogram for the sum. And the third is irrelevant. Which is which? Why?




14. A coin is tossed 1,000 times.

(a) Suppose it lands heads 529 times. Find the expected value for the number of heads, the chance error, and the standard error.

(b) Suppose it lands heads 484 times. Find the expected value for the number of heads, the chance error, and the standard error.

(c) Suppose it lands heads 514 times. Find the expected value for the number of heads, the chance error, and the standard error.

15. A survey organization takes a simple random sample of 1,500 persons from the residents of a large city. Among these sample persons, 1,035 were renters.

(a) The expected value for the percentage of sample persons who rent is _____ 69%.

(b) The SE for the percentage of sample persons who rent is _____ 1.2%.

Fill in the blanks, and explain. Options:

(i) exactly equal to (ii) estimated from the data as

## 7. SUMMARY

1. With a simple random sample, the sample percentage is used to estimate the population percentage.

2. The sample percentage will be off the population percentage, due to chance error. The SE for the sample percentage tells you the likely size of the amount off.

3. When sampling from a 0–1 box whose composition is unknown, the SD of the box can be estimated by substituting the fractions of 0's and 1's in the sample for the unknown fractions in the box. This *bootstrap estimate* is good when the sample is large.

4. A *confidence interval* for the population percentage is obtained by going the right number of SEs either way from the sample percentage. The confidence level is read off the normal curve. This method should only be used with large samples.

5. In the frequency theory of probability, parameters are not subject to chance variation. That is why confidence statements are made instead of probability statements.

6. The formulas for simple random samples may not apply to other kinds of samples. If the sample was not chosen by a probability method, watch out: SEs computed from the formulas may not mean very much.

# 22

# Measuring Employment and Unemployment

*The country is hungry for information; everything of a statistical character, or even a statistical appearance, is taken up with an eagerness that is almost pathetic; the community have not yet learned to be half skeptical and critical enough in respect to such statements.*

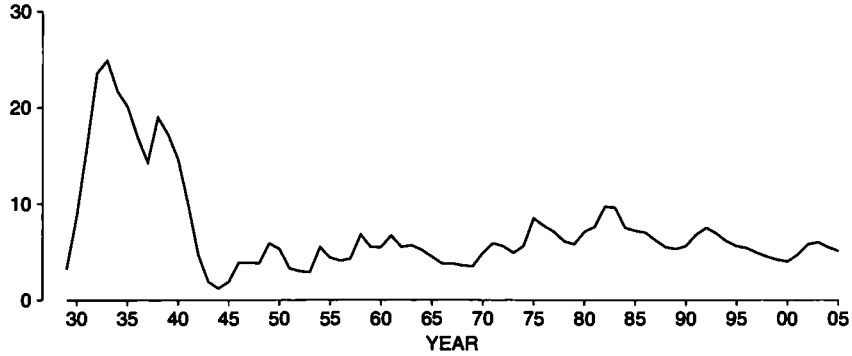
—GENERAL FRANCIS A. WALKER, SUPERINTENDENT OF THE 1870 CENSUS

## 1. INTRODUCTION

The unemployment rate is one of the most important numbers published by the government. Unemployment was only 3% in 1929, before the stock market crash (figure 1 on the next page). It reached 25% in the depths of the Depression, and remained fairly high until the U.S. entered World War II. More recently, as a result of anti-inflationary practices adopted by the Federal Reserve Board in 1981, the economy went into a deep recession in 1982–83, and the unemployment rate nearly reached 10%. By the late 1980s, the rate had dropped below 6%; in many metropolitan areas, there were shortages of skilled workers. In 2003, after the Internet bubble collapsed, the rate returned to 6%. Unemployment declined from there to the end of 2005.

The government agency in charge of the employment numbers is the Bureau of Labor Statistics. But how do they know who is employed or unemployed? Employment statistics are estimated from a sample survey—the Current Population Survey. This massive and beautifully organized sample survey is conducted

Figure 1. The unemployment rate from 1929 to 2005.




Source: *Employment and Earnings*, January 1976, table A-1; July 1989, table A-3; December 2005, table A-1.

monthly for the Bureau of Labor Statistics by the Census Bureau.[1] During the week containing the 19th day of the month, a field staff of 1,700 interviewers canvasses a nationwide probability sample of about 110,000 people. The size of the labor force, the unemployment rate, and a host of other economic and demographic statistics (like the distribution of income and educational level) are estimated from survey results, at a cost which in 2005 was about $60 million a year. The results are published in:

- *Monthly Labor Review*,
- *Employment and Earnings* (monthly),
- *The Employment Situation* (monthly),
- *Current Population Reports* (irregular),
- *Statistical Abstract of the United States* (annual),
- *Economic Report of the President* (annual).

The object of this chapter is to present the Current Population Survey in detail, from the ground up. This will illustrate and consolidate the ideas introduced in previous chapters. It should also make other large-scale surveys easier to understand. The main conclusions from this case study:

- In practice, fairly complicated probability methods must be used to draw samples. Simple random sampling is only a building-block in these designs.
- The standard-error formulas for simple random samples do not apply to these complicated designs, and other methods must be used for estimating the standard errors.

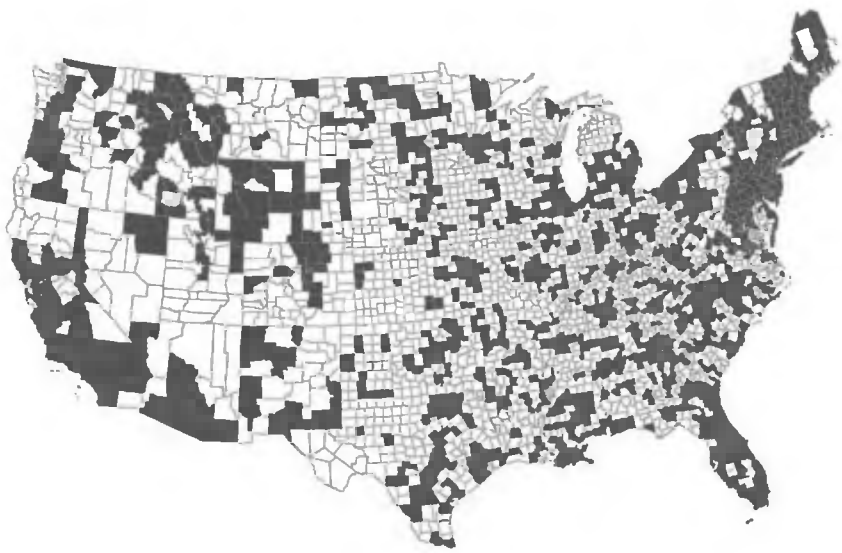
## 2. THE DESIGN OF THE CURRENT POPULATION SURVEY

The Current Population Survey is redesigned periodically by the Census Bureau, to take advantage of new information and to accomplish new objectives. There was a major redesign in the early 2000s, using data from the 2000 census.

There were 3,142 counties and independent cities in the U.S. As the first step in the redesign process, the Bureau put these together into groups to form 2,025 *Primary Sampling Units* (or PSUs, for short). Each PSU consisted either of a city, or a county, or a group of contiguous counties.[2] These PSUs were sorted into 824 *strata*, chosen so the PSUs in each stratum would resemble each other on certain demographic and economic characteristics (like unemployment at the time of stratification, the number of large households, and the number of workers in retail trade). The strata do not cross state lines. Many of the larger PSUs, like New York or Los Angeles, were put into strata by themselves.

The sample was chosen in two stages. To begin with, one PSU was chosen from each stratum, using a probability method which ensured that within the stratum, the chance of a PSU getting into the sample was proportional to its population. Since there were 824 strata, the first stage resulted in a sample of 824 PSUs. Until the next redesign (after the 2010 census), all interviewing for the Survey takes place in these 824 PSUs and in no others. The PSUs for an earlier design are shown in figure 2.

Figure 2. Primary Sampling Units for the Current Population Survey: the 1995 sample design with 792 PSUs.



Note: Alaska and Hawaii not shown.
Source: Bureau of the Census, Statistical Methods Division.

Each PSU was divided up into *Ultimate Sampling Units* (or USUs), consisting of about 4 housing units each. At the second stage, some USUs were picked at random for the sample. In the end, every person age 16 and over living in a selected USU in a selected PSU gets into the Current Population Survey. For the U.S. as a whole, the sampling rate is about 1 in 2,000. But the rate varies from about 1 in 300 for D.C. or Wyoming to 1 in 3,000 for large states like Califor-

nia, New York, and Texas.[3] The objective is to estimate unemployment rates in each of the 50 states and the District of Columbia with about the same precision.[4] This meant equalizing, at least roughly, the absolute sizes of the 51 subsamples (section 4 of chapter 20). So the ratio of sample size to population size has to be different from state to state.

The Bureau's choices for the sample to be used from 2005 to 2015 were all made well before 2005. The design even provided for people who were going to live in housing yet to be constructed. And in fact, the Bureau chose not just one sample but 16 different ones, in order to rotate part of the sample every month. After it gets into the sample, a housing unit is kept there for 4 months, dropped out for 8 months, and then brought back for a final 4 months. Why rotate the sample? For one reason, the interviewers may wear out their welcome after a while. Besides that, people's responses probably change as a result of being interviewed, progressively biasing the sample (this is called *panel bias*). For instance, there is some evidence to show that people are more likely to say they are looking for a job the first time they are interviewed than the second time. Then why not change the sample completely every month? Keeping part of it the same saves a lot of money. Besides that, having some overlap in the sample makes it easier to estimate the monthly changes in employment and unemployment.

## 3. CARRYING OUT THE SURVEY

The Survey design for 2005 produces 72,000 housing units to be canvassed each month. Of these, about 12,000 are ineligible for the sample (being vacant, or even demolished since the sample was designed). Another 4,500 or so are unavailable, because no one is at home, or because those at home will not cooperate. That leaves about 55,500 housing units in the Survey. All persons age 16 or over in these housing units are asked about their work experience in the previous week. On the basis of their answers, they are classified as:

- employed (those who did any paid work in the previous week, or were temporarily absent from a regular job);
- or unemployed (those who were not employed the previous week but were available for work and looking for work during the past four weeks);
- or outside the labor force (defying Aristotle, the Bureau defines this as the state of being neither employed nor unemployed).[5]

The employed are asked about the hours they work and the kind of job they have. The unemployed are asked about their last job, when and why they left it, and how they are looking for work. Those outside the labor force are asked whether they are keeping house, or going to school, or unable to work, or do not work for some other reason (in which case, they are asked to specify what). Results for November 2005 are shown in table 1.

Table 1. The civilian non-institutional population age 16 and over.[6] Bureau of Labor Statistics estimates, November 2005. In millions.

| | | | |
|---|---|---|---|
| Employed | 142.97 | | |
| Unemployed | 7.27 | | |
| Labor force | | 150.24 | |
| Outside the labor force | | 76.96 | |
| Total | | | 227.20 |

Source: *Employment and Earnings*, December 2005, table A-13.

By definition, the *civilian labor force* consists of the civilians who are either employed or unemployed. In November 2005, that amounted to $142.97 + 7.27 = 150.24$ million people.[7] The *unemployment rate* is the percentage of the civilian labor force which is unemployed, and that came to

$$\frac{7.27}{150.24} \times 100\% \approx 4.8\%.$$

This 4.8% is an average rate of unemployment, over all the subgroups of the population. Like many averages, it conceals some striking differences. These differences are brought out by a process of cross-tabulation. Unemployment falls more heavily on teenagers and blacks, as shown by by table 2.

Table 2. Unemployment rates by race, age, and sex. Bureau of Labor Statistics estimates, November 2005. In percent.

| | | *Age group* | | |
|---|---|---|---|---|
| *Race* | *Sex* | *16–19* | *20–64* | *65 and over* |
| White | Male | 15.1 | 3.5 | 3.2 |
| White | Female | 12.4 | 3.7 | 2.4 |
| Black | Male | 41.6 | 9.6 | 6.8 |
| Black | Female | 31.7 | 9.0 | 4.5 |

Source: *Employment and Earnings*, December 2005, table A-13.

The overall unemployment rate is quite variable, as shown in figure 1 on p. 396. But the pattern of rates in table 2 is quite stable in certain respects. For instance, the unemployment rate for blacks has been roughly double the unemployment rate for whites over the period 1961–2005. One development is worth noting. From the 1990s onwards, unemployment rates for men have become higher than those for women—a change from the past.

Unemployment numbers are published for much finer classifications than the ones shown in table 2, including marital status, race, age, sex, type of last job, reason for unemployment (for instance, fired or quit), and duration of unemployment. The Bureau starts with a huge sample. But by the time it comes down to the white men age 35–44 who quit managerial jobs, have been out of work for 5 to 14 weeks, and are looking for work by reading the newspapers, there might not be

Figure 3. Table A-32, *Employment and Earnings*, December 2005.

HOUSEHOLD DATA
NOT SEASONALLY ADJUSTED

**A-32. Unemployed persons by reason for unemployment, sex, and age**

(Numbers in thousands)

| Reason | Total, 16 years and over | | Men, 20 years and over | | Women, 20 years and over | | Both sexes, 16 to 19 years | |
|---|---|---|---|---|---|---|---|---|
| | Dec. 2004 | Dec. 2005 | Dec. 2004 | Dec. 2005 | Dec. 2004 | Dec. 2005 | Dec. 2004 | Dec. 2005 |
| **NUMBER OF UNEMPLOYED** | | | | | | | | |
| Total unemployed | 7,599 | 6,956 | 3,727 | 3,355 | 2,802 | 2,707 | 1,070 | 894 |
| Job losers and persons who completed temporary jobs | 4,166 | 3,622 | 2,573 | 2,212 | 1,433 | 1,281 | 160 | 129 |
| On temporary layoff | 1,040 | 1,013 | 709 | 679 | 255 | 287 | 77 | 47 |
| Not on temporary layoff | 3,126 | 2,609 | 1,864 | 1,534 | 1,178 | 993 | 84 | 82 |
| Permanent job losers | 2,272 | 1,866 | 1,302 | 1,072 | 908 | 743 | 61 | 51 |
| Persons who completed temporary jobs | 854 | 743 | 562 | 461 | 270 | 250 | 23 | 31 |
| Job leavers | 845 | 752 | 398 | 339 | 369 | 335 | 78 | 78 |
| Reentrants | 2,040 | 2,083 | 683 | 722 | 894 | 1,000 | 462 | 361 |
| New entrants | 548 | 499 | 74 | 82 | 105 | 91 | 369 | 325 |
| **PERCENT DISTRIBUTION** | | | | | | | | |
| Total unemployed | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Job losers and persons who completed temporary jobs | 54.8 | 52.1 | 69.0 | 65.9 | 51.2 | 47.3 | 15.0 | 14.4 |
| On temporary layoff | 13.7 | 14.6 | 19.0 | 20.2 | 9.1 | 10.6 | 7.2 | 5.3 |
| Not on temporary layoff | 41.1 | 37.5 | 50.0 | 45.7 | 42.1 | 36.7 | 7.8 | 9.2 |
| Job leavers | 11.1 | 10.8 | 10.7 | 10.1 | 13.2 | 12.4 | 7.3 | 8.6 |
| Reentrants | 26.8 | 30.0 | 18.3 | 21.5 | 31.9 | 36.9 | 43.2 | 40.4 |
| New entrants | 7.2 | 7.2 | 2.0 | 2.4 | 3.6 | 3.4 | 34.5 | 36.4 |
| **UNEMPLOYED AS A PERCENT OF THE CIVILIAN LABOR FORCE** | | | | | | | | |
| Job losers and persons who completed temporary jobs | 2.8 | 2.4 | 3.4 | 2.9 | 2.2 | 1.9 | 2.3 | 1.9 |
| Job leavers | .6 | .5 | .5 | .4 | .6 | .5 | 1.1 | 1.1 |
| Reentrants | 1.4 | 1.4 | .9 | .9 | 1.4 | 1.5 | 6.7 | 5.3 |
| New entrants | .4 | .3 | .1 | .1 | .2 | .1 | 5.4 | 4.8 |

NOTE: Beginning in January 2005, data reflect revised population controls used in the household survey.

too many cases left. Figure 3 shows estimates by reason for unemployment, sex, and age. (*Employment and Earnings* goes into much more detail.)

In general, by the time a large sample is cross-tabulated, there will be only very small subsamples in some classifications. Inferences about the corresponding subpopulations would be quite uncertain. Now, suppose that each estimate is within 1% of its true value with probability 95%, say. With a thousand estimates (which is about the number in *Employment and Earnings*), it would not be surprising if a few of them are quite a bit more than 1% off. The Bureau takes a big sample because it has to make many estimates about many subpopulations, and it wants to be reasonably confident that they are all fairly accurate. In fact, the Bureau will not make estimates when a subsample drops below a threshold size of about 50 cases.

*"The dip in sales seems to coincide with the decision to eliminate the sales staff."*

Drawing by Leo Cullum; © 2006 The New Yorker Magazine, Inc.

## 4. WEIGHTING THE SAMPLE

Suppose that one month, in the Bureau's sample of 110,000 people, there are 3,836 who are unemployed. The Bureau is sampling 1 person in 2,000 from the civilian non-institutional population age 16 and over. So it is natural to think that each person in the sample represents 2,000 people in the country. Then the way to estimate the total number of unemployed in the population is to weight up the sample number of 3,836 by the factor of 2,000:

$$2{,}000 \times 3{,}836 = 7{,}672{,}000$$

However, the Bureau does not do anything that simple. Not everybody in the sample gets the same weight. Instead, the Bureau divides the sample up into groups (by age, sex, race, and area of residence) and weights each group up separately.

There is a good reason for all the complexity. The sampling rate is different from one stratum to another, and the weights have to compensate; otherwise, the estimates could be quite biased. Moreover, the weights are used to control the impact of chance variation. For example, suppose there are too many white males age 16–19 in the sample, relative to their share in the total population. Unemployment is high in this group, which would make the overall unemployment rate in the sample too high. The Bureau has a fix: any group which is over-represented in the sample gets proportionately smaller weights, bringing the sample back into line with the population. On the other hand, if a group is under-represented, the weights are increased. Adjusting the weights this way helps to correct imbalances caused by chance variation. That reduces sampling error.[8]

## 5. STANDARD ERRORS

In estimating the unemployment rate, precision counts. For example, a definite picture of the economy is given by saying that the unemployment rate is 7.0% ± 0.1 of 1%. However, 7% ± 3% covers everything from boom to bust. So, it is important to know how good the estimates really are. Procedures we have discussed in previous chapters do not apply, because the Bureau is not using a simple random sample. In particular, at the second stage of its sampling procedure the Bureau chooses some ultimate sampling units (USUs). A USU is a *cluster* of about four adjacent housing units. Every person age 16 and over living in one of these USUs gets into the sample (section 2). A cluster is all or nothing: either everybody in the cluster gets into the sample, or nobody does. People living in the same cluster tend to be similar to one another in many ways. Information about each one says something about all the others, in terms of family background, educational history, and employment status.

With simple random sampling, by comparison, if one person in the cluster gets into the sample, the other people still have only a small chance of getting in. As a result, each person drawn into a simple random sample provides additional information, independent of the persons drawn previously. The Bureau's cluster sample of 110,000 persons contains less information than a simple random sample of the same size: cluster samples involve a lot of redundancy. Thus, clustering tends to reduce the precision of the Bureau's estimates. On the other hand, the weights improve precision. All in all, computing SEs for the Bureau's estimates is a delicate business.

As it turns out, with a cluster sample the standard errors can themselves be estimated very closely from the data, using the *half-sample method*. Although the details are complicated and take a lot of computer power, the idea is simple. If the Bureau wanted to see how accurate the Current Population Survey was, one thing to do would be make another independent survey following exactly the same procedures. The difference between the two surveys would give some idea of how reliable each set of results was.

Nobody would seriously propose to replicate the Current Population Survey, at a cost of another $60 million a year, just to see how reliable it is. But the Bureau can get almost the same effect by splitting the Survey into two independent pieces which have the same chance behavior (hence the name, "half-sample method"). Suppose for instance that one piece of the survey estimates the civilian labor force at 150.5 million, and the other comes in at 150.7 million. This difference is due to chance error. The pooled estimate of the civilian labor force is

$$\frac{150.5 + 150.7}{2} = 150.6 \text{ million}$$

The two individual estimates are 0.1 million away from their average, and the standard error is estimated by this difference of 0.1 million.

Of course, an estimated standard error based on only one split may not be too reliable. But there are many different ways to split the sample. The Bureau looks at a number of them and combines the standard errors by taking the root-mean-square. This completes the outline of the half-sample method.[9] Some of the estimated standard errors for November 2005 are shown in table 3.

Table 3. Estimated standard errors, November 2005.

| | *Estimate* | *Standard error* |
|---|---|---|
| Civilian labor force | 150.24 million | 300,000 |
| Employment | 142.97 million | 323,000 |
| Unemployment | 7.27 million | 155,000 |
| Unemployment rate | 4.8% | 0.1 of 1% |

Source: *Employment and Earnings*, December 2005, tables A13, 1B, and 1C.

How do the estimated standard errors in table 3 compare to those for a simple random sample of the same size and composition? Calculations show that for estimating the size of the labor force, the Bureau's standard error is about 8% smaller than that for a simple random sample: the weights are doing a good job. For estimating the number of unemployed, however, the Bureau's sample is about 30% worse than a simple random sample: the clustering hurts.[10]

So why doesn't the Bureau use simple random sampling? For one thing, there is no list showing all the people age 16 and over in the U.S., with current addresses. Even if there were such a list, taking a simple random sample from it would produce people spread thinly throughout the country, and the cost of interviewing them would be enormous. With the Bureau's procedure, the sample is bound to come out in clumps in relatively small and well-defined areas, so the interviewing cost is quite manageable. In 2005, this was about $100 per interview. The Bureau's sample design turns out to be amazingly cost effective.

The comparison between the Bureau's design and a simple random sample points to a real issue. To compute a standard error properly, you need more than the sample data. You need to know how the sample was picked. With a simple random sample, there is one SE. With a cluster sample, there is another. The formulas which apply to simple random samples will usually underestimate the standard errors in cluster samples. (These issues came up before, in the context of the Gallup Poll: sections 4 and 5 of chapter 21.)

> Cluster samples are less informative than simple random samples of the same size. So the simple random sample formulas for the standard error do not apply.

## Exercise Set A

1. One month, the Current Population Survey sample amounted to 100,000 people. Of them, 62,000 were employed, and 3,000 were unemployed. True or false, and explain:
   (a) 65% of the sample was in the labor force.
   (b) The Bureau would estimate that 65% of the population was in the labor force.

2. The Current Population Survey sample is split into two independent halves. From one half, the number of employed persons is estimated as 151.5 million; from the other, it is estimated as 151.3 million. Combine these two estimates, and attach a standard error to the result.

3. (Hypothetical.) The Health Department in a certain city takes a simple random sample of 100 households. In 80 of these households, all the occupants have been vaccinated against polio. So the Department estimates that for 80% of the households in that city, all the occupants have been vaccinated against polio. With this information, can you put a standard error on the 80%? Find the SE, or explain what other information is needed.

4. (Continues exercise 3.) The Department interviews every person age 25 and over in the sample households. They find 144 such persons, of whom 29 have college degrees. They estimate that 20.1% of the people age 25 and over in the city have college degrees. With this information, can you put a standard error on the 20.1%? Find the SE, or explain what other information is needed.

5. In election years, the Bureau makes a special report on voting, using the Current Population Survey sample. In 2000, about 55% of all the people of voting age in the sample said they voted; but only 52% of the total population of voting age did in fact vote.[11] Can the difference be explained as a chance error? If not, how else can it be explained? (You may assume that the Bureau's sample is the equivalent of a simple random sample of 75,000 people.)

6. In table 2 on p. 399, which estimate is more trustworthy: for white males age 20–64, or black males age 20–64? Explain briefly.

*The answers to these exercises are on pp. A84–85.*

## 6. THE QUALITY OF THE DATA

The data collected by the Survey are of very high quality: for many purposes, Survey data are considered to be more accurate than Census data. In any large-scale field operation, mistakes are inevitable. Since the Survey operates on a much smaller scale than the Census, it can afford better quality control. The key is careful selection, training, and supervision of the field staff. Interviewers are given about four days of training in survey procedures before they start work, and several hours a month of training while they are on the job. At least once a year, their work is observed by their supervisors. In addition, about 3% of the monthly sample (chosen by a separate probability sampling procedure) is reinterviewed by supervisors. All discrepancies are discussed with the interviewers. The interviewers' reports are *edited*, that is, checked for incomplete or inconsistent entries. For most items, error rates are low; and each error is reviewed with the person who made it.

## 7. BIAS

Bias is more insidious than chance error, especially if it operates more or less evenly across the sample. SEs computed by the half-sample method—or any other method—will not pick up that kind of bias. Measuring bias, even roughly, is hard work and involves going beyond the sample data.

> When bias operates more or less evenly across the sample, it cannot be detected just by looking at the data.

The Bureau has made unusually careful studies of the biases in the Current Population Survey. On the whole, these seem to be minor, although their exact sizes are not known. To begin with, the Survey design is based on Census data (section 2), and the Census misses a small percentage of the population. This percentage is not easy to pin down. Even if the Bureau knew it, they would still have a hard time adjusting the estimated number of unemployed (say) to compensate for the undercount, because the people missed by the Census are likely to be different from the ones the Census finds. A similar difficulty crops up in another place. The Survey misses about 10% of the people counted by the Census. To some extent, the weights bring these missing people back into the estimates. But non-response bias is not so easy to fix. The people missed by the Survey are probably different from the ones it finds, while the weights pretend they are the same.[12]

Next, the distinction between "employed" and "unemployed" is a little fuzzy around the edges. For example, people who have a part-time job but would like full-time work are classified as employed, but they really are partially unemployed. Moreover, people who want to work but have given up looking are classified as outside the labor force, although they probably should be classified as unemployed. The Bureau's criterion for unemployment, namely being without work, available for work, and looking for work, is necessarily subjective. In practice, it is a bit slippery. Results from the reinterview program (section 6) suggest the number of unemployed is higher than the Bureau's estimate, by several hundred thousand people. In this case, the bias is larger than the sampling error.[13] Over the period from 1980 to 2005, the number of unemployed has ranged from 5 to 10 million. Relatively speaking, both sampling error and non-sampling error are small.

## 8. REVIEW EXERCISES

*Review exercises may cover previous chapters as well.*

1. One month, there are 100,000 people in the Current Population Survey sample, of whom 63,000 are employed and 4,000 are unemployed.

   (a) True or false, and explain: the Bureau would estimate the percentage of the population who are unemployed as

   $$\frac{4{,}000}{63{,}000 + 4{,}000} \times 100\% \approx 6\%$$

   (b) What happened to the other 33,000 people?

2. One month, there are 100,000 people in the Current Population Survey sample, and the Bureau estimates the unemployment rate as 6.0%. True or false, and explain: the standard error for this percentage should be estimated as follows—

$$\text{SE for number} = \sqrt{100{,}000} \times \sqrt{0.06 \times 0.94} \approx 75$$

$$\text{SE for percent} = \frac{75}{100{,}000} \times 100\% \approx 0.08 \text{ of } 1\%$$

3. One month, the Current Population Survey sample is split into two independent replicates. Using one replicate, the number of unemployed people is estimated as 7.1 million. The other replicate produces an estimate of 6.9 million. Using this information, estimate the number of unemployed people, and attach a standard error to the estimate.

4. Using the data in exercise 3, what can you say about the bias in the estimate?

5. A simple random sample is drawn at random _____ replacement. Options: with, without.

6. A box contains 250 tickets. Two people want to estimate the percentage of 1's in the box. They agree to use the percentage of 1's in 100 draws made at random from the box. Person A wants to draw with replacement; person B wants to draw without replacement. Which procedure gives a more accurate estimate? Or does it make any difference?

7. (Hypothetical.) A survey organization draws a sample of 100 households from 10,000 in a certain town, by the following procedure. First, they divide the town into 5 districts, with 2,000 households each. Then they draw 2 districts at random. Within each of the 2 selected districts, they draw 50 households at random.

   (a) Is this a probability sample?
   (b) Is this a simple random sample?

   Answer yes or no, and explain.

8. A supermarket chain has to value its inventory at the end of every year, and this is done on a sample basis. There is a master list of all the types of items sold in the stores. Then, auditors take a sample of the items and go through the shelves, finding the amounts in stock and prices for the sample items. To draw the sample, the auditors start by choosing a number at random from 1 to 100. Suppose this turns out to be 17. The auditors take the 17th, 117th, 217th, ... items in the list for the sample. If the random number is 68, they take the 68th, 168th, 268th, ... items. And so forth.

   (a) Is this a probability sample?
   (b) Is this a simple random sample?

   Answer yes or no, and explain.

9. As part of a study on drinking, the attitudes of a sample of alcoholics are assessed by interview.[14] Cases are assigned to interviewers at random. Some of the interviewers are teetotalers, others drink. Would you expect the two groups of interviewers to reach similar conclusions? Answer yes or no, and give reasons.

10. From "The Grab Bag" by L. M. Boyd in the *San Francisco Chronicle*: "The Law of Averages says that if you throw a pair of dice 100 times, the numbers tossed will add up to just about 683." Is this right? Answer yes or no, and explain.

11. A polling organization takes a simple random sample of 750 voters from a district with 18,000 voters. In the sample, 405 voters are for. Fill in the blanks, using the options below. Explain briefly,

    (a) The observed value of the _____ is 405.
    (b) The observed value of the _____ is 54%.
    (c) The expected value of the _____ is equal to the _____.

    Options:

    (i) number of voters in the sample who are for
    (ii) percentage of voters in the sample who are for
    (iii) percentage of voters in the district who are for

12. In 2004, there were 318,390 applications to buy guns in California.[15] A criminologist takes a simple random sample of 193 out of these applications, and finds that only 2 were rejected. True or false, and explain:

    (a) 2 out of 193 is 1.04%.
    (b) The SE on the 1.04% is 0.73%.
    (c) A 95%-confidence interval for the percentage of all 318,390 applications that were rejected is 1.04% ± 1.46%.

## 9. SUMMARY

1. Unemployment rates in the U.S. are estimated using the *Current Population Survey*.

2. This survey is based on a nationwide probability sample of about 110,000 persons, who are interviewed monthly. The design is more complicated than simple random sampling.

3. The Survey reweights the sample so it agrees with Census data on age, sex, race, state of residence, and certain other characteristics influencing employment status.

4. When a sample is taken by a probability method, it is possible not only to estimate parameters, but also to figure the likely size of the chance errors in the estimates.

5. The standard errors for *cluster samples* can be obtained by the *half-sample method*, splitting the sample into two halves and seeing how well they agree.

6. The formulas for the standard error have to take into account the details of the probability method used to draw the sample. The formulas which apply to simple random samples will usually underestimate the standard errors in cluster samples.

7. When bias operates more or less evenly across the sample, it cannot be detected just by looking at the sample data. Standard errors ignore that kind of bias.

8. The Current Population Survey, like all surveys, is subject to a number of small biases. The bias in the estimate of the unemployment rate is thought to be larger than the standard error.

# 23

# The Accuracy of Averages

*Ranges are for cattle.*
—L.B.J.

## 1. INTRODUCTION

The object of this chapter is to estimate the accuracy of an average computed from a simple random sample. This section deals with a preliminary question: How much chance variability is there in the average of numbers drawn from a box? For instance, take the box

$$\boxed{1\ 2\ 3\ 4\ 5\ 6\ 7}$$

The computer was programmed to make 25 draws at random with replacement from this box:

24325 75645 44124 46472 72573

The sum of these numbers is 105, so their average is $105/25 = 4.2$. The computer did the experiment again, and the results came out differently:

51434 52177 12324 71653 66334

Now the sum is 95, so the average is $95/25 = 3.8$. The sum of the draws is subject to chance variability, therefore the average is too. The new problem is to calculate the expected value and standard error for the average of the draws. The method will be indicated by example.

*Example 1.* Twenty-five draws will be made at random with replacement from the box

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

The average of the draws will be around _____, give or take _____ or so.

*Solution.* The average of the box is 4, so the average of the draws will be around 4. The give-or-take number is the SE. To get the SE for the average, we go back to the sum. The expected value for the sum is

$$\text{number of draws} \times \text{average of box} = 25 \times 4 = 100$$

The SD of the box is 2, and the SE for the sum is

$$\sqrt{\text{number of draws}} \times \text{SD of box} = \sqrt{25} \times 2 = 10$$

The sum will be around 100, give or take 10 or so.

What does this say about the average of the draws? If the sum is one SE above expected value, or 100 + 10, the average of the 25 draws is

$$\frac{100+10}{25} = \frac{100}{25} + \frac{10}{25} = 4 + 0.4$$

On the other hand, if the sum is one SE below expected value, or 100 − 10, the average is

$$\frac{100-10}{25} = \frac{100}{25} - \frac{10}{25} = 4 - 0.4$$

The average of the draws will be about 4, give or take 0.4 or so. The 4 is the expected value for the average of the draws. The 0.4 is the standard error, completing the solution.

The idea in brief:

$$\text{sum of 25 draws} = 100 \pm 10 \text{ or so}$$

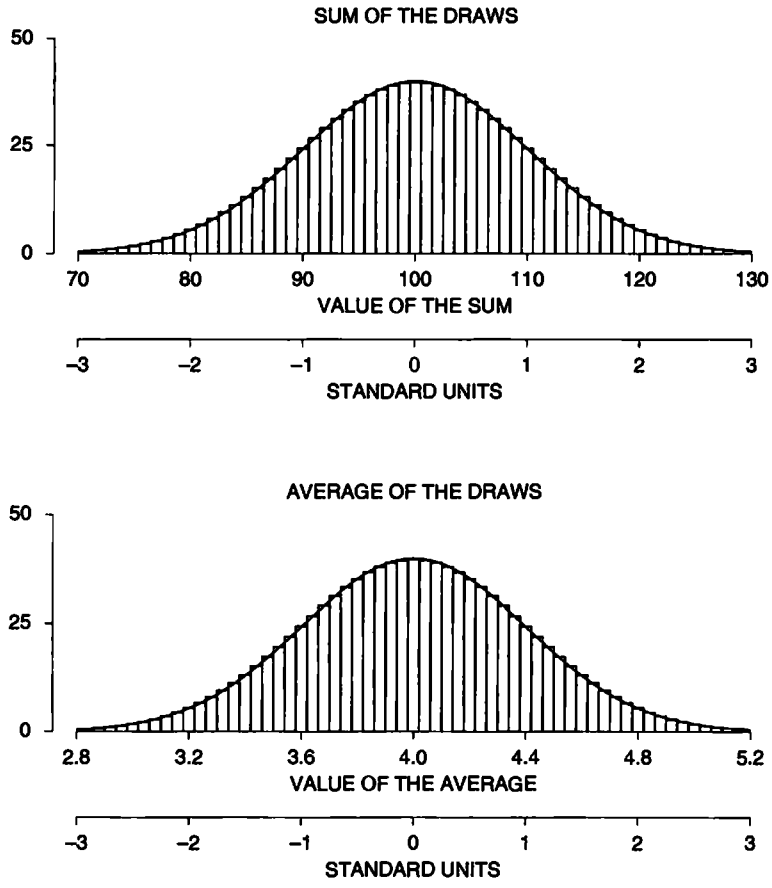
$$\text{average of 25 draws} = \frac{100}{25} \pm \frac{10}{25} \text{ or so}$$

In other words, to find the SE for the average of the draws, just go back and get the SE for the sum; then divide by the number of draws.

> When drawing at random from a box:
>
> $$\text{EV for average of draws} = \text{average of box.}$$
>
> $$\text{SE for average of draws} = \frac{\text{SE for sum}}{\text{number of draws}}.$$

The SE for the average says how far the average of the draws is likely to be from the average of the box.

Figure 1. The top panel shows a probability histogram for the sum of 25 draws from the box $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}\boxed{6}\boxed{7}$. The bottom panel shows the probability histogram for the average of the draws. In standard units, the two histograms are exactly the same.




If the number of draws is large enough, the normal curve can be used to figure chances for the average. Figure 1 (bottom panel) shows the probability histogram for the average of 25 draws from the box

$$\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}\boxed{6}\boxed{7}$$

The histogram follows the curve, so areas under the histogram can be approximated by areas under the curve.

Why does the probability histogram for the average look like the normal curve? This is a corollary of the mathematics of chapter 18. The probability histogram for the sum of the 25 draws is close to the normal curve (top panel of figure 1). The average of the draws equals their sum, divided by 25. This division is just a change of scale, and washes out in standard units. The two histograms in figure 1 have exactly the same shape, and both follow the curve.

> When drawing at random from a box, the probability histogram for the average of the draws follows the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be reasonably large.[1]

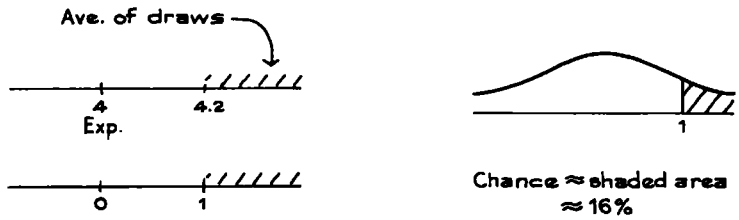
*Example 2.* One hundred draws will be made at random with replacement from the box in example 1.

(a) The average of the draws will be around _____, give or take _____ or so.

(b) Estimate the chance that the average of the draws will be more than 4.2.

*Solution.* As in example 1, the sum of the draws will be around $100 \times 4 = 400$. The give-or-take number is $\sqrt{100} \times 2 = 20$. The sum of the draws will be around 400, give or take 20 or so. The average of the draws will be around $400/100 = 4$, give or take $20/100 = 0.2$ or so. The SE for the average of 100 draws is 0.2.

Part (b) is handled by the normal approximation.




The chance is around 16%. This completes the solution.

In examples 1 and 2, when the number of draws went up by a factor of 4, from 25 to 100, the SE for the average of the draws went down by a factor of $\sqrt{4} = 2$, from 0.4 to 0.2. This is so in general.

> When drawing at random with replacement from a box of tickets, multiplying the number of draws by a factor (like 4) divides the SE for the average of the draws by the square root of that factor ($\sqrt{4} = 2$).

As the number of draws goes up, the SE for the sum gets bigger—and the SE for the average gets smaller. Here is the reason. The SE for the sum goes up, but only by the square root of the number of draws. As a result, while the SE for the sum gets bigger in absolute terms, compared to the number of draws it gets smaller. The division by the number of draws makes the SE for the average go down. Keep this difference between the two SEs in mind.

When drawing without replacement, the exact SE for the average of the draws can be found using the correction factor (section 4 of chapter 20)—

$$\text{SE without} = (\text{correction factor}) \times (\text{SE with}).$$

Usually, the number of draws is small by comparison with the number of tickets in the box, and the correction factor will be so close to 1 that it can be ignored.

## Exercise Set A

1. One hundred draws are made at random with replacement from a box.
   (a) If the sum of the draws is 7,611, what is their average?
   (b) If the average of the draws is 73.94, what is their sum?

2. A box of tickets averages out to 75, and the SD is 10. One hundred draws are made at random with replacement from this box.
   (a) Find the chance (approximately) that the average of the draws will be in the range 65 to 85.
   (b) Repeat, for the range 74 to 76.

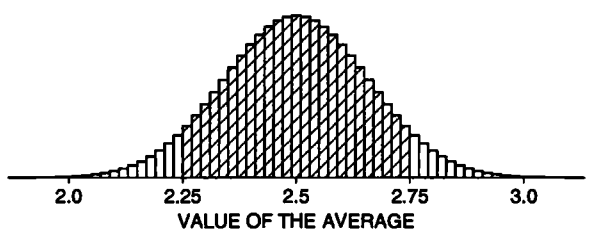
3. One hundred draws will be made at random with replacement from a box of tickets. The average of the numbers in the box is 200. The SE for the average of the draws is computed, and turns out to be 10. True or false:
   (a) About 68% of the tickets in the box are in the range 190 to 210.
   (b) There is about a 68% chance for the average of the hundred draws to be in the range 190 to 210.

4. You are drawing at random with replacement from a box of numbered tickets.
   (a) The expected value for the average of the _____ equals the average of the _____. Options: box, draws.
   (b) As the number of draws goes up, the SE for the _____ of the draws goes up but the SE for the _____ of the draws goes down. Options:

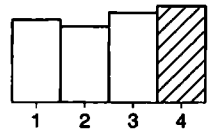
   sum average

5. A box contains 10,000 tickets. The numbers on these tickets average out to 50, and the SD is 20.
   (a) One hundred tickets are drawn at random with replacement. The average of these draws will be around _____, give or take _____ or so.
   (b) What if 100 draws are made without replacement?
   (c) What if 100 draws are made without replacement, and there are only 100 tickets in the box?

6. The figure below shows the probability histogram for the average of 50 draws from the box $\boxed{1}\boxed{2}\boxed{3}\boxed{4}$. What does the shaded area represent?



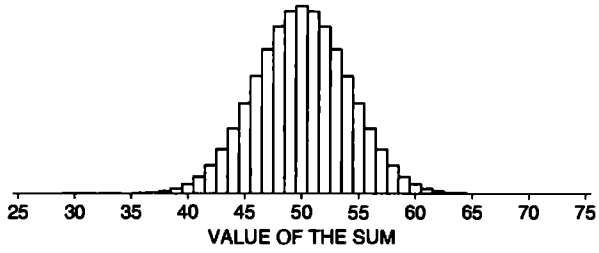
7. The figure below shows a histogram for data generated by drawing 50 times from the box in exercise 6. What does the shaded area represent?




8. (a) In the top panel of figure 1, the area of the rectangle over 90 represents what?
   (b) In the bottom panel of figure 1, the area of the rectangle over 3.6 represents what?
   (c) The rectangles in parts (a) and (b) have exactly the same area. Is that a coincidence? Discuss briefly.

9. Two hundred draws are made at random with replacement from $\boxed{1}\boxed{2}\boxed{2}\boxed{3}$. True or false, and explain:
   (a) The expected value for the average of the draws is exactly 2.
   (b) The expected value for the average of the draws is around 2, give or take 0.05 or so.
   (c) The average of the draws will be around 2, give or take 0.05 or so.
   (d) The average of the draws will be exactly 2.
   (e) The average of the box is exactly 2.
   (f) The average of the box is around 2, give or take 0.05 or so.

10. The figure below is a probability histogram for the sum of 25 draws from the box $\boxed{1}\boxed{2}\boxed{3}$. However, an investigator needs the probability histogram for the average of these draws, by midnight. A research assistant says, "There's nothing to it. All we have to do is change the numbers on the horizontal axis." Is that right? If so, the assistant should change 25 to _____, 50 to _____, and 55 to _____. If the assistant is wrong, what needs to be done? Explain your answers. (No vertical scale is needed.)




*The answers to these exercises are on p. A85.*

*Technical notes.* (i) The bottom panel in figure 1 represents what is called a *sampling distribution*. The histogram shows how the sample averages vary over the set of all possible samples. In more detail, imagine making a list of all possible samples, and computing the sample average for each one. (You would get quite a long list of averages.) Some averages come up more frequently than others. The area of the rectangle over 4.0 shows what percentage of these sample averages are 4.0, and so forth.

(ii) When drawing at random with replacement from a box, the SE for the sum of the draws is

$$\sqrt{\text{no. of draws}} \times \text{SD of box.}$$

So the SE for the average of the draws is

$$(\sqrt{\text{no. of draws}} \times \text{SD of box})/\text{no. of draws.}$$

This simplifies to (SD of box)/$\sqrt{\text{no. of draws}}$, which in most books is written $\sigma/\sqrt{n}$, where $\sigma$ is the SD and $n$ is the number of draws. The Greek letter $\sigma$ is read as "sigma."

## 2. THE SAMPLE AVERAGE

In section 1, the numbers in the box were known, and the problem was to say something about the average of the draws. This section reasons in the opposite—and more practical—direction. A random sample is taken from a box of unknown composition, and the problem is to estimate the average of the box. Naturally, the average of the draws is used as the estimate. And the SE for the sample average can be used with the normal curve to gauge the accuracy of the estimate. (Chapter 21 used the same technique for percentages.)

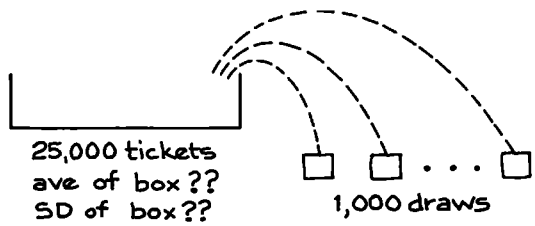
The method will be presented by example. Along the way, there will be two questions to answer:

- What's the difference between the SD of the sample and the SE for the sample average?
- Why is it OK to use the normal curve in figuring confidence levels?

Now, the example. Suppose that a city manager wants to know the average income of the 25,000 families living in his town. He hires a survey organization to take a simple random sample of 1,000 families. The total income of the 1,000 sample families turns out to be \$62,396,714. Their average income is \$62,396,714/1,000 ≈ \$62,400. The average income for all 25,000 families is estimated as \$62,400. Of course, this estimate is off by a chance error. The problem is to put a give-or-take number on the estimate:

$$\$62{,}400 \pm \$_____?$$

The SE is needed, and for that, a box model. There should be one ticket in the box for each family in the town, showing that family's income. The data are like 1,000 draws from the box.



The average income of the sample families is like the average of the draws. The SE for the average of the draws can now be found by the method of section 1. The first step is to find the SE for the sum of the draws. Since 1,000 is such a small fraction of 25,000, there is no real difference between drawing with and without replacement. The SE for the sum is

$$\sqrt{1{,}000} \times \text{SD of box.}$$

Of course, the survey organization does not know the SD of the box, but they can estimate it by the SD of the sample. (This is another example of the bootstrap method discussed in section 1 of chapter 21.)

> With a simple random sample, the SD of the sample can be used to estimate the SD of the box. The estimate is good when the sample is large.

There are 1,000 families in the sample, and the SD of their incomes turns out to be \$53,000. The SD of the box is estimated as \$53,000. The SE for the sum is estimated as

$$\sqrt{1{,}000} \times \$53{,}000 \approx \$1{,}700{,}000.$$

To get the SE for the average, we divide by the number of families in the sample: \$1,700,000/1,000 = \$1,700. That is the answer. The average of the draws is something like \$1,700 off the average of the box. So the average of the incomes of all 25,000 families in the town can be estimated as

$$\$62{,}400 \pm \$1{,}700.$$

Keep the interpretation of the \$1,700 in mind: it is the margin of error for the estimate. This completes the example.

One point is worth more discussion. The expected value for the sum of the draws—the total income of the sample families—is

$$1{,}000 \times \text{average of the box.}$$

This is unknown because the average of the box is unknown. The total income of the 1,000 sample families turned out to be \$62,396,714. This is the observed value for the sum of the draws. The SE for the sum—\$1,700,000—measures the likely size of the difference between \$62,396,714 and the expected value. In general,

$$\text{observed value} = \text{expected value} + \text{chance error.}$$

The SE measures the likely size of the chance error.

Confidence intervals for percentages (qualitative data) were discussed in section 2 of chapter 21. The same idea can be used to get confidence intervals for the average of the box (quantitative data). For example, a 95%-confidence interval for the average of the incomes of all 25,000 families in the town is obtained by going

2 SEs either way from the sample average:

$$\$62,400 \pm 2 \times \$1,700 = \$59,000 \text{ to } \$65,800.$$

("Sample average" is statistical shorthand for the average of the numbers in the sample.)

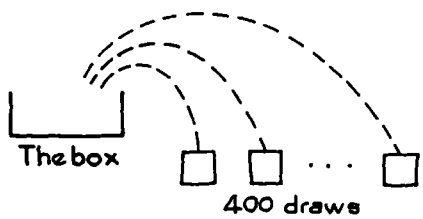
Two different numbers came up in the calculations: the SD of the sample was \$53,000, and the SE for the sample average was \$1,700. These two numbers do different things.

- The SD says how far family incomes are from average—for typical families.
- The SE says how far sample averages are from the population average—for typical samples.

People who confuse the SD with the SE might think that somehow, 95% of the families in the town had incomes in the range \$62,400 ± \$3,400. That would be ridiculous. The range \$62,400 ± \$3,400 covers only a tiny part of the income distribution: the SD is about \$53,000. The confidence interval is for something else. In about 95% of all samples, if you go 2 SEs either way from the sample average, your confidence interval will cover the average for the whole town; in the other 5%, your interval will miss. The word "confidence" is to remind you that the chances are in the sampling procedure; the average of the box is not moving around. (These issues were discussed before, in section 3 of chapter 21.)

*Example 3.* As part of an opinion survey, a simple random sample of 400 persons age 25 and over is taken in a certain town in Appalachia. The total years of schooling completed by the sample persons is 4,635. So their average educational level is $4,635/400 \approx 11.6$ years. The SD of the sample is 4.1 years. Find a 95%-confidence interval for the average educational level of all persons age 25 and over in this town.

*Solution.* First, a box model. There should be one ticket in the box for each person age 25 and over in the town, showing the number of years of schooling completed by that person; 400 draws are made at random from the box. The data are like the draws, and the sample average is like the average of the draws. That completes the model.




We need to compute the SE for the average of the draws. The SE for the sum is $\sqrt{400} \times$ SD of the box. The SD of the box is unknown, but can be estimated by

the SD of the sample, as 4.1 years. So the SE for the sum of the draws is estimated as $\sqrt{400} \times 4.1 = 82$ years. (The 82 measures the likely size of the chance error in the sum, which was 4,635.) The SE for the average is $82/400 \approx 0.2$ years. The average educational level of the persons in the sample will be off the average for the town by 0.2 years or so. An approximate 95%-confidence interval for the average educational level for all persons age 25 and over in the town is

$$11.6 \pm 0.4 \text{ years.}$$

That is the answer.

The confidence level of 95% is the area under the normal curve between $-2$ and 2. Why is the curve relevant? After all, the histogram for educational levels (p. 39) looks nothing like the curve. However, the curve is not used to approximate the histogram for the data; it is used to approximate the probability histogram for the sample average.

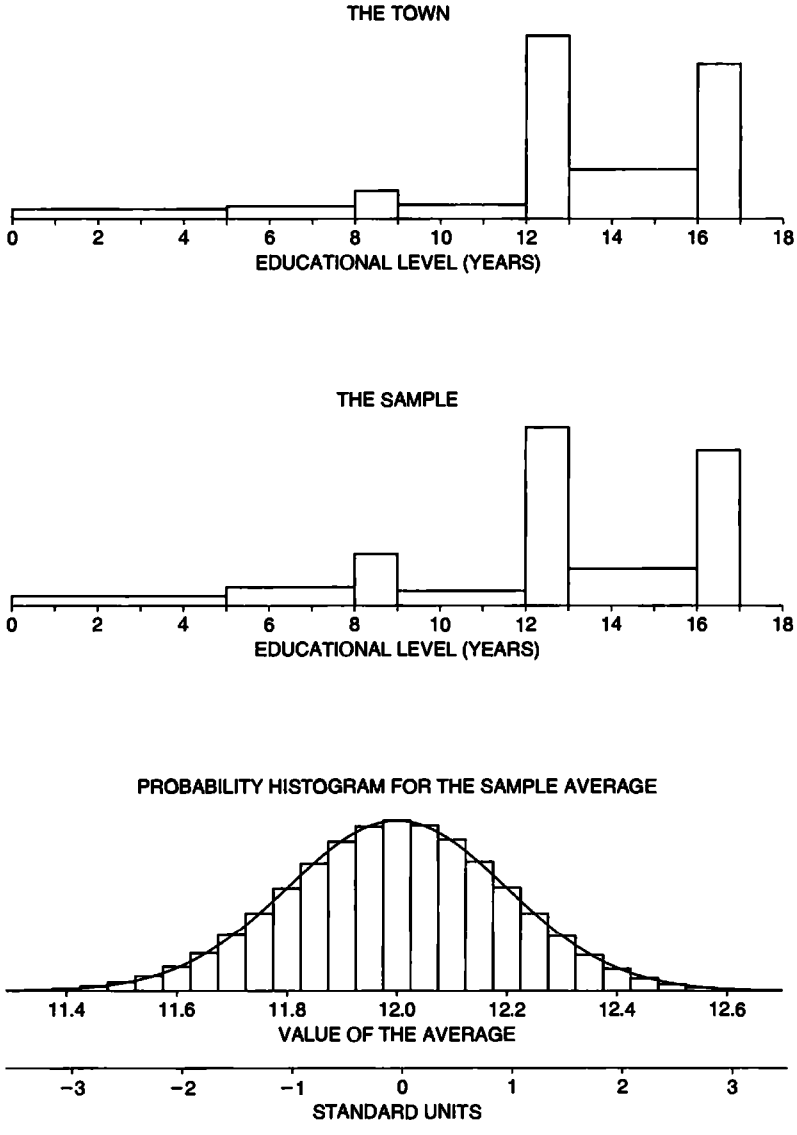
A computer simulation will help. The computer has one ticket in the box for each person age 25 or over in the town, showing his or her educational level. A histogram for the contents of the box is shown at the top of figure 2. This histogram represents the educational level of all people age 25 or over in the town. Its shape is nothing like the normal curve. (Remember, this is just a simulation; in reality, you would not know the contents of the box—but the mathematical theory can still be used.)

Now 400 draws must be made at random without replacement from the box, to get the sample. The computer was programmed to do this. A histogram for the 400 draws is shown in the second panel. This represents the distribution of educational level for the 400 sample people. It is very similar to the first histogram, although there are a few too many people with 8–9 years of education. That is a chance variation. Figure 2 indicates why the SD of the sample is a good estimate for the SD of the box. The two histograms show just about the same amount of spread.

So far, we have seen two histograms, both for data. Now a probability histogram comes in, for the average of the draws. This histogram is shown in the bottom panel. This third histogram does not represent data. Instead, it represents chances for the sample average. For instance, take the area under the probability histogram between 11.6 and 12.4 years. This area represents the chance that the average of 400 draws from the box will be between 11.6 and 12.4 years. The area works out to about 95%. For 95% of samples, the average educational level of the sample families will be in the range 11.6 to 12.4 years. For the other 5%, the sample average will be outside this range. Any area under the probability histogram can be interpreted in a similar way.

Now you can see why the normal approximation is legitimate. As the figure shows, the normal curve is a good approximation to the probability histogram for the average of the draws—even though the data do not follow the curve. That is why the curve can be used to figure confidence levels. Even with large samples, confidence levels read off the normal curve are only approximate, because they depend on the normal approximation; with a small sample, the normal curve should not be used (section 6 of chapter 26).

Figure 2. Computer simulation. The top panel shows the distribution of educational level among people age 25 or over in the whole town. The middle panel shows the distribution of educational level in the sample. These are histograms for data. The bottom panel shows the probability histogram for the average of 400 draws from the box; it is close to the normal curve. The average educational level in the town is 12.0 years, and the SD is 4.0 years; in the sample, the corresponding numbers are 11.6 and 4.1. (The endpoint convention for the data histograms: the class interval 12–13, for example, includes all the people who finished 12 years of schooling but not 13—high school graduates who did not finish a year of college.)



## Exercise Set B

1. Match each phrase on list A with one on list B.

| *List A* | *List B* |
|---|---|
| population | draws |
| population average | average of the box |
| sample | box |
| sample average | number of draws |
| sample size | average of the draws |

2. In each pair of phrases, one makes sense and one does not. Which is which? Explain briefly.
   (a) SE for box, SD of box.
   (b) SE for average of box, SE for average of draws.

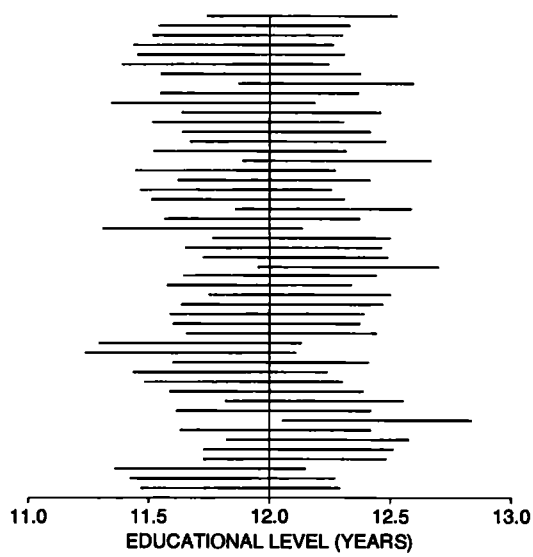
3. For the income example on pp. 415–417:
   (a) The SD of the box is _____ $53,000.
   (b) The SE for the sample average is _____ $1,700.
   (c) The _____ value for the sample average is $62,400.

   Fill in the blanks, using the options below, and explain. (At least one option will be left over.)

   | | |
   |---|---|
   | known to be | estimated from the sample as |
   | expected | observed |

4. In example 3 on p. 417, suppose 50 different survey organizations take simple random samples of 400 persons age 25 and over in the town. Each organization gets a 95%-confidence interval "sample average ± 2 SE." How many of these intervals should cover the population average?

5. The figure below is a computer simulation of the study described in exercise 4. The confidence intervals are plotted at different heights so they can be seen.



(a) Why do the intervals have different centers?

(b) Why do they have different lengths?

(c) How many of them cover the population average, marked by a vertical line at 12 years?

6. A university has 30,000 registered students. As part of a survey, 900 of these students are chosen at random. The average age of the sample students turns out to be 22.3 years, and the SD is 4.5 years.[2]

   (a) The average age of all 30,000 students is estimated as _____. This estimate is likely to be off by _____ or so.

   (b) Find a 95%-confidence interval for the average age of all 30,000 registered students.

7. The Census Bureau collects information on the housing stock as part of the decennial census. In 2000, for instance, the Bureau found that there were about 105 million occupied housing units in the U.S., one third being rental units. Typical rents varied from about $400 in Wyoming to about $800 in Hawaii.[3] (In some urban markets like east-side Manhattan or San Francisco's Nob Hill, of course, rents are much higher—if you can find an apartment in the first place.)

   A certain town has 10,000 occupied rental units. A local real estate office does a survey of these units: 250 are chosen at random, and the occupants are interviewed. Among other things, the rent paid in the previous month is determined. The 250 sample rents average out to $568, and the SD is $385. A histogram is plotted for the sample rents, and does not follow the normal curve.

   (a) If possible, find a 68%-confidence interval for the average rent paid in the previous month on all 10,000 occupied rental units in this town. If this is not possible, explain why not.

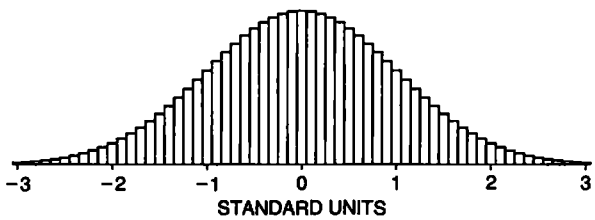
   (b) True or false, and explain: for about 68% of all the occupied rental units in this town, the rent paid in the previous month was between $544 and $592.

8. (Continues exercise 7; hard.) True or false, and explain: if another 250 occupied rental units were taken at random, there would be about a 68% chance for the new sample average to be in the range from $544 to $592.

9. (Hard.) Census data are available on 25,000 families in a certain town. For all 25,000 families, the average income is $61,700 and the SD is $50,000. A market research firm takes a simple random sample of 625 out of the 25,000 families. The figure below is a probability histogram for the average income of the sample families; the histogram is drawn in standard units. The average income of the 625 sample families turned out to be $58,700 and the SD was $49,000.

   (a) On the histogram below, +1 in standard units is _____. Options:

   $60,660 $63,700 $107,700 $111,700



9. Continued.

(b) In standard units, $58,700 is

0 −1.0 −1.5 other

Explain your answers.

*The answers to these exercises are on pp. A85–87.*

## 3. WHICH SE?

The SE always has the same interpretation: it is the likely size of a chance error. However, there seem to be many SEs. Which to use when? The best thing to do is to write down a box model, and decide what is being done to the draws. That will tell you which formula to use. There are four operations to think about: adding the draws, taking their average, classifying and counting, or taking percents. The corresponding formulas:

$$\text{SE for sum} = \sqrt{\text{number of draws}} \times \text{SD of box}$$

$$\text{SE for average} = \frac{\text{SE for sum}}{\text{number of draws}}$$

$$\text{SE for count} = \text{SE for sum, from a 0–1 box}$$

$$\text{SE for percent} = \frac{\text{SE for count}}{\text{number of draws}} \times 100\%$$

The SE for the sum is basic. The other formulas all come from that one. These formulas are exact for draws made at random with replacement from a box.

*Reasoning forward or backward.* When reasoning forward from the box to the draws, as in part V, the standard error can be computed exactly from the composition of the box. A chance quantity like the sum of the draws will be around its expected value—but will be off by an SE or so.

When reasoning backward from the draws to the box, you often have to estimate the SD of the box from the sample. So the SE itself is only approximate. However, the interpretation of the SE is almost the same. For instance, suppose the average of the sample is used to estimate the average of the box. This estimate will be off by a little, and the SE says by about how much. When the sample is reasonably large, the error in the SE itself is usually too small to matter.

> The SE shows the likely size of the amount off. It is a give-or-take number.

The terminology may be a bit confusing. Statisticians speak of the standard error for the *sample average*. The idea is that the sample average estimates the population average, but is off by a little. The SE gauges the likely size of the amount off. Statisticians also talk about confidence intervals for the *population*

*average*. The confidence interval is a range computed from the sample. This range covers the population average with some specified degree of confidence.

## Exercise Set C

*This exercise set also covers material from previous chapters.*

1. Fill in the table below, for draws made at random with replacement from the box $\boxed{0}\boxed{2}\boxed{3}\boxed{4}\boxed{6}$.

| *Number of draws* | *EV for sum of draws* | *SE for sum of draws* | *EV for average of draws* | *SE for average of draws* |
|---|---|---|---|---|
| 25 | | | | |
| 100 | | | | |
| 400 | | | | |

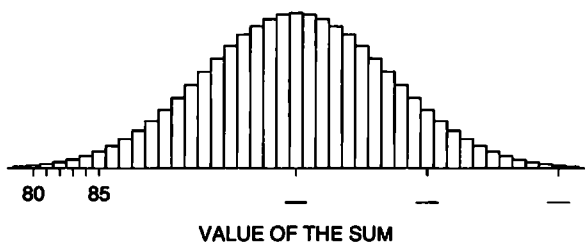
2. One hundred draws are made at random with replacement from a box. The average of the box is 3.1.
   (a) True or false: the expected value for the average of the draws is exactly equal to 3.1. If this cannot be determined from the information given, what else do you need to know, and why?
   (b) What is the SE for the average of the draws? If this cannot be determined from the information given, what else do you need to know, and why?

3. One hundred draws are made at random with replacement from a box. The average of the draws is 3.1.
   (a) The expected value for the average of the draws is _____ 3.1. Fill in the blank, using one of the options below, and explain.
      (i) exactly equal to
      (ii) estimated from the data as
   (b) What is the SE for the average of the draws? If this cannot be determined from the information given, what else do you need to know, and why?

4. Forty draws are made at random with replacement from the box

$$\boxed{1}\boxed{2}\boxed{3}\boxed{4}$$

   (a) Fill in the blanks with a word or phrase: the SE for the _____ is 7.1, and the SE for the _____ is 0.18. Explain your answers.
   (b) The figure below is a probability histogram for the sum of the draws. What numbers go into the three blanks?




5. What's the worst thing about a sample of size one?

6. There are three boxes of numbered tickets. The average of the numbers in each box is 200. However, the SD of box A is 10, the SD of box B is 20, and the SD of box C is 40. Now

   - 100 draws are made from box A,
   - 200 draws are made from box B,
   - 400 draws are made from box C.

   (The draws are made with replacement.) The average of each set of draws is computed. Here they are, in scrambled order:

   203.6 198.1 200.4

   (a) Which average comes from which box?
   (b) Could it possibly be otherwise?

   Explain briefly.

*The answers to these exercises are on p. A87.*

## 4. A REMINDER

This chapter explained how to evaluate the accuracy of an average computed from a simple random sample. Because the draws were made at random, it was possible to gauge the accuracy just from the spread in the data and the size of the sample. This is one of the major achievements of statistical theory.

The arithmetic can be carried out on any list: find the SD, multiply by the square root of the number of entries, then divide by the number of entries. However, the method gives sensible results only when the draws are made at random. If the data do not come from the right kind of sample, the result of the calculation may be nonsense (pp. 387–390, pp. 402–403).

A *sample of convenience* is a sample that is not chosen by a probability method. (An example would be some instructor's first-year psychology class.) Some people use the simple random sample formulas on samples of convenience. That could be a real blunder. With samples of convenience, the chances are hard to define; so are parameters and standard errors.

> The formulas in this chapter are for draws from a box, and should not be applied mechanically to other kinds of samples.

## Exercise Set D

*This exercise set also covers material from previous chapters.*

1. A utility company serves 50,000 households. As part of a survey of customer attitudes, they take a simple random sample of 750 of these households. The average number of television sets in the sample households turns out to be 1.86, and the

SD is 0.80. If possible, find a 95%-confidence interval for the average number of television sets in all 50,000 households.[4] If this isn't possible, explain why not.

2. Out of the 750 households in the survey of the previous exercise, 451 have computers. If possible, find a 99.7%-confidence interval for the percentage of all the 50,000 households with computers. If this isn't possible, explain why not.

3. (Continues exercises 1 and 2.) Out of the 750 households in the survey, 749 have at least one television set. If possible, find a 95%-confidence interval for the percentage of all the 50,000 households with at least one television set. If this isn't possible, explain why not.

4. As part of the survey described in exercise 1, all persons age 16 and over in the 750 sample households are interviewed. This makes 1,528 people. On the average, the sample people watched 5.20 hours of television the Sunday before the survey, and the SD was 4.50 hours. If possible, find a 95%-confidence interval for the average number of hours spent watching television on that Sunday by all persons age 16 and over in the 50,000 households. If this isn't possible, explain why not.

5. (a) As his sample, a psychology instructor takes all the students in his class. Is this a probability sample? a cluster sample?
   (b) A sociologist interviews the first 100 subjects who walk through a shopping mall one day. Does she have a probability sample? a cluster sample?

6. One hundred draws are made at random with replacement from a box. The sum of the draws is 297. Can you estimate the average of the box? Can you attach a standard error to your estimate, on the basis of the information given so far? Explain briefly.

7. A box contains 250 tickets. Two people want to estimate the average of the numbers in the box. They agree to take a sample of 100 tickets, and use the sample average as their estimate. Person A wants to draw the tickets at random without replacement; person B wants to take a simple random sample. Which procedure gives a more accurate estimate? Or does it make any difference?

*The answers to these exercises are on p. A88.*

## 5. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. A box of tickets has an average of 100, and an SD of 20. Four hundred draws will be made at random with replacement from this box.
   (a) Estimate the chance that the average of the draws will be in the range 80 to 120.
   (b) Estimate the chance that the average of the draws will be in the range 99 to 101.

2. Five hundred draws are made at random with replacement from a box with 10,000 tickets. The average of the box is unknown. However, the average of the draws was 71.3, and their SD was about 2.3. True or false, and explain:

(a) The 71.3 estimates the average of the box, but is likely to be off by 0.1 or so.
(b) A 68%-confidence interval for the average of the box is $71.3 \pm 0.1$.
(c) About 68% of the tickets in the box are in the range $71.3 \pm 0.1$.

3. A real estate office wants to make a survey in a certain town, which has 50,000 households, to determine how far the head of household has to commute to work.[5] A simple random sample of 1,000 households is chosen, the occupants are interviewed, and it is found that on average, the heads of the sample households commuted 8.7 miles to work; the SD of the distances was 9.0 miles. (All distances are one-way; if someone isn't working, the commute distance is defined to be 0.)

   (a) The average commute distance of all 50,000 heads of households in the town is estimated as _____, and this estimate is likely to be off by _____ or so.
   (b) If possible, find a 95%-confidence interval for the average commute distance of all heads of households in the town. If this isn't possible, explain why not.

4. (Continues exercise 3.) The real estate office interviewed all persons age 16 and over in the sample households; there were 2,500 such persons. On the average, these 2,500 people commuted 7.1 miles to work, and the SD of the distances was 10.2 miles. (Again, if someone isn't working, the commute distance is defined to be 0; and all distances are one-way.) If possible, find a 95%-confidence interval for the average commute distance for all people age 16 and over in this town. If this isn't possible, explain why not.

5. (Continues exercise 4.) In 721 of the sample households, the head of the household commuted by car. If possible, find a 95%-confidence interval for the percentage of all households in the town where the head of the household commutes by car. If this isn't possible, explain why not.

6. The National Assessment of Educational Progress (NAEP) periodally administers tests on different subjects to high school students.[6] In 2000, the grade 12 students in the sample averaged 301 on the mathematics test; the SD was 30. The likely size of the chance error in the 301 is about _____.

   (a) Can you fill in the blank if a cluster sample of 1,000 students was tested? If so, what is the answer? If not, why not?
   (b) Can you fill in the blank if a simple random sample of 1,000 students was tested? If so, what is the answer? If not, why not?

7. A city government did a survey of working women, to see how they felt about juggling jobs and family responsibilities. Businesses, unions, and community service organizations helped distribute the survey questionnaire to locations where the women could pick up copies. 1,678 out of 2,800 respondents, or 59.9%, checked the item "stress is a serious problem" on the questionnaire. Choose one option, and explain briefly.

(i) The standard error on the 59.9% is 0.9 of 1%.
(ii) The standard error on the 59.9% is some other number.
(iii) Neither of the above.

8. One year, there were about 3,000 institutions of higher learning in the U.S. (including junior colleges and community colleges). As part of a continuing study of higher education, the Carnegie Commission took a simple random sample of 400 of these institutions.[7] The average enrollment in the 400 sample schools was 3,700, and the SD was 6,500. The Commission estimates the average enrollment at all 3,000 institutions to be around 3,700; they put a give-or-take number of 325 on this estimate. Say whether each of the following statements is true or false, and explain. If you need more information to decide, say what you need and why.

   (a) An approximate 68%-confidence interval for the average enrollment of all 3,000 institutions runs from 3,375 to 4,025.
   (b) If a statistician takes a simple random sample of 400 institutions out of 3,000, and goes one SE either way from the average enrollment of the 400 sample schools, there is about a 68% chance that his interval will cover the average enrollment of all 3,000 schools.
   (c) About 68% of the schools in the sample had enrollments in the range $3,700 \pm 6,500$.
   (d) It is estimated that 68% of the 3,000 institutions of higher learning in the U.S. enrolled between $3,700 - 325 = 3,375$ and $3,700 + 325 = 4,025$ students.
   (e) The normal curve can't be used to figure confidence levels here at all, because the data don't follow the normal curve.

9. (Continues exercise 8.) There were about 600,000 faculty members at institutions of higher learning in the U.S. As part of its study, the Carnegie Commission took a simple random sample of 2,500 of these faculty persons.[8] On the average, these 2,500 sample persons had published 1.7 research papers in the two years prior to the survey, and the SD was 2.3 papers. If possible, find an approximate 95%-confidence interval for the average number of research papers published by all 600,000 faculty members in the two years prior to the survey. If this isn't possible, explain why not.

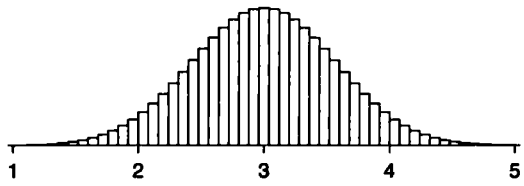
10. A survey organization takes a simple random sample of 625 households from a city of 80,000 households. On the average, there are 2.30 persons per sample household, and the SD is 1.75. Say whether each of the following statements is true or false, and explain.

    (a) The SE for the sample average is 0.07.
    (b) A 95%-confidence interval for the average household size in the sample is 2.16 to 2.44.
    (c) A 95%-confidence interval for the average household size in the city is 2.16 to 2.44.
    (d) 95% of the households in the city contain between 2.16 and 2.44 persons.

(e) The 95%-confidence level is about right because household size follows the normal curve.

(f) The 95%-confidence level is about right because, with 625 draws from the box, the probability histogram for the average of the draws follows the normal curve.

11. The figure below is a probability histogram for the average of 25 draws made at random with replacement from the box $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}$. Or is something wrong? Explain.




12. One term at the University of California, Berkeley, 400 students took the final in Statistics 2. Their scores averaged 65.3 out of 100, and the SD was 25. Now

$$\sqrt{400} \times 25 = 500, \qquad 500/400 = 1.25$$

Is $65.3 \pm 2.5$ a 95%-confidence interval? If so, for what? If not, why not?

## 6. SPECIAL REVIEW EXERCISES

*These exercises cover all of parts I–VI.*

1. An experiment was carried out to determine the effect of providing free milk to school children in a certain district (Lanarkshire, Scotland).[9] Some children in each school were chosen for the treatment group and got free milk; others were chosen for controls and got no milk. Assignment to treatment or control was done at random, to make the two groups comparable in terms of family background and health.

   After randomization, teachers were allowed to use their judgment in switching children between treatment and control, to equalize the two groups. Was it wise to let the teachers use their judgment this way? Answer yes or no, and explain briefly.

2. For the portacaval shunt (section 2 of chapter 1), survival among the controls in the poorly-designed trials was worse than survival among the controls in the randomized controlled experiments. Is it dangerous to be a control in a poorly-designed study? Answer yes or no, and explain. If your answer is no, what accounts for the difference in survival rates?

3. (a) Epidemiologists find a higher rate of oral cancer among drinkers than non-drinkers. If alcohol causes oral cancer, would that tend to create an asso-

ciation between drinking and oral cancer? Answer yes or no, and discuss briefly.

(b) Epidemiologists find an association between high levels of cholesterol in the blood and heart disease. They conclude that cholesterol causes heart disease. However, a statistician argues that smoking confounds the association, meaning that—

(i) Smoking causes heart disease.

(ii) Smoking causes heart disease, and smokers have high levels of cholesterol in their blood.

(iii) Smokers tend to eat a less healthful diet than non-smokers. Thus, smokers have high levels of cholesterol in the blood, which in turn causes heart disease.

(iv) The percentage of smokers is about the same among persons with high or low levels of cholesterol in the blood.

Choose one option, and discuss briefly.

4. A follow-back study on a large sample of death certificates shows the average age at death is smaller for left-handed people than for right-handers. (In this kind of study, surviving relatives are interviewed.)

(a) Suppose that, other things being equal (age, sex, race, income, etc.), left-handed people are more at risk from accident and disease than right handers. Could that explain a difference in average age at death?

(b) During the twentieth century, there were big changes in child-rearing practices. In the early part of the century, parents insisted on raising children to be right-handed. By mid-century, parents were much more tolerant of left-handedness. Could that explain a difference in average age at death of left-handed and right-handed people in 2005?

(c) What do you conclude from the death certificate data?

5. Before a strike in 1994, the median salary of the 746 major league baseball players was about \$500,000. The lowest salary was about \$100,000 and the highest was over \$5,000,000. Choose one option and explain:

(i) The owners were paying out around $746 \times \$500{,}000 = \$373$ million per year in salaries to the players.

(ii) The owners were paying out substantially less than \$373 million per year to the players.

(iii) The owners were paying out substantially more than \$373 million per year to the players.

6. In HANES3, the Public Health Service interviewed a representative sample of Americans. Among other things, respondents age 25 and over were asked about their geographic mobility—how often did they move? About 20% of them had moved in the last year. At the other extreme, about 25% of them had been living at the same address for 15 years or more; 5% had been at the same address for 35 years or more! The average time since the last move was

10 years, and the SD was _____. Fill in the blank using one of the options below, and explain briefly.

1 year 2 years 10 years 25 years

7. To measure water clarity in a lake, a glass plate with ruled lines is pushed down into the water until the lines cannot be seen any more. The distance below the surface of the water is called "Secchi depth." To measure pollution by algae, scientists determine the total concentration of chlorophyll in the water. In a certain lake, Secchi depth and total chlorophyll concentration are measured every Thursday at noon, from April through September. Will the correlation between these variables be positive or negative? Explain briefly.

8. An instructor standardizes her midterm and final each semester so the class average is 50 and the SD is 10 on both tests. The correlation between the tests is around 0.50. One semester, she took all the students who scored around 30 at the midterm, and gave them special tutoring. On average, they gained 10 points on the final. Can this be explained by the regression effect? Answer yes or no, and explain briefly.

9. For entering freshmen at a certain university, scores on the Math SAT and Verbal SAT can be summarized as follows:

$$\text{average M-SAT} = 555, \quad \text{SD} = 125$$
$$\text{average V-SAT} = 543, \quad \text{SD} = 115, \quad r = 0.66$$

The scatter diagram is football-shaped. One student is chosen at random and has an M-SAT of 600. You would guess his V-SAT is _____ points, and would have about a 68% chance to be right within _____ points. Fill in the blanks; explain briefly.

10. Pearson and Lee obtained the following results in a study of about 1,000 families:

$$\text{average height of husband} \approx 68 \text{ inches, SD} \approx 2.7 \text{ inches}$$
$$\text{average height of wife} \approx 63 \text{ inches, SD} \approx 2.5 \text{ inches, } r \approx 0.25$$

Among the men who were about 5 feet 4 inches tall, estimate the percentage who were shorter than their wives.

11. In a large study of the relationship between incomes of husbands and wives, the following results were obtained:

$$\text{average income of husband} \approx \$50{,}000, \text{ SD} \approx \$40{,}000$$
$$\text{average income of wife} \approx \$40{,}000, \text{ SD} \approx \$30{,}000, \; r \approx 0.33$$

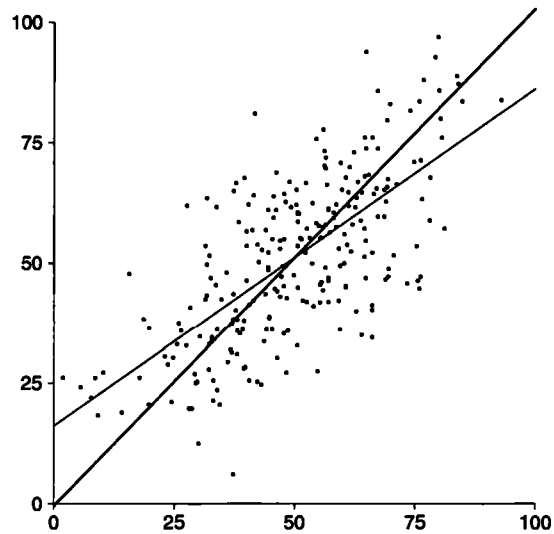
(a) The couples were divided into groups according to the income of the husbands (\$0–\$4,999, \$5,000–\$9,999, \$10,000–\$14,999, etc.). The average income for wives in each group was calculated and then plotted above the midpoint of the corresponding range (\$2,500, \$7,500, \$12,500, etc.). It was found that the points on this graph followed a straight line very closely. The slope of this line would be about

0.25 0.75 0.83 1 1.33

Explain briefly. If more information is needed, say what you need and why.

(b) For one couple in the study, the wife's income was \$37,500, but the information about her husband's income was lost. At \$40,000, the height of the line plotted in part (a) equals \$37,500. Is \$40,000 a good estimate for the husband's income? Or is the estimate likely to be too high? too low? Why?

12. The figure below shows a scatter diagram, with two lines. One estimates the average value of $y$ for each $x$. The other estimates the average value of $x$ for each $y$. Or is something wrong? Explain briefly. (The average of $x$ is 50, and the SD is 17; the statistics for $y$ are just about the same.)




13. Five cards will be dealt from a well-shuffled deck. Find the chance of getting an ace or a king among the 5 cards. (A deck has 52 cards, of which 4 are aces and 4 are kings.)

14. Out of the 300 people enrolled in a large course, 6 got a perfect score on the first midterm and 9 got a perfect score on the second midterm. One person will be chosen at random from the class. If possible with the information given, find the chance that person has a perfect score on both midterms. Otherwise, say what information is needed, and why.

15. A die is rolled 6 times. Find the chance that the first number rolled comes up 3 more times—

    (a) If the first roll is an ace.
    (b) If the first roll is a six.
    (c) If you don't know what happens on the first roll.

    (A die has 6 faces, showing 1 through 6 spots; an ace is ⚀; each face is equally likely to come up.)

16. A Nevada roulette wheel has 38 pockets. One is marked "0," another is marked "00," and the rest are numbered from 1 through 36. The wheel is spun and a ball is dropped. The ball is equally likely to end up in any one of the 38 pockets (figure 3 on p. 282). Here are two possibilities:

    (i) You win \$1 if any 7's turn up in 15 spins of the wheel.
    (ii) You win \$1 if any 7's turn up in 30 spins of the wheel.

    True or false, and explain: the second possibility gives you twice as much of a chance to win as the first.

17. A die will be rolled 20 times. The sum

    number of ones rolled + number of sixes rolled

    will be around _____, give or take _____ or so.

18. A multiple-choice quiz has 50 questions. Each question has 3 possible answers, one of which is correct. Two points are given for each correct answer, but a point is taken off for a wrong answer.

    (a) The passing score is 50. If a student answers all the questions at random, what is the chance of passing?
    (b) Repeat part (a), if the passing score is 10.

19. "Toss a hundred pennies in the air and record the number of heads that come up when they fall. Do this several thousand times and plot a histogram for the numbers that you get. You will have a histogram that closely approximates the normal curve, and the more times you toss the hundred pennies the closer your histogram will get to the curve."[10] If you keep on tossing this group of a hundred pennies, will your histogram get closer and closer to the normal curve? Or will it converge to the probability histogram for the number of heads in 100 tosses of a coin? Choose one option, and explain briefly.

20. Twenty-five draws will be made at random with replacement from the box $\boxed{1}\boxed{2}\boxed{9}$.

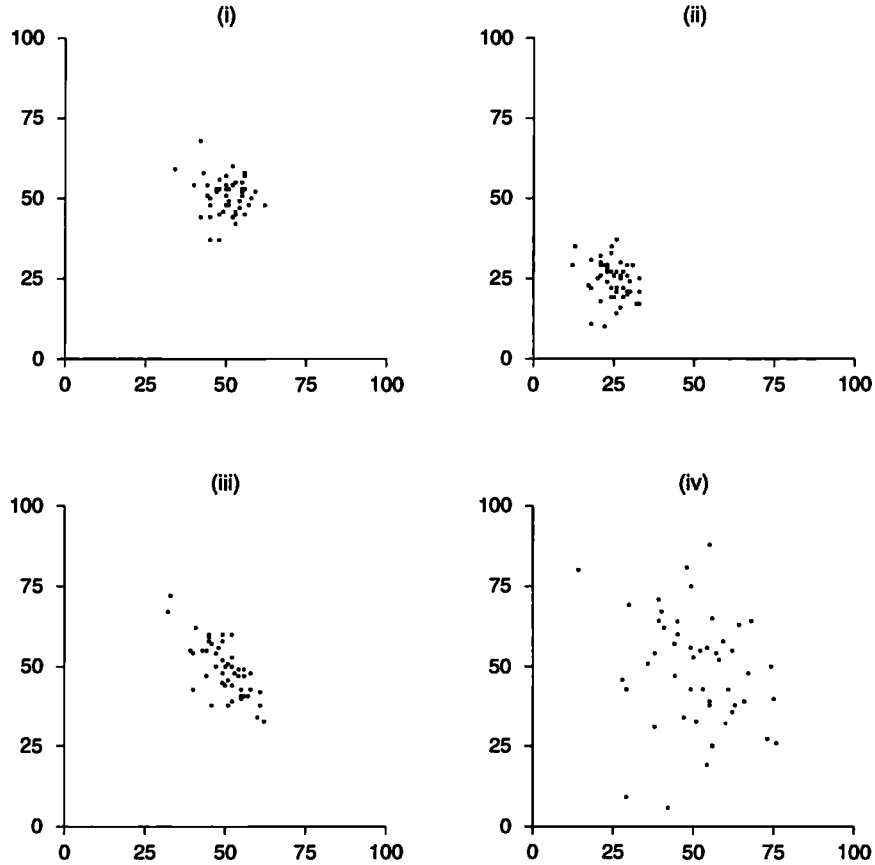
    (a) A statistician uses the normal curve to compute the chance that the sum of the draws will equal 90. The result is

    too low  too high  about right

    Choose one option, and explain.
    (b) Repeat, for the chance that the sum is between 90 and 110.

    No calculations are necessary, just look at figure 9 on p. 322.

21. Imagine making a scatter diagram from table 3 on p. 302 as follows. Plot the point whose $x$-coordinate is the number of heads in tosses #1–100, and whose $y$-coordinate is the number of heads in tosses #101–200. This gives (44, 54). Then plot the point whose $x$-coordinate is the number of heads on tosses #201–300, and whose $y$-coordinate is the number of heads in tosses #301–400. This gives (48, 53). And so on. One of the scatter diagrams on the next page plots the data. Which one? Explain briefly.



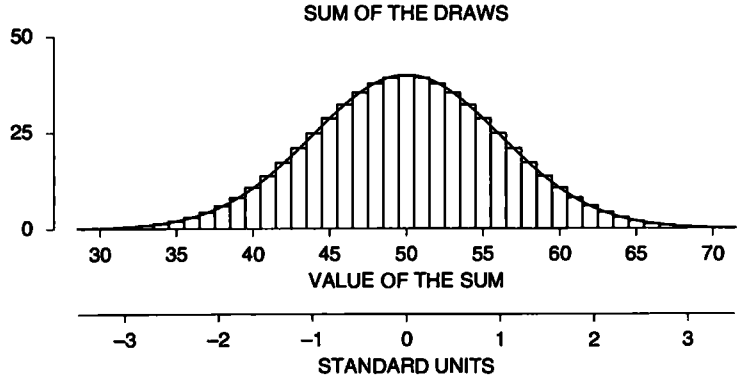

22. A box contains 10,000 marbles: 6,000 are red and 4,000 are blue; 500 marbles are drawn at random without replacement.

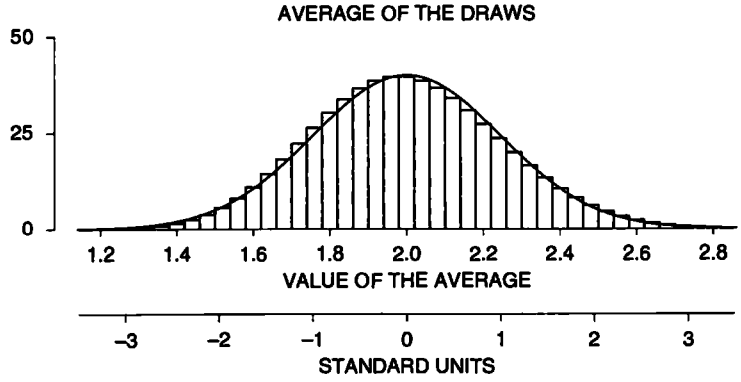
    (a) Suppose there are 218 blue marbles in the sample. Find the expected value for the percentage of blues in the sample, the observed value, the chance error, and the standard error.

    (b) Suppose there are 191 blue marbles in the sample. Find the expected value for the percentage of blues in the sample, the observed value, the chance error, and the standard error.

23. The top panel in the figure on the next page shows the probability histogram for the sum of 25 draws made at random with replacement from box A. The bottom panel shows the probability histogram for the average of 25 draws made at random with replacement from box B. Choose one option and explain briefly; if you choose (iii), say what additional information is needed.

    (i) Box A and Box B are the same.
    (ii) Box A and Box B are different.
    (iii) Can't tell without more information.




24. Draws are being made at random with replacement from a box. The number of draws is getting larger and larger. Say whether each of the following statements is true or false, and explain. ("Converges" means "gets closer and closer.")

    (a) The probability histogram for the sum of the draws (when put in standard units) converges to the normal curve.

    (b) The histogram for the numbers in the box (when put in standard units) converges to the normal curve.

    (c) The histogram for the numbers drawn (when put in standard units) converges to the normal curve.

    (d) The probability histogram for the product of the draws (when put in standard units) converges to the normal curve.

    (e) The histogram for the numbers drawn converges to the histogram for the numbers in the box.

25. (Hypothetical) A retailer has 1,000 stores nationwide. Each store has 10 to 15

employees, for a national total of 12,000. The personnel department has done a study of these employees, to assess morale. The report begins:

> Findings are based on interviews with 250 employees. We took a simple random sample of 50 stores, and interviewed 5 employees at each of the sample stores. Interviews were done by a team of occupational psychologists provided under contract by an independent survey organization. Since the interviews were anonymous, we do not know the names of the interviewees . . . .

At this point, there should be a question you want answered. What is your question, and why does it matter?

26. In 1965, the U.S. Supreme Court decided the case of *Swain v. Alabama.*[11] Swain, a black man, was convicted in Talladega County, Alabama, of raping a white woman. He was sentenced to death. The case was appealed to the Supreme Court on the grounds that there were no blacks on the jury; even more, no black "within the memory of persons now living has ever served on any petit jury in any civil or criminal case tried in Talladega County, Alabama."

    The Supreme Court denied the appeal, on the following grounds. As provided by Alabama law, the jury was selected from a panel of about 100 persons. There were 8 blacks on the panel. (They did not serve on the jury because they were "struck," through peremptory challenges by the prosecution; such challenges were constitutionally protected until 1986.) The presence of 8 blacks on the panel showed "the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes."

    At that time in Alabama, only men over the age of 21 were eligible for jury duty. There were 16,000 men over the age of 21 in Talladega County, of whom about 26% were black. If 100 people were chosen at random from this population, what is the chance that 8 or fewer would be black? What do you conclude?

27. The town of Hayward (California) has about 50,000 registered voters. A political scientist takes a simple random sample of 500 of these voters. In the sample, the breakdown by party affiliation is

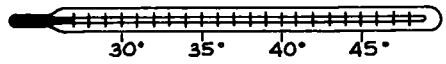
| | |
|---|---|
| Republican | 115 |
| Democrat | 331 |
| Independent | 54 |

    (a) Among all registered voters in Hayward, the percentage of independents is estimated as _____.
    (b) This estimate is likely to be off by _____ or so.
    (c) The range from _____ to _____ is a 95%-confidence interval for the percentage of independents _____.

    Fill in the blanks; explain briefly. (The first four blanks are filled in with numbers; the last blank takes a phrase—25 words or less.)

28. NAEP (National Assessment of Educational Progress) periodically tests scientific knowledge in U.S. schools.[12] Here is one question on the test, administered to students in grade 12.

> The diagram below shows a thermometer. On the diagram, fill in the thermometer so that it reads 37.5 degrees Celsius.

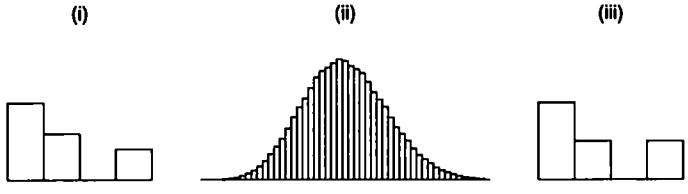



Only 64% of the students who were tested could answer this question correctly.

The superintendent of education in a certain state cannot believe these data. To check, he takes a simple random sample of 100 high schools in the state, and tests 10 randomly selected students from Grade 12 in each school. 661 out of the 1,000 students who take the test, or 66.1%, can do the problem.

With the information given above, can you put a standard error on the 66.1%? Find the SE, or say why this can't be done.

29. Twenty draws are made at random with replacement from the box $\boxed{1}\boxed{1}\boxed{2}\boxed{4}$. One of the graphs below is the probability histogram for the average of the draws. Another is the histogram for the numbers drawn. And the third is the histogram for the contents of the box. Which is which? Explain.




30. A survey research company uses random digit dialing. They have a contract, to estimate the percentage of people watching Spanish-language television in several Southwestern cities. They took a sample of size 1,000 in Austin, Texas—which has about 600,000 residents. They are satisfied with the accuracy of the estimates for Austin.

Dallas has about twice the population of Austin, but similar demographics. True or false, and explain: to get about the same accuracy in Dallas as in Austin, the company should use a sample size of 2,000.

## 7. SUMMARY AND OVERVIEW

1. When drawing at random from a box, the expected value for the average of the draws equals the average of the box. The SE for the average of the draws equals the SE for their sum, divided by the number of draws.

2. The average of the draws can be used to estimate the average of the box. The estimate will be off by some amount, due to chance error. The SE for the average tells you the likely size of the amount off.

3. Multiplying the number of draws by some factor divides the SE for their average by the square root of that factor.

4. The probability histogram for the average of the draws will follow the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be large.

5. With a simple random sample, the SD of the sample can be used to estimate the SD of the box. A confidence interval for the average of the box can be found by going the right number of SEs either way from the average of the draws. The confidence level is read off the normal curve. This method should only be used with large samples.

6. The formulas for simple random samples should not be applied mechanically to other kinds of samples.

7. With *samples of convenience*, standard errors usually do not make sense.

8. This part of the book makes the transition from probability calculations to inference. Chapter 19 distinguishes sampling error from non-sampling error, and shows how important it is to use probability methods when drawing samples. Non-sampling error is often a more subtle and important problem than sampling error. Chapter 20 develops the theory behind simple random sampling. Chapter 21 shows how to estimate population percentages from sample percentages, introducing SEs and confidence intervals based on sample data. Chapter 23 makes the extension to averages.

9. Chapters 20, 21, and 23 build on the probability theory developed in chapters 16–18. These ideas will be applied again in part VII to the study of measurement error; they will be used in part VIII to make tests of significance.

10. The Current Population Survey is discussed in chapter 22, illustrating the concepts in a real survey of some complexity.

# PART VII

# Chance Models

# 24

# A Model for Measurement Error

*Upon the whole of which it appears, that the taking of the Mean of a number of observations, greatly diminishes the chance for all the smaller errors, and cuts off almost all possibility of any great ones: which last consideration, alone, seems sufficient to recommend the use of the method, not only to astronomers, but to all others concerned in making experiments of any kind (to which the above reasoning is equally applicable). And the more observations or experiments there are made, the less will the conclusions be liable to error, provided they admit of being repeated under the same circumstances.*

—THOMAS SIMPSON (ENGLISH MATHEMATICIAN, 1710–1761)

## 1. ESTIMATING THE ACCURACY OF AN AVERAGE

In this part of the book, the frequency theory of chance will be used to study measurement error and genetics. Historically, the frequency theory was developed to handle problems of a very special kind—figuring the odds in games of chance. Some effort is needed to apply the theory to situations outside the gambling context. In each case, it is necessary to show that the situation being studied resembles a process—like drawing at random from a box—to which the theory applies. These box models are sometimes called *chance models* or *stochastic models*. The first example will be a chance model for measurement error.

To review briefly (chapter 6), any measurement is subject to chance error, and if repeated would come out a bit differently. To get at the size of the chance error, the best thing to do is to repeat the measurement several times. The spread

in the measurements, as shown by the SD, estimates the likely size of the chance error in a single measurement. Chapter 6 stopped there. This chapter continues the discussion: the focus is on the average of the measurements in the series rather than a single measurement. The problem is to estimate the likely size of the chance error in the average. If the measurements are like draws from a box, the methods of parts V and VI can be used.

Table 1 on p. 99 shows 100 measurements on NB 10. These all fell short of 10 grams, by different amounts. The table gives the amounts, in micrograms. (A microgram is one millionth of a gram, roughly the weight of a speck of dust.) The SD of the 100 numbers in the table is about 6 micrograms: a single measurement is only accurate up to 6 micrograms or so. The best guess for the weight of NB 10 is the average of all 100 measurements, which is 404.6 micrograms short of 10 grams. Since each measurement is thrown off by error, the average cannot be exactly right either. But the average is going to be more accurate than any single measurement, so it is going to be off by less than 6 micrograms.

What is the right give-or-take number to put on the average?

$$\text{average} \pm _____.$$

The answer is given by the SE for the average, which can be calculated just as in chapter 23. (The calculation rides on a box model, to be discussed in sections 2 and 3 below.) The SE for the sum of 100 measurements can be estimated as

$$\sqrt{100} \times 6 \text{ micrograms} = 60 \text{ micrograms}.$$

Then the SE for the average of the 100 measurements is

$$\frac{60 \text{ micrograms}}{100} = 0.6 \text{ micrograms}.$$

This completes the calculation. The average of all the numbers in the table is 404.6 micrograms. The likely size of the chance error in the average is estimated to be 0.6 micrograms. So NB 10 really weighs about 404.6 micrograms below 10 grams, plus or minus 0.6 micrograms or so.

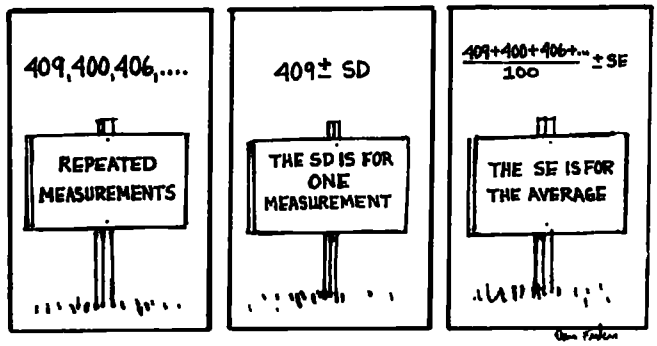
Two numbers come up in the calculation: 6 micrograms and 0.6 micrograms. The first is the SD of the 100 measurements, the second is the SE for the average. What is the difference between the two?

- The SD says that a single measurement is accurate up to 6 micrograms or so.
- The SE says that the average of all 100 measurements is accurate up to 0.6 micrograms or so.

*Example 1.* One hundred measurements are made on a certain weight. The average of these measurements is 715 micrograms above one kilogram, and the SD is 80 micrograms.

(a) Is a single measurement likely to be off the exact weight by around 8 micrograms, or 80 micrograms?

(b) Is the average of all 100 measurements likely to be off the exact weight by around 8 micrograms, or 80 micrograms?




*Solution.* A single measurement is off by a chance error similar in size to the SD of the measurements. This is 80 micrograms. The answer to (a) is 80 micrograms. For (b), the SE for the sum of the measurements is estimated as $\sqrt{100} \times 80 = 800$ micrograms. So the SE for the average is $800/100 = 8$ micrograms. That is the answer to (b).

In example 1, the give-or-take number for the average of the measurements is 8 micrograms. To make this more precise, statisticians use confidence intervals, just as in sampling. A 95%-confidence interval for the exact weight can be obtained by going 2 SEs in either direction from the average. The average is 715 micrograms above one kilogram, and 2 SEs is $2 \times 8 = 16$ micrograms. So the exact weight is somewhere between 699 and 731 micrograms above one kilogram, with confidence about 95%. The arithmetic:

$$715 - 16 = 699, \quad 715 + 16 = 731.$$

Again, the word "confidence" is there as a reminder that the chances are in the measurement process and not in the thing being measured: the exact weight is not subject to chance variability. (For a similar discussion in the sampling context, see section 3 of chapter 21.)

> The chances are in the measuring procedure, not the thing being measured.

The normal curve should be used to get confidence intervals only when there is a fairly large number of measurements. With fewer than 25 measurements, most statisticians would use a slightly different procedure, based on what is called the $t$-distribution (section 6 of chapter 26).

*Historical note.* There is a connection between the theory of measurement error and neon signs. In 1890, the atmosphere was believed to consist of nitrogen (about 80%), oxygen (a little under 20%), carbon dioxide, water vapor—and nothing else. Chemists were able to remove the oxygen, carbon dioxide, and water vapor. The residual gas should have been pure nitrogen.

Lord Rayleigh undertook to compare the weight of the residual gas with the weight of an equal volume of chemically pure nitrogen. One measurement on the weight of the residual gas gave 2.31001 grams. And one measurement of the pure nitrogen gave a bit less, 2.29849 grams. However, the difference of 0.01152 grams was rather small, and in fact was comparable to the chance errors made by the weighing procedure.

Could the difference have resulted from chance error? If not, the residual gas had to contain something heavier than nitrogen. What Rayleigh did was to replicate the experiment, until he had enough measurements to prove that the residual gas from the atmosphere was heavier than pure nitrogen.

He went on to isolate the rare gas called *argon*, which is heavier than pure nitrogen and present in the atmosphere in small quantities. Other researchers later discovered the similar gases neon, krypton, and xenon, all occurring naturally (in trace amounts) in the atmosphere. These gases are what make "neon" signs glow in different colors.[1]

## Exercise Set A

1. The total of the 100 measurements on NB 10 was 40,459 micrograms. What is the likely size of the chance error in this total?

2. Some scales use electrical *load cells*. The weight is distributed over a number of cells. Each cell converts the weight it carries to an electrical current, which is fed to a central scanner. This scanner adds up all the currents, and computes the corresponding total weight, which it prints out. This process is repeated several dozen times a second. As a result, a loaded boxcar (weighing about 100,000 pounds) can be weighed as it crosses a special track, with chance errors of only several hundred pounds in size.[2]

   Suppose 25 readings on the weight of a boxcar show an average of 82,670 pounds, and the SD is 500 pounds. The weight of the boxcar is estimated as _____; this estimate is likely to be off by _____ or so.

3. (Hypothetical.) The British Imperial Yard is sent to Paris for calibration against The Meter. Its length is determined 100 times. This sequence of measurements averages out to 91.4402 cm, and the SD is 800 microns. (A *micron* is the millionth part of a meter.)

   (a) Is a single reading off by around 80 microns, or 800 microns?
   (b) Is the average of all 100 readings off by around 80 microns, or 800 microns?
   (c) Find a 95%-confidence interval for the exact length of the Imperial Yard.

4. The 95%-confidence interval for the exact weight of NB 10 is the range from 403.4 to 405.8 micrograms below 10 grams. Say whether each of the following statements is true or false, and explain why.

   (a) About 95% of the measurements are in this range.
   (b) There is about a 95% chance that the next measurement will be in this range.
   (c) About 95% of the time that the Bureau takes 100 measurements and goes 2 SEs either way from the average, they succeed in covering the exact weight.
   (d) If the Bureau took another 100 measurements on NB 10, there is about a 95% chance that the new average would fall in the interval from 403.4 to 405.8 micrograms below 10 grams.

5. Would taking the average of 25 measurements divide the likely size of the chance error by a factor of 5, 10, or 25?

*The answers to these exercises are on p. A88.*

## 2. CHANCE MODELS

Section 1 explained how to put a standard error on the average of repeated measurements. The arithmetic is easily carried out on any list of numbers, but the method is legitimate only when the variability in the data is like the variability in repeated draws from a box.

> If the data show a trend or pattern over time, a box model does not apply.

The reason: draws from a box do not show a trend or pattern over time. The following examples illustrate this idea.

*Example 2.* Table 1 gives the population of the U.S. from 1790 to 2000. Do these numbers look like draws at random from a box?

Table 1. Population of the U.S., 1790 to 2000.

| | |
|---|---|
| 1790 | 3,929,214 |
| 1800 | 5,308,483 |
| 1810 | 7,239,881 |
| 1820 | 9,638,453 |
| 1830 | 12,866,020 |
| 1840 | 17,069,453 |
| 1850 | 23,191,876 |
| 1860 | 31,443,321 |
| 1870 | 39,818,449 |
| 1880 | 50,189,209 |
| 1890 | 62,979,766 |
| 1900 | 76,212,168 |
| 1910 | 92,228,496 |
| 1920 | 106,021,537 |
| 1930 | 123,202,624 |
| 1940 | 132,164,569 |
| 1950 | 151,325,798 |
| 1960 | 179,323,175 |
| 1970 | 203,302,031 |
| 1980 | 226,542,199 |
| 1990 | 248,718,302 |
| 2000 | 281,422,602 |

Notes: Resident population. From 1950 onwards, includes Alaska and Hawaii. Revised figures for 1870–1940. Source: *Statistical Abstract*, 2006, Table 1.

*Solution.* No. The population of the U.S. has been going up steadily. Numbers drawn at random from a box don't do that: sometimes they go up and other times they go down.

*Example 3.* The 22 numbers in table 1 average out to 94.7 million, and the SD is 89.3 million. An investigator attaches a standard error to the average, by the following procedure:

$$\text{SE for the sum} \approx \sqrt{22} \times 89.3 \text{ million} \approx 419 \text{ million}$$
$$\text{SE for average} \approx 419/22 \approx 19.0 \text{ million.}$$

Is this sensible?

*Solution.* The average and SD make sense, as descriptive statistics. They summarize part of the information in table 1, although they miss quite a bit—for instance, the fact that the numbers increase steadily. The SE of 19 million, however, is silly. If the investigator wants to know the average of the 22 numbers in the table, that has been computed, and there is no need to worry about chance error. Of course, something else may be involved, like the average of a list showing the population of the U.S. in every year from 1790 to 2000. (Every tenth number on that list is shown in table 1; the numbers in between are known with less precision, because the Census is only taken every ten years.) The investigator would then be making an inference, using the average from table 1 to estimate that other average. And the estimate would be off by some amount. But the square root law cannot help much with the margin of error. The reason is that the numbers in table 1 are not like draws from a box.

| The square root law only applies to draws from a box. |
|---|

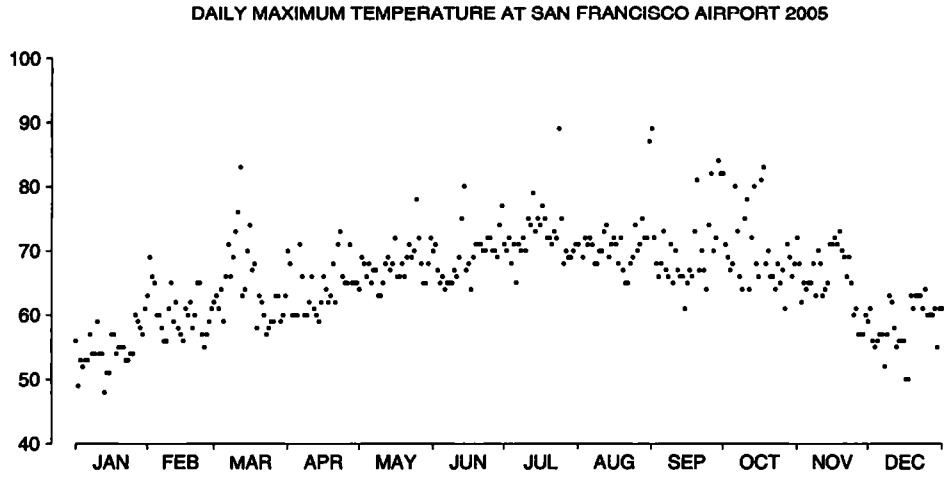
*Example 4.* A list is made, showing the daily maximum temperature at San Francisco airport. Are these data like draws from a box?

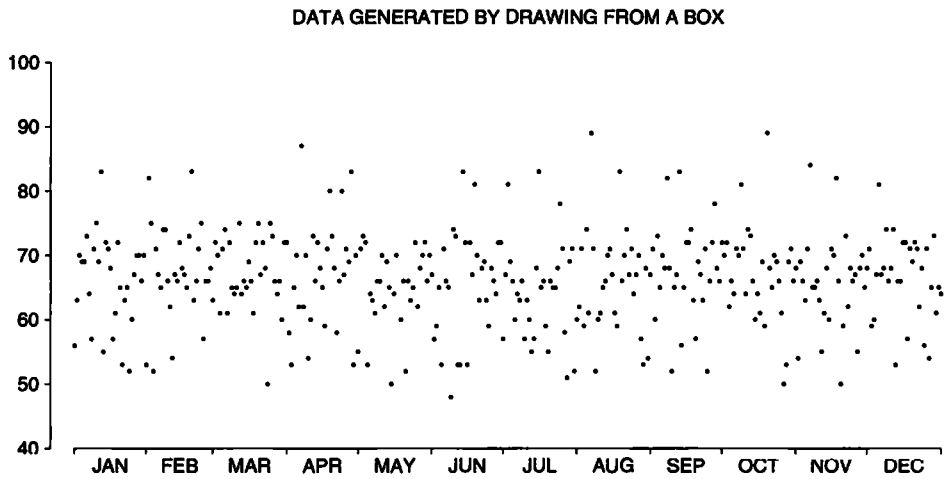
*Solution.* No, there is a definite seasonal pattern to these data—warmer in the summer, colder in the winter. There even are local patterns to the data. The temperature on one day tends to be like the temperature on the day before.

The temperature data are graphed in the top panel of figure 1. There is a dot above each day of the year for 2005, showing the maximum temperature on that day. The seasonal pattern is clear. On the whole, the dots are higher in the summer than in the winter. Also, there is an irregular wavy pattern within each season. The crest of a wave represents a stretch of warm days—a warm spell. The cold spells are in the troughs.

By comparison, the second panel in figure 1 is for a mythical airport where the climate is on average like San Francisco, but the daily maximum temperatures are like draws from a box. These data are random: they show no trend or pattern over the year. With this sort of climate, weather forecasting would be hopeless.

Figure 1. Temperature and box models. The first panel shows the daily maximum temperature at San Francisco airport in 2005.[3] There is a seasonal pattern to the data, warmer in summer than winter. Also, there are local patterns: warm spells and cold spells. A box model would not apply. The second panel shows what the temperatures would look like if they were generated by drawing from a box.

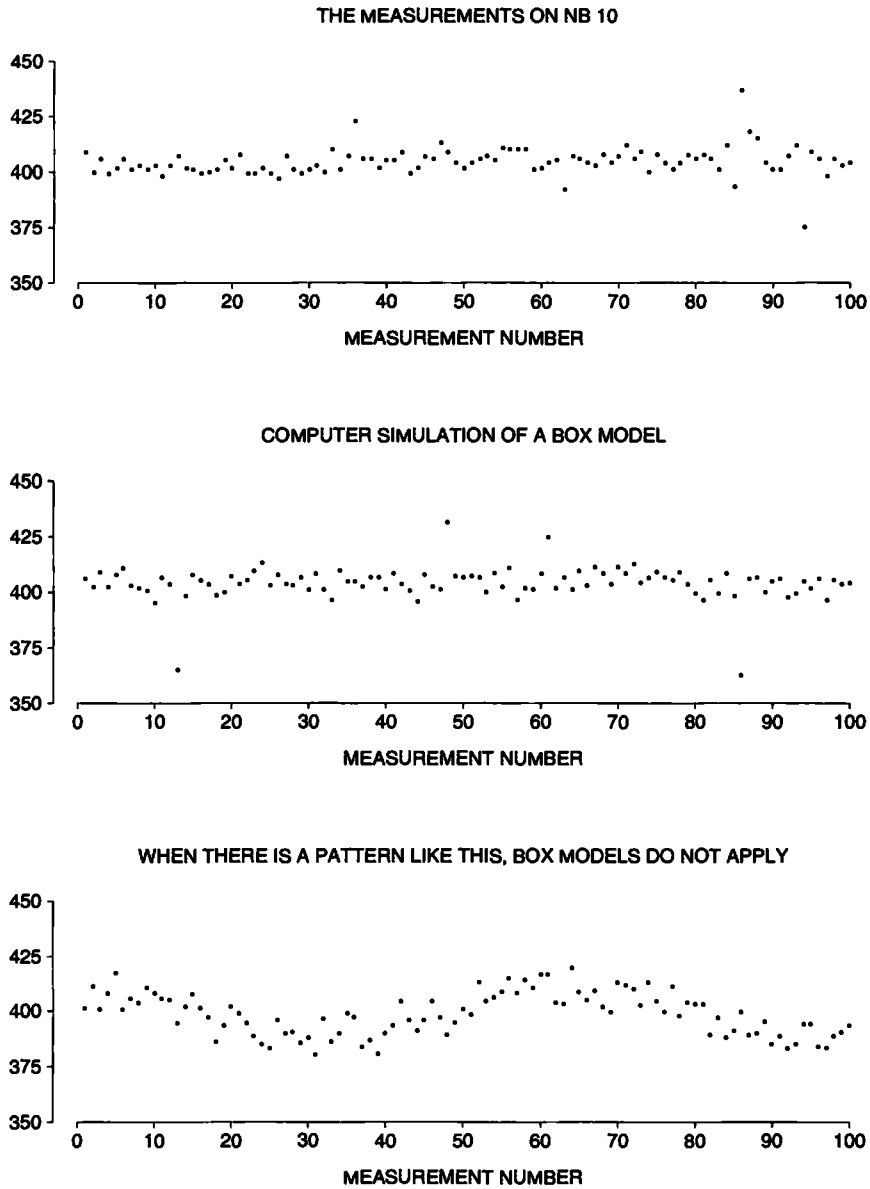



In section 1, we acted as if the measurements on NB 10 were like draws from a box. Was this sensible? The top panel in figure 2 (next page) is a graph of the data. There is one point for each measurement. The $x$-coordinate says which

measurement it was: first, or second, or third, and so on. The $y$-coordinate says how many micrograms below 10 grams the measurement was. The points do not show any trend or pattern over time; they look as random as draws from a box. In

Figure 2. The top panel graphs the repeated measurements on NB 10 (p. 99). The middle panel graphs hypothetical data, generated by computer simulation of a box model. These two panels are very similar, showing how well the box model represents the real data. The bottom panel graphs data showing a strong pattern: a box model would not apply.



fact, the second panel shows hypothetical data generated on the computer using a box model.[4] If you did not know which was which, it would be hard to tell the difference between these two panels. By comparison the third panel (also for computer-generated data) shows a strong pattern: a box model would not apply.

It is no accident that the data on NB 10 look like draws from a box. Investigators at the Bureau use pictures of the data, like the top panel in figure 2, to check their work. A trend or pattern is a signal that something is wrong and needs to be fixed. This idea is basic to precision measurement work—and to quality control in manufacturing, where the number of defective units is plotted against time.

## Exercise Set B

1. A thumbtack is thrown in the air. It lands either point up or point down.

   Someone proposes the following box model: drawing with replacement from the box [U D], where U means "point up" and D means "point down." Someone else suggests the box [U D D]. How could you decide which box was better?

2. In San Francisco, it rains on about 17% of the days in an average year. Someone proposes the following chance model for the sequence of dry and rainy days: draw with replacement from a box containing one card marked "rainy" and five cards marked "dry." Is this a good model?

3. Someone goes through the phone book, and makes a list showing the last digit of each phone number. Can this be modeled by a sequence of draws (with replacement) from the box

   [0 1 2 3 4 5 6 7 8 9]

   What about a list of first digits?

4. Someone makes a list showing the first letter of each family name in the phone book, going name by name through the book in order. Is it sensible to model this sequence of letters by drawing at random with replacement from a box? (There would be 26 tickets in the box, each ticket marked with one letter of the alphabet.) Explain.

5. "The smart professional gambler, when heads comes up four times in a row, will bet that it comes up again. A team that's won six in a row will win seven. *He believes in the percentages.* The amateur bettor will figure that heads can't come up again, that tails is 'due.' He'll bet that a team on a losing streak is 'due' to win. *The amateur believes in the law of averages.*"

   —Jimmy the Greek, *San Francisco Chronicle*, July 2, 1975

   Kerrich's coin (chapter 16) will be tossed until it lands heads four times in a row. Suppose Jimmy the Greek offers 5 to 4 that the coin will land heads on the next toss. (On heads, he pays you $5; on tails, you pay him $4.) Do you take the bet?

*The answers to these exercises are on p. A89.*

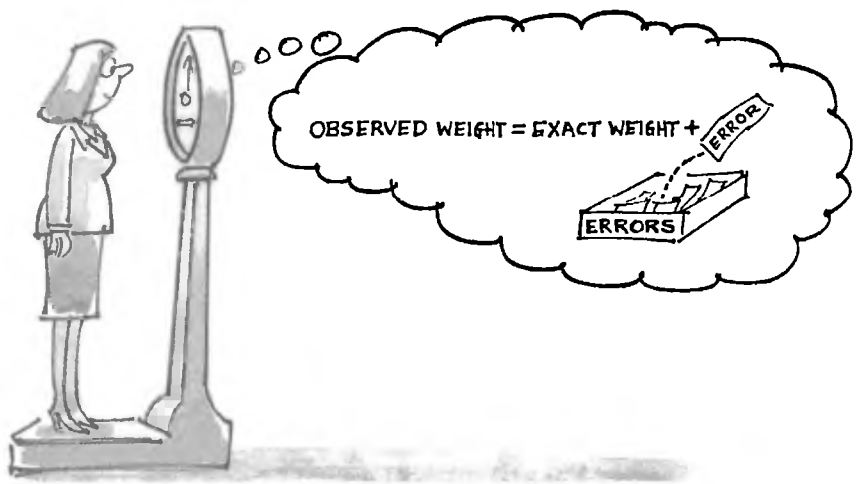
## 3. THE GAUSS MODEL

The box model for measurement error will now be described in more detail. The basic situation is that a series of repeated measurements are made on some quantity. According to the model, each measurement differs from the exact value by a chance error; this error is like a draw made at random from a box of tickets—the *error box*. Successive measurements are done independently and under the same conditions, so the draws from the error box are made with replacement. To capture the idea that the chance errors aren't systematically positive or systematically negative, it is assumed that the average of the numbers in the error box equals 0. This model is named after Carl Friedrich Gauss (Germany, 1777–1855), who worked on measurement error in astronomical data.

> In the Gauss model, each time a measurement is made, a ticket is drawn at random with replacement from the error box. The number on the ticket is the chance error. It is added to the exact value to give the actual measurement. The average of the error box is equal to 0.

In the model, it is the SD of the box which gives the likely size of the chance errors. Usually, this SD is unknown and must be estimated from the data. Take the 100 measurements on NB 10, for example. According to the model, each measurement is around the exact weight, but it is off by a draw from the error box:

$$\begin{aligned}
\text{1st measurement} &= \text{exact weight} + \text{1st draw from error box}\\
\text{2nd measurement} &= \text{exact weight} + \text{2nd draw from error box}\\
&\vdots\\
\text{100th measurement} &= \text{exact weight} + \text{100th draw from error box}
\end{aligned}$$



Carl Friedrich Gauss (Germany, 1777–1855)

Wolff-Leavenworth collection, courtesy of the Syracuse University Art Collection.

With the NB 10 data, the SD of the 100 draws would be a fine estimate for the SD of the error box.[5] The catch is that the draws cannot be recovered from the data, because the exact weight is unknown. However, the variability in the measurements equals the variability in the draws, because the exact weight does not change from measurement to measurement. More technically, adding the exact value to all the errors does not change the SD (pp. 92–93). That is why statisticians use the SD of the measurements when computing the SE. And that completes the reasoning behind the calculation in section 1.[6]

> When the Gauss model applies, the SD of a series of repeated measurements can be used to estimate the SD of the error box. The estimate is good when there are enough measurements.

There may be another way to get at the SD of the error box. When there is a lot of experience with the measurement process, it is better to estimate the SD from all the past data rather than a few current measurements. The reason: the error box belongs to the measurement process, not the thing being measured.

*Example 5.* (Hypothetical.) After making several hundred measurements on NB 10 and finding the SD to be about 6 micrograms, the Bureau's investigators misplace this checkweight. They go out and buy a new one. They measure its weight by exactly the same procedure as for NB 10, and on the same scale. After a week, they accumulate 25 measurements. These average out to 605 micrograms above 10 grams, and the SD is 7 micrograms. Assuming the Gauss model, the new weight is 605 micrograms above 10 grams, give or take about

6 micrograms  7 micrograms  1.2 micrograms  1.4 micrograms.

*Solution.* According to the model, the chance error in each measurement is like a draw from the error box. The error box belongs to the scales, not the weight. The SD of the error box should be estimated by the SD of the large amount of past data on NB 10, not the small amount of current data on the new weight. The SD of the error box is estimated as 6 micrograms. This tells the likely size of the chance error in a single measurement. But the likely size of the chance error in the average of 25 measurements is smaller. The SE for the average is 1.2 micrograms. That is the answer.

In the model, the error box belongs to the scales, not the weight. This seems reasonable for chunks of metal which are similar in size. However, if we change from a 10-gram weight to a 100-gram weight, the error box could change too. And for weights that wiggle around more actively—like babies—the separation between "true value" and "chance error" might not be so convincing.

A final point. The version of the Gauss model presented here makes the assumption that there is no bias in the measuring procedure. When bias is present, each measurement is the sum of three terms:

$$\text{exact value} + \text{bias} + \text{chance error}.$$

Then the SE for the average no longer says how far the average of the measurements is from the exact value, but only how far it is from

$$\text{exact value} + \text{bias}.$$

The methods of this chapter are no help in judging bias. We did not take it into account for the measurements on NB 10, because other lines of reasoning suggest that the bias in precision weighing at the National Bureau of Standards is negligible. In other situations, bias can be more serious than chance errors—and harder to detect.[7]

## Exercise Set C

1. (a) A 10-gram checkweight is being weighed. Assume the Gauss model with no bias. If the exact weight is 501 micrograms above 10 grams, and the number drawn from the error box is 3 micrograms, what would the measurement be?
   (b) Repeat, if the exact weight is 510 micrograms above 10 grams, and the number drawn from the error box is −6 micrograms.

2. The first measurement on NB 10 was 409 micrograms below 10 grams. According to the Gauss model (with no bias),

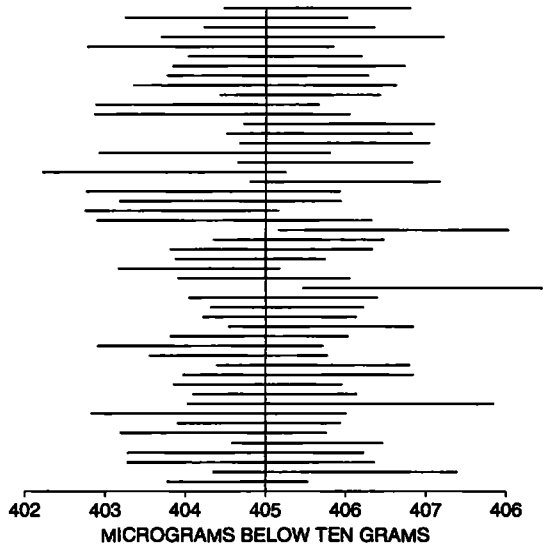
$$409 = \text{exact value} + \text{chance error}.$$

   Can you figure out the numerical value for each of the two terms? Explain briefly.

3. In the Gauss model for the measurements on NB 10, the SD of the error box is _____ 6 micrograms. Fill in the blank using one of the two phrases below, and explain briefly.

   known to be  estimated from the data as

4. The figure below shows the result of a computer simulation: 50 imaginary investigators set out to weigh NB 10, following the procedure used by the Bureau. Each investigator takes 100 measurements and computes the average, the SD, and the SE for the average. The 50 confidence intervals "average ± 2 SE" are plotted at different heights in the figure so they can be seen. In the simulation, the exact weight is taken as 405 micrograms below 10 grams.
   (a) Why do the intervals have different centers?
   (b) Why do they have different lengths?
   (c) How many should cover the exact weight?
   (d) How many do?




5. The Bureau is about to weigh a one-kilogram checkweight 100 times, and take the average of the measurements. They are willing to assume the Gauss model, with no bias, and on the basis of past experience they estimate the SD of the error box to be 50 micrograms.
   (a) The average of all 100 measurements is likely to be off the exact weight by _____ or so.
   (b) The SD of all 100 measurements is likely to be around _____.
   (c) Estimate the probability that the average of all 100 measurements will be within 10 micrograms of the exact weight.

6. Suppose you sent a nominal 10-gram weight off to the Bureau, asking them to weigh it 25 times and tell you the average. They will use the same procedure as on NB 10, where the SD of several hundred measurements was about 6 micrograms. The 25 measurements average out to 307 micrograms above 10 grams, and the SD is about 5 micrograms. Your weight is 307 micrograms above 10 grams, give or take around

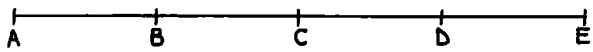
   5 micrograms  6 micrograms  1 microgram  1.2 micrograms

   (You may assume the Gauss model, with no bias.)

7. Twenty-five measurements are made on the speed of light. These average out to 300,007 and the SD is 10, the units being kilometers per second. Fill in the blanks in part (a), then say whether each of (b–f) is true or false. Explain your answers briefly. (You may assume the Gauss model, with no bias.)

   (a) The speed of light is estimated as _____. This estimate is likely to be off by _____ or so.
   (b) The average of all 25 measurements is off 300,007 by 2 or so.
   (c) Each measurement is off 300,007 by 10 or so.
   (d) A 95%-confidence interval for the speed of light is 300,007 ± 4.
   (e) A 95%-confidence interval for the average of the 25 measurements is 300,007 ± 4.
   (f) If a 26th measurement were made, there is a 95% chance that it would be off the exact value for the speed of light by less than 4.

8. A surveyor is measuring the distance between five points A, B, C, D, E. They are all on a straight line. She finds that each of the four distances AB, BC, CD, and DE measures one mile, give or take an inch or so. These four measurements are made independently, by the same procedure.




   The distance from A to E is about four miles; but this estimate is likely to be off by around

   4 inches  2 inches  1 inch  1/2 inch  1/4 inch.

   Explain briefly. (You may assume the Gauss model, with no bias.)

9. The concept of measurement error is often applied to the results of psychological tests. The equation is

   actual test score = true test score + chance error.

   The chance error term reflects accidental factors, like the mood of the subject, or luck. Do you think that the Gauss model applies?

*The answers to these exercises are on pp. A89–90.*

## 4. CONCLUSION

NB 10 is just a chunk of metal. It is weighed on a contraption of platforms, gears, and levers. The results of these weighings have been subjected to a statistical analysis involving the standard error, the normal curve, and confidence intervals. It is the Gauss model which connects the mathematics to NB 10. The chance errors are like draws from a box; their average is like the average of the draws. The number of draws is so large that the probability histogram for the average will follow the normal curve very closely. Without the model there would be no box, no standard error, and no confidence levels.

*Statistical inference* uses chance methods to draw conclusions from data. Attaching a standard error to an average is an example. Now it is always possible to go through the SE procedure mechanically. Many computer programs will do the work for you. It is even possible to label the output as a "standard error."

Do not get hypnotized by the arithmetic or the terminology. The procedure only makes sense because of the square root law. The implicit assumption is that the data are like the results of drawing from a box (an old point, but worth repeating). Many investigators don't pay attention to assumptions. The resulting "standard errors" are often meaningless.[8]

> Statistical inference can be justified by putting up an explicit chance model for the data. No box, no inference.

Parts II and III focused on *descriptive statistics*—drawing diagrams or calculating numbers which summarize data and bring out the salient features. Such techniques can be used very generally, because they do not involve any hidden assumptions about where the data came from. For statistical inference, however, models are basic.

## 5. REVIEW EXERCISES

1. Laser altimeters can measure elevation to within a few inches, without bias, and with no trend or pattern to the measurements. As part of an experiment, 25 readings were made on the elevation of a mountain peak. These averaged out to 81,411 inches, and their SD was 30 inches. Fill in the blanks in part (a), then say whether each of (b–f) is true or false. Explain your answers briefly.
   (a) The elevation of the mountain peak is estimated as _____; this estimate is likely to be off by _____ or so.
   (b) 81,411 ± 12 inches is a 95%-confidence interval for the elevation of the mountain peak.
   (c) 81,411 ± 12 inches is a 95%-confidence interval for the average of the 25 readings.
   (d) There is about a 95% chance that the next reading will be in the range 81,411 ± 12 inches.
   (e) About 95% of the readings were in the range 81,411 ± 12 inches.
   (f) If another 25 readings are made, there is about a 95% chance that their average will be in the range 81,411 ± 12 inches.

2. The first measurement on NB 10 was 409 micrograms below 10 grams. The average of all 100 measurements was 404.6 micrograms below 10 grams, with an SD of 6.4 micrograms; the data are shown in figure 2 on p. 102. Fill in the blanks with a word or phrase. Explain briefly. You may assume the Gauss model, with no bias.
   (a) 404.6 ± 2 × 0.64 is a 95%-confidence interval for the weight of NB 10 because, with 100 draws from the box, the _____ follows the normal curve.
   (b) 409 ± 2 × 6.4 isn't a 95%-confidence interval for the weight of NB 10 because the _____ doesn't follow the normal curve.

3. The speed of light was measured 2,500 times. The average was 299,774 kilometers per second, and the SD was 14 kilometers per second.[9] Assume the Gauss model, with no bias. Find a 95%-confidence interval for the speed of light.

4. In exercise 3, light was timed as it covered a certain distance. The distance was measured 57 times, and the average of these measurements was 1.594265 kilometers. What else do you need to know to decide how accurate this value is?

5. Exercise 4 points to one possible source of bias in the measurements described in exercise 3. What is it?

6. In 2005, the average of the daily maximum temperature at San Francisco airport was 65.8 degrees, and the SD was 7.0 degrees (figure 1, p. 447). Now

$$\sqrt{365} \times 7.0 \approx 134 \text{ degrees}, \quad 134/365 \approx 0.4 \text{ degrees}.$$

True or false: a 95%-confidence interval for the average daily maximum temperature at San Francisco airport is $65.8 \pm 0.8$ degrees. Explain briefly.

7. A calibration laboratory has been measuring a one-kilogram checkweight by the same procedure for several years. They have accumulated several hundred measurements, and the SD of these measurements is 18 micrograms. Someone now sends in a one-kilogram weight to be calibrated by the same procedure. The lab makes 50 measurements on the new weight, which average 78.1 micrograms above a kilogram, and their SD is 20 micrograms. If possible, find a 95%-confidence interval for the value of this new weight. (You may assume the Gauss model, with no bias.)

8. In a long series of trials, a computer program is found to take on average 58 seconds of CPU time to execute, and the SD is 2 seconds. There is no trend or pattern in the data. It will take about _____ seconds of CPU time to execute the program 100 times, give or take _____ seconds or so. (The CPU is the "central processing unit," where the machine does logic and arithmetic.)

9. A machine makes sticks of butter whose average weight is 4.0 ounces; the SD of the weights is 0.05 ounces. There is no trend or pattern in the data. There are 4 sticks to a package.

   (a) A package weighs _____, give or take _____ or so.
   (b) A store buys 100 packages. Estimate the chance that they get 100 pounds of butter, to within 2 ounces.

10. True or false, and explain: "If the data don't follow the normal curve, you can't use the curve to get confidence levels."

11. "All measurements were made twice. If two staff members were present, the duplicate measurements were made by different people. In order to minimize gross errors, discrepancies greater than certain arbitrary limits were measured a third time, and if necessary a fourth, until two measurements were obtained which agreed within the set limits. In cases of discrepancy, the mea-

surers decided which of the three or four results was most 'representative' and designated it for inclusion in the statistical record. In cases of satisfactory agreement, the statistical record was based routinely on the first measurement recorded." Comment briefly.[10]

## 6. SUMMARY

1. According to the *Gauss model* for measurement error, each time a measurement is made, a ticket is drawn at random with replacement from the *error box*. The number on the ticket is the chance error. It is added to the exact value of the thing being measured, to give the actual measurement. The average of the error box is equal to 0. Here, bias is assumed to be negligible.

2. When the Gauss model applies, the SD of many repeated measurements is an estimate for the SD of the error box. This tells the likely size of the chance error in an individual measurement.

3. The average of the series is more precise than any individual measurement, by a factor equal to the square root of the number of measurements. The calculation assumes that the data follow the Gauss model.

4. An approximate confidence interval for the exact value of the thing being measured can be found by going the right number of SEs either way from the average of the measurements; the confidence level is taken from the normal curve. The approximation is good provided the Gauss model applies, with no bias, and there are enough measurements.

5. With the Gauss model, the chance variability is in the measuring process, not the thing being measured. The word "confidence" is to remind you of this.

6. If the model does not apply, neither does the procedure for getting confidence intervals. In particular, if there is any trend or pattern in the data, the formulas may give silly answers.

7. *Statistical inference* is justified in terms of an explicit *chance model* for the data.

# 25

# Chance Models in Genetics

*I shall never believe that God plays dice with the world.*
—ALBERT EINSTEIN (1879–1955)

## 1. HOW MENDEL DISCOVERED GENES

This chapter is hard, and it can be skipped without losing the thread of the argument in the book. It is included for two reasons:

- Mendel's theory of genetics is great science.
- The theory shows the power of simple chance models in action.

In 1865, Gregor Mendel published an article which provided a scientific explanation for heredity, and eventually caused a revolution in biology.[1] By a curious twist of fortune, this paper was ignored for about thirty years, until the theory was simultaneously rediscovered by three men, Correns in Germany, de Vries in Holland, and Tschermak in Austria. De Vries and Tschermak are now thought to have seen Mendel's paper before they published, but Correns apparently found the idea by himself.

Mendel's experiments were all carried out on garden peas; here is a brief account of one of these experiments. Pea seeds are either yellow or green. (As the phrase suggests, seed color is a property of the seed itself,[2] and not of the parental plant: indeed, one parent often has seeds of both colors.) Mendel bred a pure yellow strain, that is, a strain in which every plant in every generation had

Gregor Mendel (Austria, 1822–1884).
From the collection of the Moravian Museum, Brno.

only yellow seeds. Separately, he bred a pure green strain. He then crossed plants of the pure yellow strain with plants of the pure green strain. For instance, he used pollen from the yellows to fertilize ovules on plants of the green strain. (The alternative method, using pollen from the greens to fertilize plants of the yellow strain, gave exactly the same results.) The seeds resulting from a yellow-green cross, and the plants into which they grow, are called *first-generation hybrids.* First-generation hybrid seeds are all yellow, indistinguishable from seeds of the pure yellow strain. The green seems to have disappeared completely.

These first-generation hybrid seeds grew into first-generation hybrid plants which Mendel crossed with themselves, producing *second-generation hybrid* seeds. Some of these second-generation seeds were yellow, but some were green. So the green disappeared for one generation, but reappeared in the second. Even more surprising, the green reappeared in a definite, simple proportion. Of the second-generation hybrid seeds, about 75% were yellow and 25% were green.

What is behind this regularity? To explain it, Mendel postulated the existence of the entities now called *genes.*[3] According to Mendel's theory, there were two different variants of a gene which paired up to control seed color. They will be denoted here by *y* (for yellow) and *g* (for green). It is the gene-pair in the seed—not the parent—which determines what color the seed will be, and all the cells making up a seed contain the same gene-pair.

There are four different gene-pairs: $y/y$, $y/g$, $g/y$, and $g/g$. Gene-pairs control seed color by the rule

- $y/y$, $y/g$, and $g/y$ make yellow,
- $g/g$ makes green.

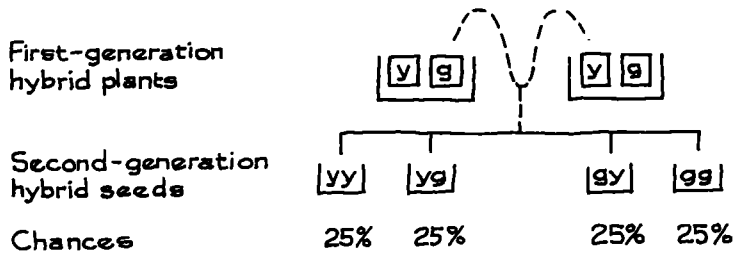
As geneticists say, $y$ is *dominant* and $g$ is *recessive*. This completes the first part of the model.

Now the seed grows up and becomes a plant. All the cells in this plant will also carry the seed's color gene-pair—with one exception. Sex cells, either sperm or eggs, contain only one gene of the pair.[4] For instance, a plant whose ordinary cells contain the gene-pair $y/y$ will produce sperm cells containing the gene $y$. Similarly, the plant will produce egg cells containing the gene $y$. On the other hand, a plant whose ordinary cells contain the gene-pair $y/g$ will produce some sperm cells containing the gene $y$, and some sperm cells containing the gene $g$. In fact, half its sperm cells will contain $y$, and the other half will contain $g$; half its eggs will contain $y$, the other half will contain $g$.

This model accounts for the experimental results. Plants of the pure yellow strain have the color gene-pair $y/y$, so the sperm and eggs all just contain the gene $y$. Similarly, plants of the pure green strain have the gene-pair $g/g$, so their pollen and ovules just contain the gene $g$. Crossing a pure yellow with a pure green amounts for instance to fertilizing a $g$-egg by a $y$-sperm, producing a fertilized cell having the gene-pair $y/g$. This cell reproduces itself and eventually becomes a seed, in which all the cells have the gene-pair $y/g$ and are yellow in color. The model has explained why all first-generation hybrid seeds are yellow, and none are green.

What about the second generation? A first-generation hybrid seed grows into a first-generation hybrid plant, with the gene-pair $y/g$. This plant produces sperm cells, of which half will contain the gene $y$ and the other half will contain the gene $g$. The plant also produces eggs, of which half will contain $y$ and the other half will contain $g$. When two first-generation hybrids are crossed, a resulting second-generation hybrid seed gets one gene at random from each parent—because the seed is formed by the random combination of a sperm cell and an egg. From the point of view of the seed, it's as if one ticket was chosen at random from each of two boxes. In each box, half the tickets are marked $y$ and the other half are marked $g$. The tickets are the genes, and there is one box for each parent (figure 1).

Figure 1. Mendel's chance model for the genetic determination of seed-color: one gene is chosen at random from each parent. The chance of each combination is shown. (The sperm gene is listed first; in terms of seed color, the combinations $y/g$ and $g/y$ are not distinguishable after fertilization.[5])



As shown in Figure 1, the seed has a 25% chance to get a gene-pair with two $g$'s and be green. The seed has a 75% chance to get a gene-pair with one or two $y$'s and be yellow. The number of seeds is small by comparison with the number of pollen grains, so the selections for the various seeds are essentially independent. The conclusion: the color of second-generation hybrid seeds will be determined as if by a sequence of draws with replacement from the box

| yellow | yellow | yellow | green |
|---|---|---|---|

And that is how the model accounts for the reappearance of green in the second generation, for about 25% of the seeds.

Mendel made a bold leap from his experimental evidence to his theoretical conclusions. His reconstruction of the chain of heredity was based entirely on statistical evidence of the kind discussed here. And he was right. Modern research in genetics and molecular biology is uncovering the chemical basis of heredity, and has provided ample direct proof for the existence of Mendel's hypothetical entities. As we know today, genes are segments of DNA on chromosomes—the dark patches in Figure 2 on the next page.

Essentially the same mechanism of heredity operates in all forms of life, from dolphins to fruit flies. So the genetic model proposed by Mendel unlocks one of the great mysteries of life. How is it that a pea-seed always produces a pea, and never a tomato or a whale? Furthermore, the answer turns out to involve chance in a crucial way, despite Einstein's quote at the opening of the chapter.

## Exercise Set A

1. In some experiments, a first-generation hybrid pea is "back-crossed" with one parent. If a $y/g$ plant is crossed with a $g/g$, about what percentage of the seeds will be yellow? Of 1,600 such seeds, what is the chance that over 850 will be yellow?

2. Flower color in snapdragons is controlled by one gene-pair. There are two variants of the gene, $r$ (for red) and $w$ (for white). The rules are:

   $r/r$ makes red flowers,
   $r/w$ and $w/r$ make pink flowers,
   $w/w$ makes white flowers.

   So neither $r$ nor $w$ is dominant. Their effects are *additive*, like mixing red paint with white paint.

   (a) Work out the expected percentages of red-, pink-, and white-flowered plants resulting from the following crosses: white × red, white × pink, pink × pink.
   (b) With 400 plants from pink × pink crosses, what is the chance that between 190 and 210 will be pink-flowered?

3. Snapdragon leaves come in three widths: wide, medium, and narrow. In breeding trials, the following results are obtained:

   wide × wide → 100% wide
   wide × medium → 50% wide, 50% medium
   wide × narrow → 100% medium
   medium × medium → 25% narrow, 50% medium, 25% wide.

[Exercise continues on p. 463.]

Figure 2. Photomicrograph. These cells are from the root tip of a pea plant, and are magnified about 2,000 times. The cell shown in the center is about to divide. At this stage, each individual chromosome consists of two identical pieces, lying side by side. There are fourteen chromosomes arranged in seven homologous pairs, indicated by the Roman numerals from I to VII. The gene-pair controlling seed-color is located on chromosome pair I, one of the genes being on each chromosome.[6]

Source: New York State Agriculture Experiment Station, Geneva, N.Y.

(a) Can you work out a genetic model to explain these results?

(b) What results would you expect from each of the following crosses: narrow × narrow, narrow × medium?

4. Eye color in humans is determined by one gene-pair, with brown dominant and blue recessive. In a certain family, the husband had a blue-eyed father; he himself has brown eyes. The wife has blue eyes. They plan on having three children. What is the chance that all three will have brown eyes? (It is better to work this out exactly rather than using the normal approximation.)

*The answers to these exercises are on p. A90.*

## 2. DID MENDEL'S FACTS FIT HIS MODEL?

Mendel's discovery ranks as one of the greatest in science. Today, his theory is amply proved and extremely powerful. But how good was his own experimental proof? Did Mendel's data prove his theory? Only too well, answered R. A. Fisher:

> ...the general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions. The data have evidently been sophisticated systematically, and after examining various possibilities, I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made.[7]

Leave the gardener aside for now. Fisher is saying that Mendel's data were fudged. The reason: Mendel's observed frequencies were uncomfortably close to his expected frequencies, much closer than ordinary chance variability would permit.

In one experiment, for instance, Mendel obtained 8,023 second-generation hybrid seeds. He expected $\frac{1}{4} \times 8{,}023 \approx 2{,}006$ of them to be green, and observed 2,001, for a discrepancy of 5. According to his own chance model, the data on seed color are like the results of drawing 8,023 times with replacement from the box

| yellow | yellow | yellow | green |
|---|---|---|---|

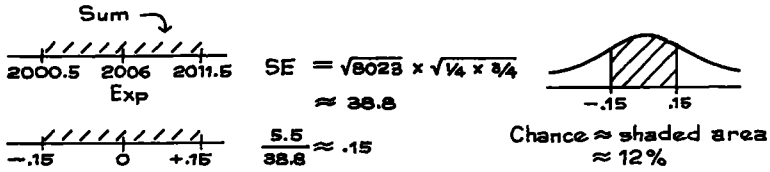
In this model, what is the chance of observing a discrepancy of 5 or less between the number of greens and the expected number? In other words, what is the probability that the number of greens will be

$$\text{between } \tfrac{1}{4} \times 8{,}023 - 5 \approx 2{,}001 \text{ and } \tfrac{1}{4} \times 8{,}023 + 5 \approx 2{,}011?$$

That is like drawing 8,023 times with replacement from the box

| 0 | 0 | 0 | 1 |
|---|---|---|---|

and asking for the chance that the sum will be between 2,001 and 2,011 inclusive. This chance can be estimated using the normal approximation, keeping track of the edges of the rectangles, as on p. 317.




About 88% of the time, chance variation would cause a discrepancy between Mendel's expectations and his observations greater than the one he reported.

By itself, this evidence is not very strong. The trouble is, every one of Mendel's experiments (with an exception to be discussed below) shows this kind of unusually close agreement between expectations and observations. Using the $\chi^2$-test to pool the results (chapter 28), Fisher showed that the chance of agreement as close as that reported by Mendel is about four in a hundred thousand. To put this another way, suppose millions of scientists were busily repeating Mendel's experiments. For each scientist, imagine measuring the discrepancy between his observed frequencies and the expected frequencies by the $\chi^2$-statistic. Then by the laws of chance, about 99,996 out of every 100,000 of these imaginary scientists would report a discrepancy between observations and expectations greater than the one reported by Mendel. That leaves two possibilities:

- either Mendel's data were massaged
- or he was pretty lucky.

The first is easier to believe.

One aspect of Fisher's argument deserves more attention. However, the discussion is technical, and readers can skip to the beginning of the next section. Mendel worked with six characteristics other than seed color. One of them, for instance, was the shape of the pod, which was either inflated (the dominant form) or constricted (the recessive form). The hereditary mechanism is very similar to that for seed color. Pod shape is controlled by one gene-pair. There are two variants of the shape-gene, denoted by $i$ (inflated) and $c$ (constricted). The gene $i$ is dominant, so $i/i$ or $i/c$ or $c/i$ make inflated pods, and $c/c$ makes constricted pods. (The gene-pair controlling seed color acts independently of the pair controlling pod shape.)

There is one difference between seed color and pod shape. Pod shape is a characteristic of the parent plant, and is utterly unaffected by the fertilizing pollen. Thus, if a plant of a pure strain showing the recessive constricted form of seed pods is fertilized with pollen from a plant of pure strain showing the dominant inflated form, all the resulting seed pods will have the recessive constricted form. But when the seeds of this cross grow up into mature first-generation hybrid plants and make their own seed pods, they will all exhibit the dominant inflated form.

If first-generation hybrids are crossed with each other, of the second-generation hybrid plants about 3/4 will exhibit the dominant form and 1/4 the recessive form. As Figure 1 shows, of the second-generation hybrid plants

with the dominant inflated form, about

$$\frac{25\%}{25\% + 25\% + 25\%} = \frac{1}{3}$$

should be $i/i$'s and the other 2/3 should be $i/c$ or $c/i$. Mendel checked this out on 600 plants, finding 201 $i/i$'s, a result too close to the expected 200 for comfort.[8] (The chance of such close agreement is only 10%.)

But worse is yet to come. You can't tell the $i/i$'s from the $i/c$'s or $c/i$'s just by looking, the appearances are identical. So how did Mendel classify them? Well, if undisturbed by naturalists, a pea plant will pollinate itself. So Mendel took his second-generation hybrid plants showing the dominant inflated form, and selected 600 at random. He then raised 10 offsprings from each of his selected plants. If the plant bred true and all 10 offsprings showed the dominant inflated form, he classified it as $i/i$. If the plant produced any offspring showing the recessive constricted form, he classified it as $i/c$ or $c/i$.

There is one difficulty with this scheme, which Mendel seems to have overlooked. As Figure 1 shows, the chance that the offspring of a self-fertilized $i/c$ will contain at least one dominant gene $i$, and hence show the dominant inflated form, is 3/4. So the chance that 10 offsprings of an $i/c$ crossed with itself will all show the dominant form is $(3/4)^{10} \approx 6\%$. Similarly for $c/i$'s. The expected frequency of plants classified as $i/i$ is therefore a bit higher than 200, because about 6% of the 400 $i/c$'s and $c/i$'s will be incorrectly classified as $i/i$. Indeed, the expected frequency of plants classified as $i/i$—correctly or incorrectly—is

$$200 + 0.06 \times 400 = 224.$$

Mendel's observed frequency (201 classified as $i/i$) is rather too far from expectation: the chance of such a large discrepancy is only about 5%. As Fisher concludes, "There is no easy way out of the difficulty."

## 3. THE LAW OF REGRESSION

This section is difficult, and readers can skip to the next section. Part III discussed Galton's work on heredity, and presented his finding that on the average a child is halfway between the parent and the average. In 1918, Fisher proposed a chance model[9] based on Mendel's ideas, which explained Galton's finding on regression as well as the approximate normality of many biometric characteristics like height (chapter 5). The model can be made quite realistic at the expense of introducing complications. This section begins with a stripped-down version which is easier to understand; later, some refinements will be mentioned. The model will focus on heights, although exactly the same argument could be made for other characteristics. The first assumption in the model is

(1) height is controlled by one gene-pair.

The second assumption is

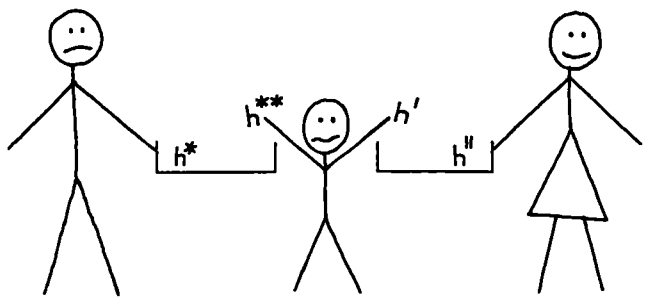
(2) the genes controlling height act in a *purely additive* way.

The symbols $h^*$, $h^{**}$, $h'$, $h''$ will be used to denote four typical variants of the height-gene. (Variants of a gene are called "alleles.") Assumption (2) means, for instance, that $h^*$ always contributes a fixed amount to an individual's height, whether it is combined with another $h^*$, or with an $h'$, or with any other variant of the height gene. These genes act very differently from the $y$'s and $g$'s controlling seed color in Mendel's peas: $g$ contributes green to the seed color when it is combined with another $g$, but when it is combined with a $y$ it has no effect. The height genes are more like the snapdragon genes in exercises 2 and 3 on p. 461 above.

With assumption (2), each gene contributes a fixed amount to an individual's height. This contribution (say in inches) will be denoted by the same letter as used to denote the gene, but in capitals. Thus, an individual with the gene-pair $h^*/h'$ will have height equal to the sum $H^* + H'$. In the first instance, the letters refer to the genes; in the second, to the contributions to height.

Fisher assumed with Mendel that a child gets one gene of the pair controlling height at random from each parent (figure 3). To be more precise, the father has a gene-pair controlling height, and so does the mother. Then one gene is drawn at random from the father's pair and one from the mother's pair to make up the child's pair.

Figure 3. The simplified Mendel-Fisher model for the genetic determination of height. Height is controlled by one gene-pair, with purely additive genetic effects. One gene is drawn at random from each parent's gene-pair to make up the child's gene-pair.



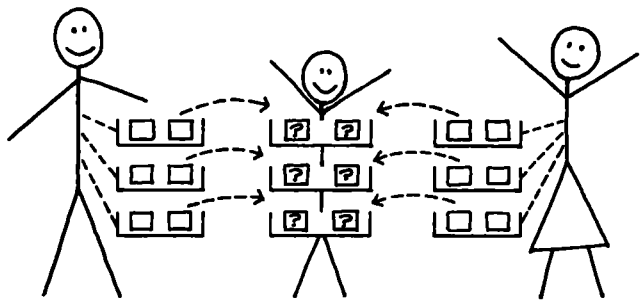

For the sake of argument, suppose the father has the gene-pair $h^*/h^{**}$, and the mother has the gene-pair $h'/h''$. The child has chance $1/2$ to get $h^*$ and chance $1/2$ to get $h^{**}$ from the father. Therefore, the father's expected contribution to the child's height is $\frac{1}{2}H^* + \frac{1}{2}H^{**} = \frac{1}{2}(H^* + H^{**})$, namely one-half the father's height. Similarly, the mother's expected contribution equals one-half her height. If you take a large number of children of parents whose father's height is fixed at one level, and mother's height is fixed at another level, the average height of these children must be about equal to

(3) $\frac{1}{2}$(father's height + mother's height).

The expression (3) is called the *mid-parent height*. For instance, with many families where the father is 72 inches tall and the mother is 68 inches tall, the mid-parent height is $\frac{1}{2}(72 + 68) = 70$, and on the average the children will be about 70 inches tall at maturity, give or take a small chance error. This is the biological explanation for Galton's law of regression to mediocrity (pp. 169–173).

The assumption (1), that height is controlled by one gene-pair, isn't really needed in the argument; it was made to avoid complicated sums. If three gene-pairs are involved, you only have to assume additivity of the genetic effects and randomness in drawing one gene from each pair for the child (figure 4).

Figure 4. The simplified Mendel-Fisher model for the genetic determination of height, assuming three gene-pairs with purely additive effects. One gene is drawn at random from each gene-pair of each parent to make up the corresponding gene-pair of the child.



So far, the model has not taken into account sex differences in height. One way to get around this is by "adjusting" women's heights, increasing them by around 8% so that women are just as tall as men—at least in the equations of the model. More elegant (and more complicated) methods are available too.

*How well does the model fit?* For the Pearson–Lee study (p. 119), the regression of son's height on parents' heights was approximately[10]

$$\text{(4)} \qquad \text{estimated son's ht.} = 15'' + 0.8 \times \frac{\text{father's ht.} + 1.08 \times \text{mother's ht.}}{2}$$

The regression coefficient of 0.8 is noticeably lower than the 1.0 predicted by a purely additive genetic model. Some of the discrepancy may be due to environmental effects, and some to nonadditive genetic effects. Furthermore, the sons averaged 1 inch taller than the fathers. This too cannot be explained by a purely additive genetic model.[11]

The regression of son's height on father's height was very nearly

$$\text{(5)} \qquad \text{estimated son's height} = 35'' + 0.5 \times \text{father's height.}$$

Equation (5) can be derived from equation (3) in the additive model, by assuming that there is no correlation between the heights of the parents.[12] Basically, however, this is a case of two mistakes cancelling. The additive model is a bit off,

and the heights of parents are somewhat correlated; but these two facts work in opposite directions, and balance out in equation (5).

*Technical note.* To derive equation (3) from the model, no assumptions are necessary about the independence of draws from different gene-pairs; all that mattered was each gene having a 50% chance to get drawn. No assumptions are necessary about statistical relationships between the genes in the different parents (such as independence). And no assumptions are necessary about the distribution of the genes in the population (like equilibrium).

## 4. AN APPRECIATION OF THE MODEL

Genetics represents one of the most satisfying applications of statistical methods. To review the development, Mendel found some striking empirical regularities—like the reappearance of a recessive trait in one-fourth of the second-generation hybrids. He made up a chance model involving what are now called genes to explain his results. He discovered these entities by pure reasoning—he never saw any. Independently, Galton and Pearson found another striking empirical regularity: on the average, a son is halfway between his father and the overall average for sons.

At first sight, the Galton-Pearson results look very different from Mendel's, and it is hard to see how they can both be explained by the same biological mechanism. But Fisher was able to do it. He explained why the average height of children equaled mid-parent height, and even why there are deviations from average. These deviations are caused by chance variation, when genes are chosen at random to pass from the parents to the children.

Chance models are now used in many fields. Usually, the models only assert that certain entities behave as if they were determined by drawing tickets at random from a box, and little effort is spent establishing a physical basis for the claim of randomness. Indeed, the models seldom say explicitly what is like the box, or what is like the tickets.

The genetic model is quite unusual, in that it answers such questions. There are two main sources of randomness in the model:

(i) the random allotment of chromosomes (one from each pair) to sex cells;

(ii) the random pairing of sex cells to produce the fertilized egg.

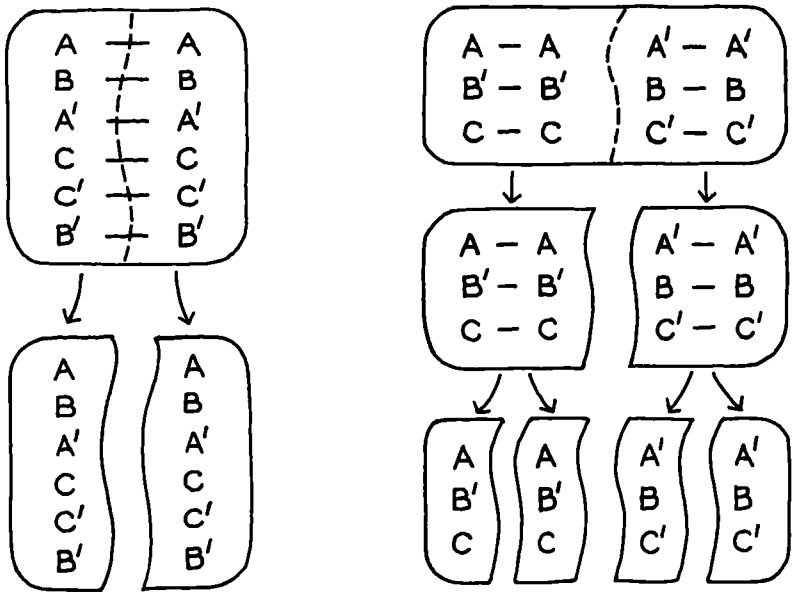
The two sources of randomness will now be discussed in more detail.

Chromosomes naturally come in *homologous* pairs. The one matching $C$ is denoted $C'$; the chromosomes $C$ and $C'$ are similar, but not identical. A gene-pair has one gene located on each chromosome of a homologous chromosome-pair. A body cell can divide to form other cells. As a preliminary step, each chromosome in the parent cell doubles itself, as shown in figure 2 and (schematically) in figure 5. When chromosome $C$ is in this doubled condition, it will be denoted $C$-$C$. The two pieces are chemically identical and loosely joined together.

The production of ordinary body cells is shown in figure 5a. The parent cell splits in two. Each fragment becomes a separate cell with one-half of each doubled chromosome, winding up with exactly the same complement of chromo-

Figure 5. Production of sex-cells and body-cells by splitting. Chromosomes are denoted here by capital letters, like $C$. Chromosomes come in homologous pairs, as indicated by primes. Thus, $C$ and $C'$ form a homologous pair; they are chemically similar but not identical. As a preliminary to splitting, a cell doubles all its chromosomes. When doubled, $C$ will be denoted by $C$-$C$. The two pieces are chemically identical, and loosely attached. Similarly, $C'$ doubles to $C'$-$C'$.[13]

(a) Splitting to make body cells (b) Splitting to make sex cells




somes as the parent cell (before doubling). There is nothing random about the resulting chromosomes—it is a matter of copying the whole set. Homologous chromosomes are not treated in any special way.

The production of sex cells is shown in figure 5b. The doubled chromosomes move into position, with one doubled chromosome from each homologous pair on opposite sides of the line along which the cell will split (top of figure 5b). Which side of the line? This seems to be random, like coin-tossing. Sometimes one side, sometimes the other, just as a coin sometimes lands heads, sometimes tails. In the model, the choice of side is assumed to be random.

The cell then splits as shown in the middle of figure 5b. Each fragment contains doubled chromosomes—but only one chromosome of each homologous pair is represented. Finally, each of these fragments splits again, as shown at the bottom of figure 5b, and the results of the second split are the sex cells.[14] The lining-up of the homologous pairs (top of figure 5b) is a critical step. The sex cell contains ordinary, undoubled chromosomes—but only one chromosome out of each homologous pair. Which one? One chosen at random. This is one physical source of randomness in Mendelian genetics.

A fertilized egg results from the union of one male sex cell and one female, out of the many which are produced. Which ones? This seems to be random, like

drawing tickets at random from a box. In the model, the pairing is assumed to be random. This is the second main physical source of randomness in Mendelian genetics.

When thinking about any other chance model, it is good to ask two questions:

- What are the physical entities which are supposed to act like the tickets and the box?
- Do they really act like that?

## 5. REVIEW EXERCISES

1. Mendel discovered that for peas, the unripe pods are green or yellow. Their color is controlled by one gene-pair, with variants $g$ for green and $y$ for yellow, $g$ being dominant. In a set of breeding trials, plants with known pod color but unknown genetic makeup were crossed. The results are tabulated below.[15] For each line of the table, guess the genetic makeup of the parents:
   (i) $g/g$ (ii) $y/g$ or $g/y$ (iii) $y/y$

| *Pod color of parents* | *Number of progeny with* | |
|---|---|---|
| | *green pods* | *yellow pods* |
| green × yellow | 82 | 78 |
| green × green | 118 | 39 |
| yellow × yellow | 0 | 50 |
| green × yellow | 74 | 0 |
| green × green | 90 | 0 |

2. Mendel found that pea seeds were either smooth or wrinkled. He bred a pure smooth strain and a pure wrinkled strain. Interbreeding these two strains gave first-generation hybrids, which all turned out to be smooth. Mendel crossed the first-generation hybrids with themselves to get second-generation hybrids; of 7,324 second-generation hybrid plants, 5,474 turned out to be smooth, and 1,850 were wrinkled. Make up a genetic model to account for these results. In the model, what is the chance of agreement between the expected frequency of smoothies and the observed frequency as close as that reported by Mendel?

3. Peas flower at three different times: early, intermediate, and late.[16] Breeding trials gave the following results:

   early × early → early
   early × late → intermediate
   late × late → late.

   Suppose you have 2,500 plants resulting from the cross

   intermediate × intermediate.

   What is the chance that 1,300 or more are intermediate-flowering?

4. In humans, there is a special chromosome-pair which determines sex. Males

have the pair $X$-$Y$, while females have the pair $X$-$X$. A child gets one $X$-chromosome automatically, from the mother; from the father, it has half a chance to get an $X$-chromosome and be female, half a chance to get $Y$ and be male. Some genes are carried only on the $X$-chromosome: these are said to be *sex-linked*. An example is the gene for male-pattern baldness. (Color blindness and hemophilia are other sex-linked characteristics; the model for baldness is simplified.)

(a) If a man has a bald father, is he more likely to go bald?
(b) If a man's maternal grandfather was bald, is he more likely to go bald?

Explain briefly.

5. Sickle-cell anemia is a genetic disease. In the U.S., it is especially prevalent among blacks: one person in four hundred suffers from it. The disease is controlled by one gene-pair, with variants $A$ and $a$, where $a$ causes the disease but is recessive:

$A/A$, $A/a$, $a/A$—healthy person
$a/a$—sickle-cell anemia.

(a) Suppose one parent has the gene-pair $A/A$. Can the child have sickle-cell anemia? How?
(b) Suppose neither parent has sickle-cell anemia. Can the child have it? How?
(c) Suppose both parents have sickle-cell anemia. Can the child avoid having it? How?

## 6. SUMMARY AND OVERVIEW

1. Whenever reproduction is sexual, the mechanism of heredity is based on gene-pairs. The offspring gets one gene of each pair drawn at random from the corresponding pair in the maternal organism, and one at random from the corresponding pair in the paternal organism. The two genes in a pair are very similar, but not identical.

2. Gene-pairs can control biological characteristics in several ways. One way is *dominance*. In this case, there may be only two varieties (alleles) of the gene, say $d$ and $r$. The gene-pairs $d/d$, $d/r$, $r/d$ all produce the dominant characteristic, while $r/r$ produces the recessive characteristic. (Seed color in peas is an example.) Another way is *additivity*. In this case, each variety of the gene has an effect, and the effect of the gene-pair is the sum of the individual effects of the two genes in the pair. (Flower color in snapdragons is an example.)

3. Fisher showed that Galton's law of regression was a mathematical consequence of Mendel's rules, assuming additive genetic effects.

4. The genetic model explains (at least part of the reason) why children resemble their parents, and also why they differ.

5. This part of the book discussed two chance models: the Gauss model for measurement error and Mendel's model for genetics. These models show how complicated phenomena can be analyzed using the techniques built up in parts II and IV–VI.

6. Chance models are now used in many fields. Usually, the models only assert that some things behave like tickets drawn at random from a box. The genetic model is unusual, because it establishes a physical basis for the claim of randomness.

7. In the next part of the book, we will look at some of the procedures statisticians use for testing models.

PART VIII

# Tests of Significance

# 26

# Tests of Significance

*Who would not say that the glosses [commentaries on the law] increase doubt and ignorance? It is more of a business to interpret the interpretations than to interpret the things.*

—MICHEL DE MONTAIGNE (FRANCE, 1533–1592)[1]

## 1. INTRODUCTION

Was it due to chance, or something else? Statisticians have invented *tests of significance* to deal with this sort of question. Nowadays, it is almost impossible to read a research article without running across tests and significance levels. Therefore, it is a good idea to find out what they mean. The object in chapters 26 through 28 is to explain the ideas behind tests of significance, and the language. Some of the limitations will be pointed out in chapter 29. This section presents a hypothetical example, where the arguments are easier to follow.

Suppose two investigators are arguing about a large box of numbered tickets. Dr. Nullsheimer says the average is 50. Dr. Altshuler says the average is different from 50. Eventually, they get tired of arguing, and decide to look at some data. There are many, many tickets in the box, so they agree to take a sample—they'll draw 500 tickets at random. (The box is so large that it makes no difference whether the draws are made with or without replacement.) The average of the draws turns out to be 48, and the SD is 15.3.

*Dr. Null* The average of the draws is nearly 50, just like I thought it would be.

*Dr. Alt* The average is really below 50.

*Dr. Null* Oh, come on, the difference is only 2, and the SD is 15.3. The difference is tiny relative to the SD. It's just chance.

*Dr. Alt* Hmmm. Dr. Nullsheimer, I think we need to look at the SE not the SD.

*Dr. Null* Why?

*Dr. Alt* Because the SE tells us how far the average of the sample is likely to be from its expected value—the average of the box.

*Dr. Null* So, what's the SE?

*Dr. Alt* Can we agree to estimate the SD of the box as 15.3, the SD of the data?

*Dr. Null* I'll go along with you there.

*Dr. Alt* OK, then the SE for the sum of the draws is about $\sqrt{500} \times 15.3 \approx 342$. Remember the square root law.

*Dr. Null* But we're looking at the average of the draws.

*Dr. Alt* Fine. The SE for the average is $342/500 \approx 0.7$.

*Dr. Null* So?

*Dr. Alt* The average of the draws is 48. You say it ought to be 50. If your theory is right, the average is about 3 SEs below its expected value.

*Dr. Null* Where did you get the 3?

*Dr. Alt* Well,

$$\frac{48-50}{0.7} \approx -3.$$

*Dr. Null* You're going to tell me that 3 SEs is too many SEs to explain by chance.

*Dr. Alt* That's my point. You can't explain the difference by chance. The difference is real. In other words, the average of tickets in the box isn't 50, it's some other number.

*Dr. Null* I thought the SE was about the difference between the sample average and its expected value.

*Dr. Alt* Yes, yes. But the expected value of the sample average *is* the average of the tickets in the box.

Our first pass at testing is now complete. The issue in the dialog comes up over and over again: one side thinks a difference is real but the other side might say it's only chance. The "it's only chance" attack can fended off by a calculation, as in the dialog. This calculation is called a *test of significance*. The key idea: if an observed value is too many SEs away from its expected value, that is hard to explain by chance. Statisticians use rather technical language when making this sort of argument, and the next couple of sections will introduce the main terms: *null hypothesis, alternative hypothesis, test statistic*, and *P-value*.[2]

## Exercise Set A

1. Fill in the blanks. In the dialog—

(a) The SD of the box was _____ 15.3. Options: known to be, estimated from the data as
(b) The 48 is an _____ value. Options: observed, expected

2. In the dialog, suppose the 500 tickets in the sample average 48 but the SD is 33.6. Who wins now, Dr. Null or Dr. Alt?

3. In the dialog, suppose 100 tickets are drawn, not 500. The sample average is 48 and the SD is 15.3. Who wins now, Dr. Null or Dr. Alt?

4. A die is rolled 100 times. The total number of spots is 368 instead of the expected 350. Can this be explained as a chance variation, or is the die loaded?

5. A die is rolled 1,000 times. The total number of spots is 3,680 instead of the expected 3,500. Can this be explained as a chance variation, or is the die loaded?

*The answers to these exercises are on p. A91.*

## 2. THE NULL AND THE ALTERNATIVE

In the example of the previous section, there was sample data for 500 tickets. Both sides saw the sample average of 48. In statistical shorthand, the 48 was "observed." The argument was about the interpretation: what does the sample tell us about the other tickets in the box? Dr. Altshuler claimed that the observed difference was "real." That may sound odd. Of course 48 is different from 50. But the question was whether the difference just reflected chance variation—as Dr. Nullsheimer thought—or whether the average for all the tickets in the box was different from 50, as Dr. Altshuler showed.

In statistical jargon, the null hypothesis and the alternative hypothesis are statements about the box, not just the sample. Each hypothesis represents one side of the argument.

- Null hypothesis—the average of the box equals 50.
- Alternative hypothesis—the average of the box is less than 50.

In the dialog, Dr. Nullsheimer is defending the null hypothesis. According to him, the average of the box was 50. The sample average turned out to be lower than 50 just by the luck of the draw. Dr. Altshuler was arguing for the alternative hypothesis. She thinks the average of the box is lower than 50. Her argument in a nutshell: the sample average is so many SEs below 50 that Dr. Nullsheimer almost has to be wrong. Both sides agreed about the data. They disagreed about the box.

> The null hypothesis corresponds to the idea that an observed difference is due to chance. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box, corresponding to the idea that the observed difference is real.

The terminology may be unsettling. The "alternative hypothesis" is often what someone sets out to prove. The "null hypothesis" is then an alternative (and dull) explanation for the findings, in terms of chance variation. However, there is no way out: the names are completely standard.

Every legitimate test of significance involves a box model. The test gets at the question of whether an observed difference is real, or just chance variation. A real difference is one that says something about the box, and isn't just a fluke of sampling. In the dialog, the argument was about all the numbers in the box, not the 500 numbers in the sample. A test of significance only makes sense in a debate about the box. This point will be discussed again, in section 4 of chapter 29.

## Exercise Set B

1. In order to test a null hypothesis, you need
   (i) data
   (ii) a box model for the data
   (iii) both of the above
   (iv) none of the above

2. The _____ hypothesis says that the difference is due to chance but the _____ hypothesis says that the difference is real. Fill in the blanks. Options: null, alternative.

3. In the dialog of section 1, Dr. Alt needed to make a test of significance because
   (i) she knew what was in the box but didn't know how the data were going to turn out, or
   (ii) she knew how the data had turned out but didn't know what was in the box.
   Choose one option, and explain briefly.

4. In the dialog, the null hypothesis says that the average of the _____ is 50. Options: sample, box.

5. One hundred draws are made at random with replacement from a box. The average of the draws is 22.7, and the SD is 10. Someone claims that the average of the box equals 20. Is this plausible?

*The answers to these exercises are on p. A91.*

## 3. TEST STATISTICS AND SIGNIFICANCE LEVELS

In the dialog of section 1, Dr. Altshuler temporarily assumed the null hypothesis to be right (the average of the box is 50). On this basis, she calculated how many SEs away the observed value of the sample average was from its expected value:

$$\frac{48 - 50}{0.7} \approx -3.$$

This is an example of a *test statistic*.

> A test statistic is used to measure the difference between the data and what is expected on the null hypothesis.

Dr. Altshuler's test statistic is usually called $z$:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

Tests using the $z$-statistic are called *$z$-tests*. Keep the interpretation in mind.

> $z$ says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis.

It is the null hypothesis which told Dr. Altshuler to use 50 as the benchmark, and not some other number, in the numerator of $z$. That is the exact point where the null hypothesis comes into the procedure. Other null hypotheses will give different benchmarks in the numerator of $z$. The null hypothesis did not tell us the SD of the box. That had to be estimated from the data, in order to compute the SE in the denominator of $z$.

The $z$-statistic of $-3$ stopped Dr. Nullsheimer cold. Why was it so intimidating? After all, 3 is not a very big number. The answer, of course, is that the area to the left of $-3$ under the normal curve is ridiculously small. The chance of getting a sample average 3 SEs or more below its expected value is about 1 in 1,000.

(From the normal table on p. A104, the area is 0.135 of 1%; rounding off, we get 0.1 of 1%; this is 0.1 of 0.01 = 0.001 = 1/1,000.)

The chance of 1 in 1,000 forced Dr. Nullsheimer to concede that the average of the box—not just the average of the sample—was below 50. This chance of 1 in 1,000 is called an *observed significance level*. The observed significance level is often denoted $P$, for probability, and referred to as a *$P$-value*. In the example, the $P$-value of the test was about 1 in 1,000.

Why look at the area to the left of $-3$? The first point to notice: the data could have turned out differently, and then $z$ would have been different too. For instance, if the sample average is 47.2 and the SD is 14.1,

$$z = \frac{47.2 - 50}{0.63} \approx -4.4$$

This is stronger evidence against the null hypothesis: 4.4 SEs below 50 is even worse for "it's just chance" than 3 SEs. On the other hand, if the sample average is 46.9 and the SD is 37,

$$z = \frac{46.9 - 50}{1.65} \approx -1.9$$

This is weaker evidence. The area to the left of $-3$ represents the samples which give even more extreme $z$-values than the observed one, and stronger evidence against the null hypothesis.

> The observed significance level is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null.

The $z$-test can be summarized as follows:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}, \quad P \approx$$

Since the test statistic $z$ depends on the data, so does $P$. That is why $P$ is called an "observed" significance level.

At this point, the logic of the $z$-test can be seen more clearly. It is an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must therefore be rejected. You look at the data, compute the test statistic, and get the observed significance level. Take, for instance, a $P$ of 1 in 1,000. To interpret this number, you start by assuming that the null hypothesis is right. Next, you imagine many other investigators repeating the experiment.

What the 1 in 1,000 says is that your test statistic is really far out. Only one investigator in a thousand would get a test statistic as extreme as, or more extreme than, the one you got. The null hypothesis is creating absurdities, and should be rejected. In general, the smaller the observed significance level, the more you want to reject the null. The phrase "reject the null" emphasizes the point that with a test of significance, the argument is by contradiction.

Our interpretation of $P$ may seem convoluted. It is convoluted. Unfortunately, simpler interpretations turn out to be wrong. If there were any justice in the world, $P$ would be the probability of the null hypothesis given the data. However, $P$ is computed using the null. Even worse, according to the frequency theory, there is no way to define the probability of the null hypothesis being right.

The null is a statement about the box. No matter how often you do the draws, the null hypothesis is either always right or always wrong, because the box does not change.[3] (A similar point for confidence intervals is discussed in section 3 of

chapter 21.) What the observed significance level gives is the chance of getting evidence against the null as strong as the evidence at hand—or stronger—if the null is true.

> The $P$-value of a test is the chance of getting a big test statistic—assuming the null hypothesis to be right. $P$ is not the chance of the null hypothesis being right.

The $z$-test is used for reasonably large samples, when the normal approximation can be used on the probability histogram for the average of the draws. (The average has already been converted to standard units, by $z$.) With small samples, other techniques must be used, as discussed in section 6 below.

## Exercise Set C

1. (a) Other things being equal, which of the following $P$-values is best for the null hypothesis? Explain briefly.

   0.1 of 1% 3% 17% 32%

   (b) Repeat, for the alternative hypothesis.

2. According to one investigator's model, the data are like 50 draws made at random from a large box. The null hypothesis says that the average of the box equals 100. The alternative says that the average of the box is more than 100. The average of the draws is 107.3 and the SD is 22.1. The SE for the sample average is 3.1. Now

   $$z = (107.3 - 100)/3.1 = 2.35 \text{ and } P = 1\%.$$

   True or false, and explain:

   (a) If the null hypothesis is right, there is only a 1% chance of getting a $z$ bigger than 2.35.
   (b) The probability of the null hypothesis given the data is 1%.

3. True or false, and explain:

   (a) The observed significance level depends on the data.
   (b) If the observed significance level is 5%, there are 95 chances in 100 for the alternative hypothesis to be right.

4. According to one investigator's model, the data are like 400 draws made at random from a large box. The null hypothesis says that the average of the box equals 50; the alternative says that the average of the box is more than 50. In fact, the data averaged out to 52.7, and the SD was 25. Compute $z$ and $P$. What do you conclude?

5. In the previous exercise, the null hypothesis says that the average of the _____ is 50. Fill in the blank, and explain briefly. Options: box, sample

6. In the dialog of section 1, suppose the two investigators had only taken a sample of 10 tickets. Should the normal curve be used to compute $P$? Answer yes or no, and explain briefly.

7. Many companies are experimenting with "flex-time," allowing employees to choose their schedules within broad limits set by management.[4] Among other things, flex-time is supposed to reduce absenteeism. One firm knows that in the past few years, employees have averaged 6.3 days off from work (apart from vacations). This year, the firm introduces flex-time. Management chooses a simple random sample of 100 employees to follow in detail, and at the end of the year, these employees average 5.5 days off from work, and the SD is 2.9 days. Did absenteeism really go down, or is this just chance variation? Formulate the null and alternative hypotheses in terms of a box model, then answer the question.

8. Repeat exercise 7 for a sample average of 5.9 days and an SD of 2.9 days.

*The answers to these exercises are on pp. A91–92.*

## 4. MAKING A TEST OF SIGNIFICANCE

Making a test of significance is a complicated job. You have to

- set up the null hypothesis, in terms of a box model for the data;
- pick a test statistic, to measure the difference between the data and what is expected on the null hypothesis;
- compute the observed significance level $P$.

The choice of test statistic depends on the model and the hypothesis being considered. So far, we've discussed the "one-sample $z$-test." Two-sample $z$-tests will be covered in chapter 27. There are also "$t$-tests" based on the $t$-statistic (section 6), "$\chi^2$-tests" based on the $\chi^2$-statistic (chapter 28), and many other tests not even mentioned in this book. However, all tests follow the steps outlined above, and their $P$-values can be interpreted in the same way.

It is natural to ask how small the observed significance level has to be before you reject the null hypothesis. Many investigators draw the line at 5%.

- If $P$ is less than 5%, the result is called *statistically significant* (often shortened to *significant*).

There is another line at 1%.

- If $P$ is less than 1%, the result is called *highly significant*.

These somewhat arbitrary lines will be discussed again in section 1 of chapter 29.

Do not let the jargon distract you from the main idea. When the data are too far from the predictions of a theory, that is bad for the theory. In statistics, the null hypothesis is rejected when the observed value is too many SEs away from the expected value.

## Exercise Set D

1. True or false:
   (a) A "highly significant" result cannot possibly be due to chance.
   (b) If a difference is "highly significant," there is less than a 1% chance for the null hypothesis to be right.

(c) If a difference is "highly significant," there is better than a 99% chance for the alternative hypothesis to be right.

2. True or false:

   (a) If $P$ is 43%, the null hypothesis looks plausible.
   (b) If $P$ is 0.43 of 1%, the null hypothesis looks implausible.

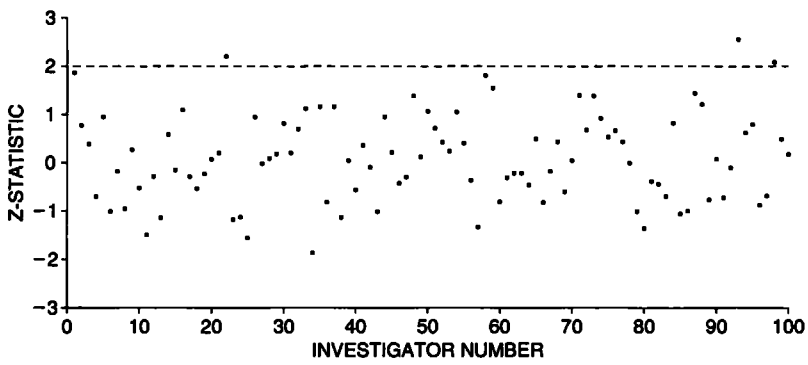
3. True or false:

   (a) If the observed significance level is 4%, the result is "statistically significant."
   (b) If the $P$-value of a test is 1.1%, the result is "highly significant."
   (c) If a difference is "highly significant," then $P$ is less than 1%.
   (d) If the observed significance level is 3.6%, then $P = 3.6\%$.
   (e) If $z = 2.3$, then the observed value is 2.3 SEs above what is expected on the null hypothesis.

4. An investigator draws 250 tickets at random with replacement from a box. What is the chance that the average of the draws will be more than 2 SEs above the average of the box?

5. One hundred investigators set out to test the null hypothesis that the average of the numbers in a certain box equals 50. Each investigator takes 250 tickets at random with replacement, computes the average of the draws, and does a $z$-test. The results are plotted in the diagram. Investigator #1 got a $z$-statistic of 1.9, which is plotted as the point (1, 1.9). Investigator #2 got a $z$-statistic of 0.8, which is plotted as (2, 0.8), and so forth. Unknown to the investigators, the null hypothesis is true.

   (a) True or false, and explain: the $z$-statistic is positive when the average of the draws is more than 50.
   (b) How many investigators should get a positive $z$-statistic?
   (c) How many of them should get a $z$-statistic bigger than 2? How many of them actually do?
   (d) If $z = 2$, what is $P$?




*The answers to these exercises are on p. A92.*

## 5. ZERO-ONE BOXES

The $z$-test can also be used when the situation involves classifying and counting. It is a matter of putting 0's and 1's in the box (section 5 of chapter 17). This

section will give an example. Charles Tart ran an experiment at the University of California, Davis, to demonstrate ESP.[5] Tart used a machine called the "Aquarius." The Aquarius has an electronic random number generator and 4 "targets." Using its random number generator, the machine picks one of the 4 targets at random. It does not indicate which. Then, the subject guesses which target was chosen, by pushing a button. Finally, the machine lights up the target it picked, ringing a bell if the subject guessed right. The machine keeps track of the number of trials and the number of correct guesses.

Tart selected 15 subjects who were thought to be clairvoyant. Each of the subjects made 500 guesses on the Aquarius, for a total of $15 \times 500 = 7{,}500$ guesses. Out of this total, 2,006 were right. Of course, even if the subjects had no clairvoyant abilities whatsoever, they would still be right about 1/4 of the time. In other words, about $1/4 \times 7{,}500 = 1{,}875$ correct guesses are expected, just by chance. True, there is a surplus of $2{,}006 - 1{,}875 = 131$ correct guesses, but can't this be explained as a chance variation?

Tart could—and did—fend off the "it's only chance" explanation by making a test of significance. To set up a box model, he assumed that the Aquarius generates numbers at random, so each of the 4 targets has 1 chance in 4 to be chosen. He assumed (temporarily) that there is no ESP. Now, a guess has 1 chance in 4 to be right.

The data consist of a record of the 7,500 guesses, showing whether each one is right or wrong. The null hypothesis says that the data are like 7,500 draws from the box

$$\boxed{\boxed{1}\,\boxed{0}\,\boxed{0}\,\boxed{0}} \qquad 1 = \text{right}, \quad 0 = \text{wrong}$$

The number of correct guesses is like the sum of 7,500 draws from the box. This completes the box model for the null hypothesis.

The machine is classifying each guess as right or wrong, and counting the number of correct guesses. That is why a zero-one box is needed. Once the null hypothesis has been translated into a box model, the $z$-test can be used:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

The "observed" is 2,006, the number of correct guesses. The expected number of correct guesses comes from the null hypothesis, and is 1,875. The numerator of the $z$-statistic is $2{,}006 - 1{,}875 = 131$, the surplus number of correct guesses.

Now for the denominator. You need the SE for the number of correct guesses. Look at the box model. In this example, the null hypothesis tells you exactly what is in the box: a 1 and three 0's. The SD of the box is $\sqrt{0.25 \times 0.75} \approx 0.43$. The SE is $\sqrt{7{,}500} \times 0.43 \approx 37$. So

$$z = 131/37 \approx 3.5$$

The observed value of 2,006 is 3.5 SEs above the expected value. And $P$ is tiny:

P ≈

≈ 2 in 10,000

The surplus of correct guesses is hard to dismiss as a chance variation. This looks like strong evidence for ESP. However, there are other possibilities to consider. For example, the Aquarius random number generator may not be very good (section 5 of chapter 29). Or the machine may be giving the subject some subtle clues as to which target it picked. There may be many reasonable explanations for the results, besides ESP. But chance variation isn't one of them. That is what the test of significance shows, finishing the ESP example.

The same $z$-statistic is used here as in section 1:

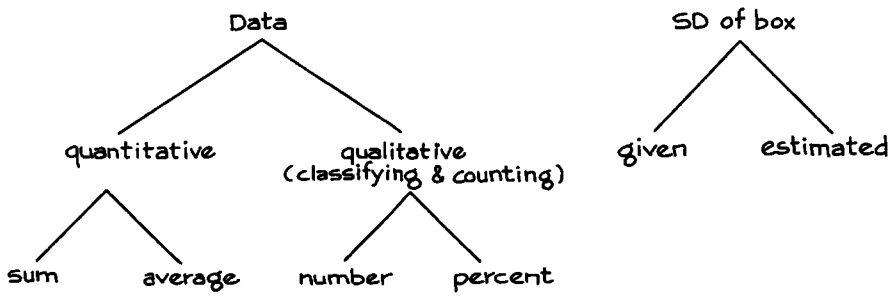
$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

Although the formula is the same, there are some differences between the $z$-test in this section and the $z$-test in section 1.

1) In section 1, the SE was for an average. Here, the SE is for the number of correct guesses. To work out $z$, first decide what is "observed" in the numerator. Are you dealing with a sum, an average, a number, or a percent? That will tell you which SE to use in the denominator. In the ESP example, the number of correct guesses was observed. That is why the SE for the number goes into the denominator, as indicated by the sketch.

$$z = \frac{\overset{\text{number}}{\text{observed}} - \overset{\text{number}}{\text{expected}}}{\underset{\text{for number}}{\text{SE}}}$$

2) In section 1, the SD of the box was unknown. The investigators had to estimate it from the data. Here, the SD of the box is given by the null hypothesis. You do not have to estimate it. The diagram summarizes points 1) and 2).




3) In section 1, there was an alternative hypothesis about the box: its average was below 50. With ESP, there is no sensible way to set up the alternative hypothesis as a box model. The reason: if the subjects do have ESP, the chance for each guess to be right may well depend on the previous trials, and may change from trial to trial. Then the data will not be like draws from a box.[6]

4) In section 1, the data were like draws from a box, because the investigators agreed to take a simple random sample of tickets. The argument was only about the average of the box. Here, part of the question is *whether* the data are like draws from a box—any box.

Chapters 19–24 were about estimating parameters from data, and getting margins of error. *Testing*, the topic of this chapter, is about another kind of question. For example, is a parameter equal to some prespecified value, or isn't it? Estimation and testing are related, but the goals are different.

## Exercise Set E

*This exercise set also covers material from previous sections.*

1. In Tart's experiment, the null hypothesis says that _____. Fill in the blank, using one of the options below.
   (i) The data are like 7,500 draws from the box [0 0 0 1].
   (ii) The data are like 7,500 draws from the box [0 0 1].
   (iii) The fraction of 1's in the box is 2,006/7,500.
   (iv) The fraction of 1's among the draws is 2,006/7,500.
   (v) ESP is real.

2. As part of a statistics project in the early 1970s, Mr. Frank Alpert approached the first 100 students he saw one day on Sproul Plaza at the University of California, Berkeley, and found out the school or college in which they enrolled. There were 53 men in his sample. From Registrar's data, 25,000 students were registered at Berkeley that term, and 67% were male. Was his sampling procedure like taking a simple random sample?

   Fill in the blanks. That will lead you step by step to the box model for the null hypothesis. (There is no alternative hypothesis about the box.)
   (a) There is one ticket in the box for each _____.
   person in the sample  student registered at Berkeley that term
   (b) The ticket is marked _____ for the men and _____ for the women.
   (c) The number of tickets in the box is _____ and the number of draws is _____. Options: 100, 25,000.
   (d) The null hypothesis says that the sample is like _____ _____ made at random from the box. (The first blank must be filled in with a number; the second, with a word.)
   (e) The percentage of 1's in the box is _____. Options: 53%, 67%.

3. (This continues exercise 2.) Fill in the blanks. That will lead you step by step to $z$ and $P$.
   (a) The observed number of men is _____.
   (b) The expected number of men is _____.
   (c) If the null hypothesis is right, the number of men in the sample is like the _____ of the draws from the box. Options: sum, average.
   (d) The SE for the number of men is _____.
   (e) $z =$ _____ and $P =$ _____.

4. (This continues exercises 2 and 3.) Was Alpert's sampling procedure like taking a simple random sample? Answer yes or no, and explain briefly.

5. This also continues exercises 2 and 3.

   (a) In 3(b), the expected number was _____.

   computed from the null hypothesis  estimated from the data

   (b) In 3(d), the SE was _____.

   computed from the null hypothesis  estimated from the data

6. Another ESP experiment used the "Ten Choice Trainer." This is like the Aquarius, but with 10 targets instead of 4. Suppose that in 1,000 trials, a subject scores 173 correct guesses.

   (a) Set up the null hypothesis as a box model.
   (b) The SD of the box is _____. Fill in the blank, using one of the options below, and explain briefly.

   $$\sqrt{0.1 \times 0.9} \qquad \sqrt{0.173 \times 0.827}$$

   (c) Make the $z$-test.
   (d) What do you conclude?

7. A coin is tossed 10,000 times, and it lands heads 5,167 times. Is the chance of heads equal to 50%? Or are there too many heads for that?

   (a) Formulate the null and alternative hypotheses in terms of a box model.
   (b) Compute $z$ and $P$.
   (c) What do you conclude?

8. Repeat exercise 7 if the coin lands heads 5,067 times, as it did for Kerrich (section 1 of chapter 16).

9. One hundred draws are made at random with replacement from a box of tickets; each ticket has a number written on it. The average of the draws is 29 and the SD of the draws is 40. You see a statistician make the following calculation:

   $$z = \frac{29 - 20}{4} = 2.25, \quad P \approx 1\%$$

   (a) She seems to be testing the null hypothesis that the average of the _____ is 20. Options: box, sample.
   (b) True or false: there is about a 1% chance for the null hypothesis to be right.

   Explain briefly.

10. A colony of laboratory mice consisted of several hundred animals. Their average weight was about 30 grams, and the SD was about 5 grams. As part of an experiment, graduate students were instructed to choose 25 animals haphazardly, without any definite method.[7] The average weight of these animals turned out to be around 33 grams, and the SD was about 7 grams. Is choosing animals haphazardly the same as drawing them at random? Or is 33 grams too far above average for that? Discuss briefly; formulate the null hypothesis as a box model; compute $z$ and $P$. (There is no need to formulate an alternative hypothesis about the box; you must decide whether the null hypothesis tells you the SD of the box: if not, you have to estimate the SD from the data.)

11. (Hard.) Discount stores often introduce new merchandise at a special low price in order to induce people to try it. However, a psychologist predicted that this practice would actually reduce sales. With the cooperation of a discount chain, an experiment was performed to test the prediction.[8] Twenty-five pairs of stores were selected, matched according to such characteristics as location and sales volume. These stores did not advertise, and displayed their merchandise in similar ways.

    A new kind of cookie was introduced in all 50 stores. For each pair of stores, one was chosen at random to introduce the cookies at a special low price, the price increasing to its regular level after two weeks. The other store in the pair introduced the cookies at the regular price. Total sales of the cookies were computed for each store for six weeks from the time they were introduced.

    In 18 of the 25 pairs, the store which introduced the cookies at the regular price turned out to have sold more of them than the other store. Can this result be explained as a chance variation? Or does it support the prediction that introducing merchandise at a low price reduces long-run sales? (Formulate the null hypothesis as a box model; there is no alternative hypothesis about the box.)

*The answers to these exercises are on pp. A92–93.*

## 6. THE $t$-TEST

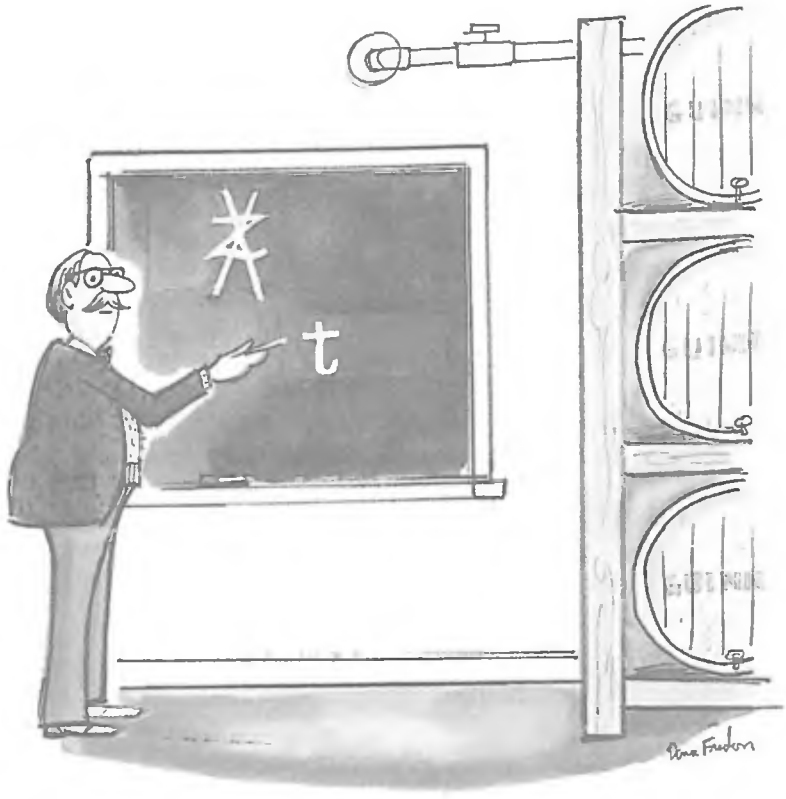
With small samples, the $z$-test has to be modified. Statisticians use the $t$-test, which was invented by W. S. Gosset (England, 1876–1936). Gosset worked as an executive in the Guinness Brewery, where he went after taking his degree at Oxford. He published under the pen name "Student" because his employers didn't want the competition to realize how useful the results could be.[9]

This section will show how to do the $t$-test, by example. However, the discussion is a bit technical, and can be skipped. In Los Angeles, many studies have been conducted to determine the concentration of CO (carbon monoxide) near freeways with various conditions of traffic flow. The basic technique involves capturing air samples in special bags, and then determining the CO concentrations in the bag samples by using a machine called a *spectrophotometer*. These machines can measure concentrations up to about 100 ppm (parts per million by volume) with errors on the order of 10 ppm. Spectrophotometers are quite delicate and have to be calibrated every day. This involves measuring CO concentration in a manufactured gas sample, called *span gas*, where the concentration is precisely controlled at 70 ppm. If the machine reads close to 70 ppm on the span gas, it's ready for use; if not, it has to be adjusted. A complicating factor is that the size of the measurement errors varies from day to day. On any particular day, however, we assume that the errors are independent and follow the normal curve; the SD is unknown and changes from day to day.[10]

One day, a technician makes five readings on span gas, and gets

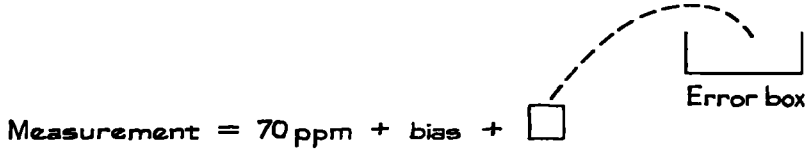
78 83 68 72 88

Four out of five of these numbers are higher than 70, and some of them by quite a




bit. Can this be explained on the basis of chance variation? Or does it show bias, perhaps from improper adjustment of the machine?

A test of significance is called for, and a box model is needed. The one to use is the Gauss model (section 3 of chapter 24). According to this model, each measurement equals the true value of 70 ppm, plus bias, plus a draw with replacement from the error box. The tickets in the error box average out to 0, and the SD is unknown.




The key parameter is the bias. The null hypothesis says that the bias equals 0. On this hypothesis, the average of the 5 measurements has an expected value of 70 ppm; the difference between the average and 70 ppm is explained as a chance variation. The alternative hypothesis says that the bias differs from 0, so the difference between the average of the measurements and 70 ppm is real.

As before, the appropriate test statistic to use is

$$\frac{\text{observed} - \text{expected}}{\text{SE}}$$

The average of the 5 measurements is 77.8 ppm, and their SD is 7.22 ppm. It seems right to estimate the SD of the error box by 7.22 ppm. The SE for the sum of the draws is $\sqrt{5} \times 7.22 \approx 16.14$ ppm. And the SE for the average is $16.14/5 \approx 3.23$ ppm. The test statistic is

$$\frac{77.8 - 70}{3.23} \approx 2.4$$

In other words, the average of the sample is about 2.4 SEs above the value expected on the null hypothesis. Now the area to the right of 2.4 under the normal curve is less than 1%. This *P*-value looks like strong evidence against the null hypothesis.

There is something missing, however. The SD of the measurements is only an estimate for the SD of the error box. And the number of measurements is so small that the estimate could be way off. There is extra uncertainty that has to be taken into account, which is done in two steps.

*Step 1.* When the number of measurements is small, the SD of the error box should not be estimated by the SD of the measurements. Instead, $\text{SD}^+$ is used[11]

$$\text{SD}^+ = \sqrt{\frac{\text{number of measurements}}{\text{number of measurements} - \text{one}}} \times \text{SD}.$$

This estimate is larger. (See p. 74 for the definition of $\text{SD}^+$, and p. 495 for the logic behind the definition.)

In the example, the number of measurements is 5 and their SD is 7.22 ppm. So $\text{SD}^+ \approx \sqrt{5/4} \times 7.22 \approx 8.07$ ppm. Then, the SE is figured in the usual way. The SE for the sum is $\sqrt{5} \times 8.07 \approx 18.05$ ppm; the SE for the average is $18.05/5 = 3.61$ ppm. The test statistic becomes

$$\frac{77.8 - 70}{3.61} \approx 2.2$$

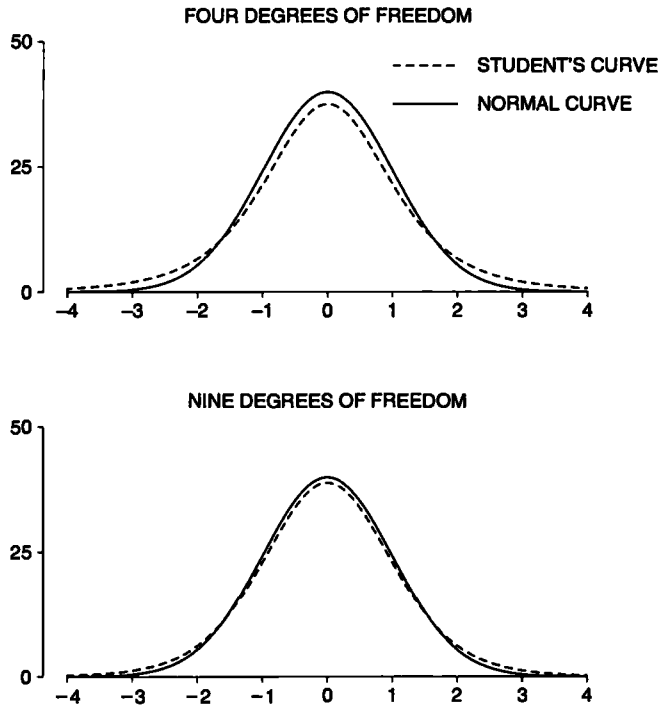
*Step 2.* The next step is to find the *P*-value. With a large number of measurements, this can be done using the normal curve. But with a small number of measurements, a different curve must be used, called *Student's curve*. As it turns out, the *P*-value from Student's curve is about 5%. That is quite a bit more than the 1% from the normal curve.

Using Student's curve takes some work. Actually, there is one of these curves for each number of *degrees of freedom*. In the present context,

$$\text{degrees of freedom} = \text{number of measurements} - \text{one}.$$

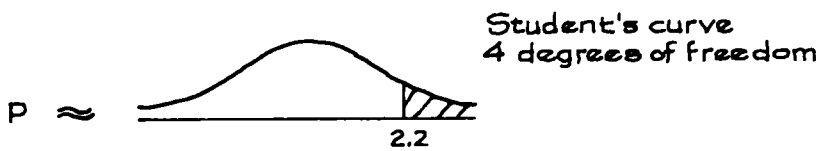
Student's curves for 4 and 9 degrees of freedom are shown in figure 1, with the

Figure 1. Student's curves. The dashed line is Student's curve for 4 degrees of freedom (top panel) or 9 degrees of freedom (bottom). The solid line is a normal curve, for comparison.




normal curve for comparison. Student's curves look quite a lot like the normal curve, but they are less piled up in the middle and more spread out. As the number of degrees of freedom goes up, the curves get closer and closer to the normal, reflecting the fact that the SD of the measurements is getting closer and closer to the SD of the error box. The curves are all symmetric around 0, and the total area under each one equals 100%.[12]

In the example, with 5 measurements there are $5-1=4$ degrees of freedom. To find the $P$-value, we need to find the area to the right of 2.2 under Student's curve with 4 degrees of freedom:




The area can be found with the help of a special table (p. A105), part of which is shown in table 1 (next page). The rows are labeled by degrees of freedom. Look

across the row for 4 degrees of freedom. The first entry is 1.53, in the column headed 10%. This means the area to the right of 1.53 under Student's curve with 4 degrees of freedom equals 10%. The other entries can be read the same way.

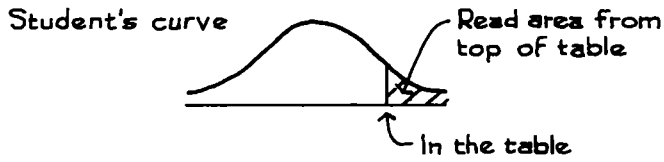
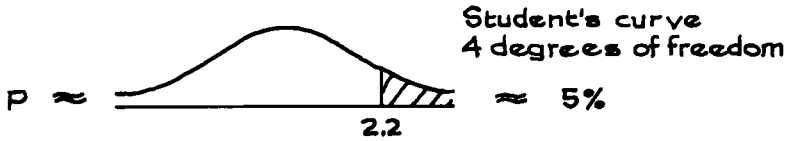



Table 1. A short $t$-table.

| *Degrees of freedom* | *10%* | *5%* | *1%* |
|---|---|---|---|
| 1 | 3.08 | 6.31 | 31.82 |
| 2 | 1.89 | 2.92 | 6.96 |
| 3 | 1.64 | 2.35 | 4.54 |
| 4 | 1.53 | 2.13 | 3.75 |
| 5 | 1.48 | 2.02 | 3.36 |

In the example, there are 4 degrees of freedom, and $t$ is 2.2. From table 1, the area under Student's curve to right of 2.13 is 5%. So the area to the right of 2.2 must be about 5%. The $P$-value is about 5%.




The evidence is running against the null hypothesis, though not very strongly. This completes the example.

Student's curve should be used under the following circumstances.

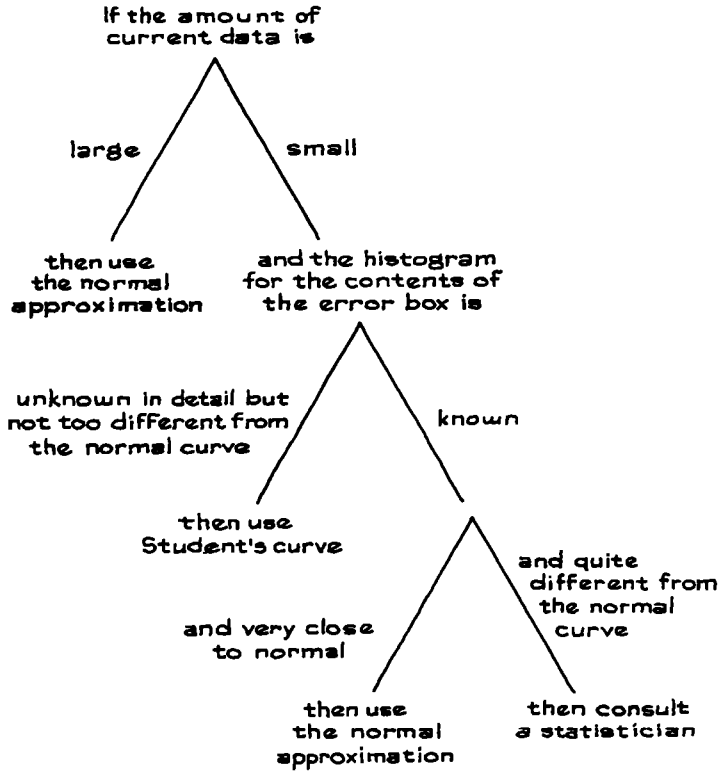
- The data are like draws from a box.
- The SD of the box is unknown.
- The number of observations is small, so the SD of the box cannot be estimated very accurately.
- The histogram for the contents of the box does not look too different from the normal curve.

With a large number of measurements (say 25 or more), the normal curve would ordinarily be used. If the SD of the box is known, and the contents of the box follow the normal curve, then the normal curve can be used even for small samples.[13]

*Example 1.* On another day, 6 readings on span gas turn out to be

72 79 65 84 67 77.

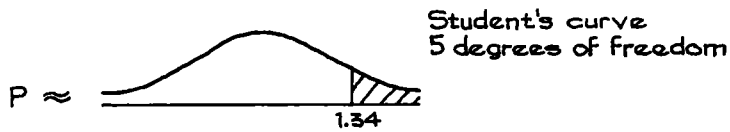
Is the machine properly calibrated? Or do the measurements show bias?



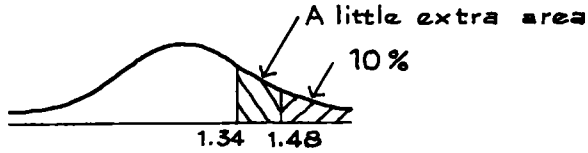

*Solution.* The model is the same as before. The average of the new measurements is 74 ppm, and their SD is 6.68 ppm. Since there are only 6 observations, the SD of the error box should be estimated by $SD^+$ of the data, not the SD. The $SD^+$ is $\sqrt{6/5} \times 6.68 \approx 7.32$ ppm, so the SE for the average is 2.99 ppm. Now

$$t = \frac{74 - 70}{2.99} \approx 1.34$$

To compute the $P$-value, Student's curve is used instead of the normal, with $6 - 1 = 5$ degrees of freedom.




From table 1, the area to the right of 1.34 under Student's curve with 5 degrees of freedom is a little more than 10%. There does not seem to be much evidence of bias. The machine is ready to use. The reasoning on the 10%: from the table, the




area to the right of 1.48 is 10%. And 1.34 is just to the left of 1.48. So the area to the right of 1.34 is a little more than 10%.

## Exercise Set F

1. Find the area under Student's curve with 5 degrees of freedom:
   (a) to the right of 2.02
   (b) to the left of −2.02
   (c) between −2.02 and 2.02
   (d) to the left of 2.02.

2. The area to the right of 4.02 under Student's curve with 2 degrees of freedom is

   less than 1% between 1% and 5% more than 5%

   Choose one option, and explain.

3. True or false, and explain: to make a $t$-test with 4 measurements, use Student's curve with 4 degrees of freedom.

4. Each (hypothetical) data set below represents some readings on span gas. Assume the Gauss model, with errors following the normal curve. However, bias may be present. In each case, make a $t$-test to see whether the instrument is properly calibrated or not. In one case, this is impossible. Which one, and why?
   (a) 71, 68, 79
   (b) 71, 68, 79, 84, 78, 85, 69
   (c) 71
   (d) 71, 84

5. A new spectrophotometer is being calibrated. It is not clear whether the errors follow the normal curve, or even whether the Gauss model applies. In two cases, these assumptions should be rejected. Which two, and why? The numbers are replicate measurements on span gas.
   (a) 71, 70, 72, 69, 71, 68, 93, 75, 68, 61, 74, 67
   (b) 71, 73, 69, 74, 65, 67, 71, 69, 70, 75, 71, 68
   (c) 71, 69, 71, 69, 71, 69, 71, 69, 71, 69, 71, 69

6. A long series of measurements on a checkweight averages out to 253 micrograms above ten grams, and the SD is 7 micrograms. The Gauss model is believed to apply, with negligible bias. At this point, the balance has to be rebuilt, which may introduce bias as well as changing the SD of the error box. Ten measurements on the checkweight, using the rebuilt scale, show an average of 245 micrograms above ten grams, and the SD is 9 micrograms. Has bias been introduced? Or is this chance variation? (You may assume that the errors follow the normal curve.)

7. Several thousand measurements on a checkweight average out to 512 micrograms above a kilogram; the SD is 50 micrograms. Then, the weight is cleaned. The next 100 measurements average out to 508 micrograms above one kilogram; the SD is 52 micrograms. Apparently, the weight got 4 micrograms lighter. Or is this chance variation? (You may assume the Gauss model with no bias.)
   (a) Formulate the null and alternative hypotheses as statements about a box model.
   (b) Would you estimate the SD of the box as 50 or 52 micrograms?
   (c) Would you make a $z$-test or a $t$-test?
   (d) Did the weight get lighter? If so, by how much?

*The answers to these exercises are on p. A94.*

*Technical notes.* (i) The term "degrees of freedom" is slightly baroque; here is the idea behind the phrase. The SE for the average depends on the SD of the measurements, and that in turn depends on the deviations from the average. But the sum of the deviations has to be 0, so they cannot all vary freely. The constraint that the sum equals 0 eliminates one degree of freedom. For example, with 5 measurements, the sum of the 5 deviations is 0. If you know 4 of them, you can compute the 5th—so there are only 4 degrees of freedom.

(ii) Why use $SD^+$? Suppose we have some draws made at random with replacement from a box whose SD is unknown. If we knew the average of the box, the r.m.s. difference between the sample numbers and the average of the box could be used to estimate the SD of the box. However, we usually do not know the average of the box and must estimate that too, using the average of the draws. Now there is a little problem. The average of the draws follows the draws around; deviations from the average of the *draws* tend to be smaller than deviations from the average of the *box*. $SD^+$ corrects this problem.

## 7. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. True or false, and explain:
   (a) The $P$-value of a test equals its observed significance level.
   (b) The alternative hypothesis is another way of explaining the results; it says the difference is due to chance.

2. With a perfectly balanced roulette wheel, in the long run, red numbers should turn up 18 times in 38. To test its wheel, one casino records the results of 3,800 plays, finding 1,890 red numbers. Is that too many reds? Or chance variation?
   (a) Formulate the null and alternative hypotheses as statements about a box model.
   (b) The null says that the percentage of reds in the box is _____. The alternative says that the percentage of reds in the box is _____. Fill in the blanks.

(c) Compute $z$ and $P$.
(d) Were there too many reds?

3. One kind of plant has only blue flowers and white flowers. According to a genetic model, the offsprings of a certain cross have a 75% chance to be blue-flowering, and a 25% chance to be white-flowering, independently of one another. Two hundred seeds of such a cross are raised, and 142 turn out to be blue-flowering. Are the data consistent with the model? Answer yes or no, and explain briefly.

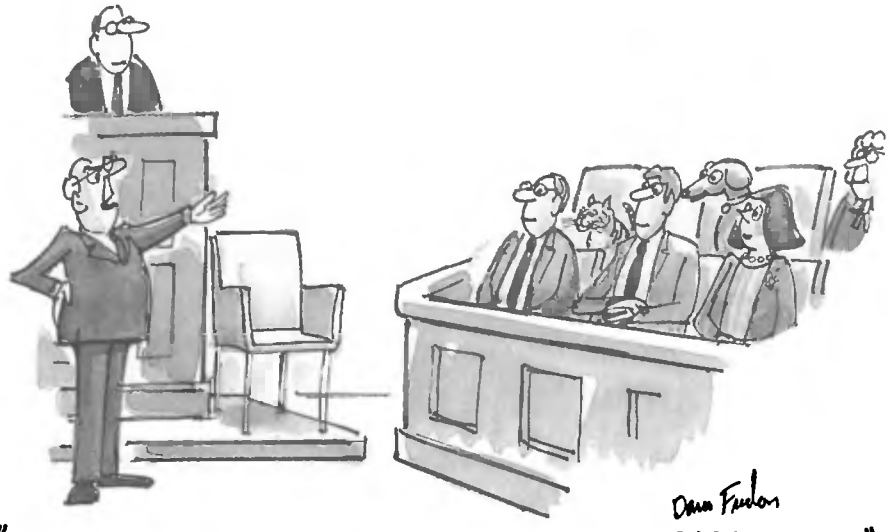
4. One large course has 900 students, broken down into section meetings with 30 students each. The section meetings are led by teaching assistants. On the final, the class average is 63, and the SD is 20. However, in one section the average is only 55. The TA argues this way:

> If you took 30 students at random from the class, there is a pretty good chance they would average below 55 on the final. That's what happened to me—chance variation.

Is this a good defense? Answer yes or no, and explain briefly.

5. A newspaper article says that on the average, college freshmen spend 7.5 hours a week going to parties.[14] One administrator does not believe that these figures apply at her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen, and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours. Is the difference between 6.6 and 7.5 real?
   (a) Formulate the null and alternative hypotheses in terms of a box model.
   (b) Fill in the blanks. The null says that the average of the box is _____. The alternative says that average of the box is _____.
   (c) Now answer the question: is the difference real?

6. In 1969, Dr. Spock came to trial before Judge Ford, in Boston's federal court house. The charge was conspiracy to violate the Military Service Act. "Of all defendants, Dr. Spock, who had given wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury."[15] The jury was drawn from a "venire," or panel, of 350 persons selected by the clerk. This venire included only 102 women, although a majority of the eligible jurors in the district were female. At the next stage in selecting the jury to hear the case, Judge Ford chose 100 potential jurors out of these 350 persons. His choices included 9 women.
   (a) 350 people are chosen at random from a large population, which is over 50% female. How likely is it that the sample includes 102 women or fewer?
   (b) 100 people are chosen at random (without replacement) from a group consisting of 102 women and 248 men. How likely is it that the sample includes 9 women or fewer?
   (c) What do you conclude?



"YOUR HONOR, THE PROSECUTION OBJECTS TO THE COMPOSITION OF THIS JURY."

7. I. S. Wright and associates did a clinical trial on the effect of anticoagulant therapy for coronary heart disease.[16] Eligible patients who were admitted to participating hospitals on odd days of the month were given the therapy; eligible patients admitted on even days were the controls. In total, there were 580 patients in the therapy group and 442 controls. An observer says,

   > Since the odd-even assignment to treatment or control is objective and impartial, it is just as good as tossing a coin.

   Do you agree or disagree? Explain briefly. Assume the trial was done in a month with 30 days.

8. Bookstores like education, one reason being that educated people are more likely to spend money on books. National data show the nationwide average educational level to be 13 years of schooling completed, with an SD of about 3 years, for persons age 18 and over.[17]

   A bookstore is doing a market survey in a certain county, and takes a simple random sample of 1,000 people age 18 and over. They find the average educational level to be 14 years, and the SD is 5 years. Can the difference in average educational level between the sample and the nation be explained by chance variation? If not, what other explanation can you give?

9. A computer is programmed to make 100 draws at random with replacement from the box $\boxed{0\ 0\ 0\ 0\ 1}$, and take their sum. It does this 144 times; the average of the 144 sums is 21.13. The program is working fine. Or is it?

   Working fine  Something is wrong

   Choose one option, and explain your reason.

10. On November 9, 1965, the power went out in New York City, and stayed out for a day—the Great Blackout. Nine months later, the newspapers suggested that New York was experiencing a baby boom. The table below shows the number of babies born every day during a 25 day period, centered nine months and ten days after the Great Blackout.[18] These numbers average out to 436. This turns out not to be unusually high for New York. But there is an interesting twist to the data: the 3 Sundays only average 357. How likely is it that the average of 3 days chosen at random from the table will be 357 or less? Is chance a good explanation for the difference between Sundays and weekdays? If not, how would you explain the difference?

*Number of births in New York, August 1–25, 1966*

| *Date* | *Day* | *Number* | *Date* | *Day* | *Number* |
|---|---|---|---|---|---|
| 1 | Mon. | 451 | 15 | Mon. | 451 |
| 2 | Tues. | 468 | 16 | Tues. | 497 |
| 3 | Wed. | 429 | 17 | Wed. | 458 |
| 4 | Thur. | 448 | 18 | Thur. | 429 |
| 5 | Fri. | 466 | 19 | Fri. | 434 |
| 6 | Sat. | 377 | 20 | Sat. | 410 |
| 7 | Sun. | 344 | 21 | Sun. | 351 |
| 8 | Mon. | 448 | 22 | Mon. | 467 |
| 9 | Tue. | 438 | 23 | Tues. | 508 |
| 10 | Wed. | 455 | 24 | Wed. | 432 |
| 11 | Thur. | 468 | 25 | Thur. | 426 |
| 12 | Fri. | 462 | | | |
| 13 | Sat. | 405 | | | |
| 14 | Sun. | 377 | | | |

11. According to the census, the median household income in Atlanta (1.5 million households) was $52,000 in 1999.[19] In June 2003, a market research organization takes a simple random sample of 750 households in Atlanta; 56% of the sample households had incomes over $52,000. Did median household income in Atlanta increase over the period 1999 to 2003?

    (a) Formulate null and alternative hypotheses in terms of a box model.
    (b) Calculate the appropriate test statistic and $P$.
    (c) Did median family income go up?

12. (Hard.) Does the psychological environment affect the anatomy of the brain? This question was studied experimentally by Mark Rosenzweig and his associates.[20] The subjects for the study came from a genetically pure strain of rats. From each litter, one rat was selected at random for the treatment group, and one for the control group. Both groups got exactly the same kind of food and drink—as much as they wanted. But each animal in the treatment group lived with 11 others in a large cage, furnished with playthings which were changed daily. Animals in the control group lived in isolation, with no toys. After a month, the experimental animals were killed and dissected.

*Cortex weights (in milligrams) for experimental animals. The treatment group (T) had an enriched environment. The control group (C) had a deprived environment.*

| *Expt. #1* | | *Expt. #2* | | *Expt. #3* | | *Expt. #4* | | *Expt. #5* | |
|---|---|---|---|---|---|---|---|---|---|
| *T* | *C* | *T* | *C* | *T* | *C* | *T* | *C* | *T* | *C* |
| 689 | 657 | 707 | 669 | 690 | 668 | 700 | 662 | 640 | 641 |
| 656 | 623 | 740 | 650 | 701 | 667 | 718 | 705 | 655 | 589 |
| 668 | 652 | 745 | 651 | 685 | 647 | 679 | 656 | 624 | 603 |
| 660 | 654 | 652 | 627 | 751 | 693 | 742 | 652 | 682 | 642 |
| 679 | 658 | 649 | 656 | 647 | 635 | 728 | 578 | 687 | 612 |
| 663 | 646 | 676 | 642 | 647 | 644 | 677 | 678 | 653 | 603 |
| 664 | 600 | 699 | 698 | 720 | 665 | 696 | 670 | 653 | 593 |
| 647 | 640 | 696 | 648 | 718 | 689 | 711 | 647 | 660 | 672 |
| 694 | 605 | 712 | 676 | 718 | 642 | 670 | 632 | 668 | 612 |
| 633 | 635 | 708 | 657 | 696 | 673 | 651 | 661 | 679 | 678 |
| 653 | 642 | 749 | 692 | 658 | 675 | 711 | 670 | 638 | 593 |
| | | 690 | 621 | 680 | 641 | 710 | 694 | 649 | 602 |

On the average, the control animals were heavier and had heavier brains, perhaps because they ate more and got less exercise. However, the treatment group had consistently heavier cortexes (the "grey matter," or thinking part of the brain). This experiment was repeated many times; results from the first 5 trials are shown in the table: "T" means treatment, and "C" is for control. Each line refers to one pair of animals. In the first pair, the animal in treatment had a cortex weighing 689 milligrams; the one in control had a lighter cortex, weighing only 657 milligrams. And so on.

Two methods of analyzing the data will be presented in the form of exercises. Both methods take into account the pairing, which is a crucial feature of the data. (The pairing comes from randomization within litter.)

(a) *First analysis.* How many pairs were there in all? In how many of these pairs did the treatment animal have a heavier cortex? Suppose treatment had no effect, so each animal of the pair had a 50–50 chance to have the heavier cortex, independently from pair to pair. Under this assumption, how likely is it that an investigator would get as many pairs as Rosenzweig did, or more, with the treatment animal having the heavier cortex? What do you infer?

(b) *Second analysis.* For each pair of animals, compute the difference in cortex weights "treatment − control." Find the average and SD of all these differences. The null hypothesis says that these differences are like draws made at random with replacement from a box whose average is 0—the treatment has no effect. Make a $z$-test of this hypothesis. What do you infer?

(c) To ensure the validity of the analysis, the following precaution was taken. "The brain dissection and analysis of each set of littermates was done in immediate succession but in a random order and identified only

by code number so that the person doing the dissection does not know which cage the rat comes from." Comment briefly on the following: What was the point of this precaution? Was it a good idea?

## 8. SUMMARY

1. A *test of significance* gets at the question of whether an observed difference is real (the *alternative hypothesis*) or just a chance variation (the *null hypothesis*).

2. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box.

3. A *test statistic* measures the difference between the data and what is expected on the null hypothesis. The *z-test* uses the statistic

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

The expected value in the numerator is computed on the basis of the null hypothesis. If the null hypothesis determines the SD of the box, use this information when computing the SE in the denominator. Otherwise, you have to estimate the SD from the data.

4. The *observed significance level* (also called $P$, or the $P$-value) is the chance of getting a test statistic as extreme as or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. Therefore, $P$ does not give the chance of the null hypothesis being right.

5. Small values of $P$ are evidence against the null hypothesis: they indicate something besides chance was operating to make the difference.

6. Suppose that a small number of tickets are drawn at random with replacement from a box whose contents follow the normal curve, with an average of 0 and an unknown SD. Each draw is added to an unknown constant to give a measurement. The null hypothesis says that this unknown constant equals some given value $c$. An alternative hypothesis says that the unknown constant is bigger than $c$. The SD of the box is estimated by the $\text{SD}^+$ of the data. Then the SE for the average of the draws is computed. The test statistic is

$$t = \frac{\text{average of draws} - c}{\text{SE}}$$

The observed significance level is obtained not from the normal curve but from one of the Student's curves, with

$$\text{degrees of freedom} = \text{number of measurements} - \text{one}.$$

This procedure is a *t-test*.

# 27

# More Tests for Averages

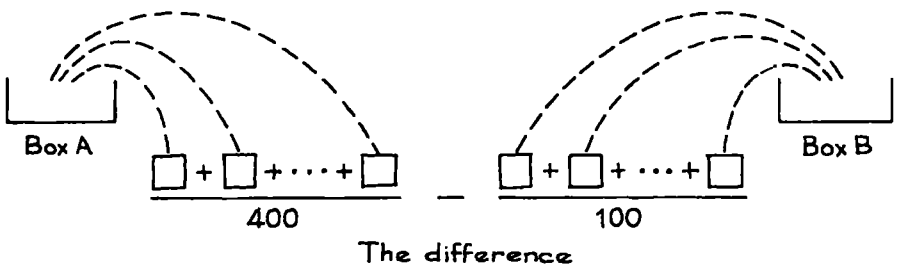
*Vive la différence!*

## 1. THE STANDARD ERROR FOR A DIFFERENCE

This chapter is about comparing two samples. The SE for the difference between their averages is needed. We begin with an example to illustrate the mathematics. (Real examples come later.) Suppose two boxes A and B have the averages and SDs shown below.

| Box A | Box B |
|---|---|
| Average = 110 | Average = 90 |
| SD = 60 | SD = 40 |

Four hundred draws are made at random with replacement from box A, and independently 100 draws are made at random with replacement from box B.



The problem is to find the expected value and standard error for the difference between the two sample averages. The first step is to compute the expected value and SE for each average separately (section 1 of chapter 23):

$$\text{average of 400 draws from box A} = 110 \pm 3 \text{ or so}$$
$$\text{average of 100 draws from box B} = 90 \pm 4 \text{ or so}$$

The expected value for the difference is just $110 - 90 = 20$. The next problem is how to put the SEs together:

$$(110 \pm 3) - (90 \pm 4) = 20 \pm _____?$$

A natural guess is to add the SEs: $3 + 4 = 7$. This ignores the possibility of cancellation in the two chance errors. The right SE, which is noticeably less than 7, can be found using a square root law.[1]

> The standard error for the difference of two independent quantities is $\sqrt{a^2 + b^2}$, where
> - $a$ is the SE for the first quantity;
> - $b$ is the SE for the second quantity.

In the example, the draws from the two boxes are made independently, so the two averages are independent, and the square root law applies. Now $a$ is 3 and $b$ is 4. So the SE for the difference between the two averages is

$$\sqrt{3^2 + 4^2} = \sqrt{25} = 5.$$

*Example 1.* One hundred draws are made at random with replacement from box C, shown below. Independently, 100 draws are made at random with replacement from box D. Find the expected value and SE for the difference between the number of 1's drawn from box C and the number of 4's drawn from box D.

(C) $\boxed{1\ 2}$ (D) $\boxed{3\ 4}$

*Solution.* The number of 1's will be around 50, give or take 5 or so. The number of 4's will also be around 50, give or take 5 or so. The expected value for the difference is $50 - 50 = 0$. The draws are made independently, so the two numbers are independent, and the square root law applies. The SE for the difference is $\sqrt{5^2 + 5^2} \approx 7$.

*Example 2.* One hundred draws are made at random with replacement from the box

$$\boxed{1\ 2\ 3\ 4}$$

The expected number of 1's is 25, with an SE of 4.3. The expected number of 4's is also 25 with an SE of 4.3. True or false: the SE for the difference between the number of 1's and the number of 4's is $\sqrt{4.3^2 + 4.3^2}$.

*Solution.* This is false. The two numbers are dependent: if one is large, the other is likely to be small. The square root law does not apply.

## Exercise Set A

1. Two hundred draws are made at random with replacement from the box in example 2. Someone is thinking about the difference

   "number of 1's in draws 1–100" − "number of 5's in draws 101–200"

   True or false, and explain: the SE for the difference is $\sqrt{4^2 + 4^2}$.

2. Box A has an average of 100 and an SD of 10. Box B has an average of 50 and an SD of 18. Now 25 draws are made at random with replacement from box A, and independently 36 draws are made at random with replacement from box B. Find the expected value and standard error for the difference between the average of the draws from box A and the average of the draws from box B.

3. A coin is tossed 500 times. Find the expected value and SE for the difference between the percentage of heads in the first 400 tosses and the percentage of heads in the last 100 tosses.

4. A coin is tossed 500 times. True or false, and explain.
   (a) The SE for the percentage of heads among the 500 tosses is 2.2 percentage points.
   (b) The SE for the percentage of tails among the 500 tosses is 2.2 percentage points.
   (c) The SE for the difference

   percentage of heads − percentage of tails

   is $\sqrt{2.2^2 + 2.2^2} \approx 3.1$ percentage points.

5. A box contains 5,000 numbered tickets, which average out to 50; the SD is 30. Two hundred tickets are drawn at random without replacement. True or false, and explain: the SE for the difference between the average of the first 100 draws and the average of the second 100 draws is approximately $\sqrt{3^2 + 3^2}$. (Hint: What if the draws were made with replacement?)

6. One hundred draws are made at random with replacement from box F: the average of these draws is 51 and their SD is 3. Independently, 400 draws are made at random with replacement from box G: the average of these draws is 48 and their SD is 8. Someone claims that both boxes have the same average. What do you think?

*The answers to these exercises are on pp. A94–95.*

## 2. COMPARING TWO SAMPLE AVERAGES

The National Assessment of Educational Progress (NAEP) monitors trends in school performance. Each year, NAEP administers tests on several subjects to a nationwide sample of 17-year-olds who are in school.[2] The reading test was given in 1990 and again in 2004. The average score went down from 290 to 285. The difference is 5 points. Is this real, or just a chance variation?

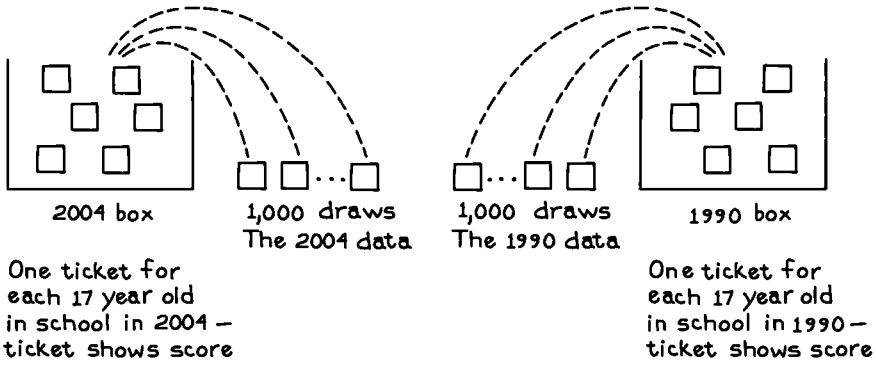
A $z$-test can be used, but the calculation is more complicated than it was in chapter 26. There, a sample average was compared to an external standard. Here, there are two samples, and the difference of their averages is the issue:

average score in 2004 sample − average score in 1990 sample.

Both averages are subject to chance variability, and the SE for the difference must take that into account. The method of section 1 can be used.

To compute standard errors, you need a box model, and the model depends on the design of the sample. In fact, the NAEP design was quite complicated, but a simplified version can be presented here. Suppose that in 2004 and in 1990, the test was administered to a nationwide simple random sample of one thousand 17-year-olds currently enrolled in school.

With this design, the model is straightforward. There have to be two boxes, one for each of the two test years. The 2004 box has millions of tickets—one for each person who was 17 years old back then, and enrolled in school. The number on the ticket shows what that person would have scored, if he or she had taken the NAEP reading test. The 2004 data are like 1,000 draws at random from the box. The 1990 box is set up the same way. That completes the model.




The null hypothesis says that the averages of the two boxes are equal. On that basis, the difference between the sample averages is expected to be 0, and the observed difference just reflects the luck of the draw. Schools are not getting worse. The alternative hypothesis says that the average of the 2004 box is smaller than the average of the 1990 box—reading scores really did go down, and that is why the two sample averages are different. The two-sample $z$-statistic will help in choosing between these hypotheses.

$$z = \frac{\text{observed}^{\text{difference}} - \text{expected}^{\text{difference}}}{\text{SE}_{\text{for difference}}}$$

We begin with the numerator of the $z$-statistic. It is the difference between the sample averages that is observed: $285 - 290 = -5$ points. Therefore, the relevant benchmark in the numerator of $z$ is the expected value of the difference. The

expected value is computed using the null hypothesis. On that basis, the difference between the two sample averages is expected to be 0. So the numerator of the $z$-statistic is

$$-5 - 0 = -5$$

Now the denominator. The SE for the difference between the sample averages is needed. Take the samples one at a time. In 2004, the SD of the 1,000 test scores turned out to be 37. So the SD of the 2004 box is estimated as 37. The SE for the sum of the 1,000 test scores in 2004 is estimated as $\sqrt{1{,}000} \times 37 \approx 1{,}170$. The SE for the average is $1{,}170/1{,}000 \approx 1.2$. In 1990, the SD was 40 and the SE for the 1990 average is 1.3. The SE for the difference can be computed using the method of the previous section, because the samples are independent:

$$\sqrt{1.2^2 + 1.3^2} \approx 1.8$$

Finally,

$$z \approx -5/1.8 \approx -2.8$$

In other words, the difference between 2004 and 1990 was about 2.8 SEs below the value expected on the null hypothesis—pushing the envelope of chance variation. We reject the null hypothesis, and are left with the alternative hypothesis that the difference is real. On the other hand, the difference is small, and other measures of school performance give more optimistic results. Chapter 29 continues the discussion.

The two-sample $z$-statistic is computed from—
- the sizes of the two samples,
- the averages of the two samples,
- the SDs of the two samples.

The test assumes two independent simple random samples.

With NAEP, the samples are big enough so that the probability histogram for each sample average follows the normal curve. Then $z$ follows the normal curve. The two-sample $z$-test can also be used for percents, as the next example shows.[3]

*Example 3.* In 1999, NAEP found that 13% of the 17-year-old students had taken calculus, compared to 17% in 2004. Is the difference real, or a chance variation?

*Solution.* Again, let's assume that in each of the two years, NAEP took a nationwide simple random sample of one thousand 17-year-olds currently enrolled in school. There are two samples here, so the two-sample $z$-test is needed rather than the one-sample test (chapter 26). As in example 2, there are two boxes, one for 2004 and one for 1999. The data are qualitative rather than quantitative, so 0's and 1's go on the tickets. A ticket is marked 1 if the student took calculus and 0 otherwise. The 2004 data are like 1,000 draws from 2004 box. The 1999 box can be set up the same way. The null hypothesis says that the percentage of 1's in

the two boxes is the same. The alternative hypothesis says that the percentage for the 2004 box is bigger than the percentage for the 1999 box.

To make the $z$-test, we need to put an SE on the difference between the sample percentages. Take the samples one at a time. The SE for the number of 1's in the 2004 sample is estimated as

$$\sqrt{1{,}000} \times \sqrt{0.17 \times 0.83} \approx 11$$

The SE for the percentage is

$$\frac{11}{1{,}000} \times 100\% = 1.2\%$$

Similarly, the SE for the 1999 percentage is 1.1%. The SE for the difference is

$$\sqrt{1.2^2 + 1.1^2} \approx 1.6\%$$

On the null hypothesis, the expected difference is 0%. The observed difference is $17 - 13 = 4\%$. So the test statistic is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}} \approx \frac{4\% - 0\%}{1.6\%} \approx 2.5$$

Again, we reject the null: $P \approx 1\%$.

## Exercise Set B

1. "Is the difference between two sample averages just due to chance?" To help answer this question, statisticians use a _____ _____ $z$-test. Fill in the blanks, and explain briefly.

2. In 1990 and 2004, NAEP tested the 17-year-olds on mathematics as well as reading. The average score went up from 305 to 307. You may assume the NAEP took simple random samples of size 1,000 in each of the two years; the SD for the 1990 data was 34, and the SD for the 2004 data was 27. (In fact, NAEP used a more complicated sample design.[4]) Can the difference between the 305 and 307 be explained as a chance variation?

   (a) Should you make a one-sample $z$-test or a two-sample $z$-test? Why?
   (b) Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? Do the tickets show test scores, or 0's and 1's. Why?
   (c) Now answer the main question: is the difference real, or can it be explained by chance?

3. In 1970, 59% of college freshmen thought that capital punishment should be abolished; by 2005, the percentage had dropped to 35%.[5] Is the difference real, or can it be explained by chance? You may assume that the percentages are based on two independent simple random samples, each of size 1,000.

4. A study reports that freshmen at public universities work 10.2 hours a week for pay, on average, and the SD is 8.5 hours; at private universities, the average is 8.1

hours and the SD is 6.9 hours. Assume these data are based on two independent simple random samples, each of size 1,000.[6] Is the difference between the averages due to chance? If not, what else might explain it?

5. A university takes a simple random sample of 132 male students and 279 females; 41% of the men and 17% of the women report working more than 10 hours during the survey week. To find out whether the difference in percentages is statistically significant, the investigator starts by computing $z = (41 - 17)/.048$. Is anything wrong?

6. Cycle III of the Health Examination Survey used a nationwide probability sample of youths age 12 to 17. One object of the survey was to estimate the percentage of youths who were illiterate.[7] A test was developed to measure literacy. It consisted of seven brief passages, with three questions about each, like the following:

   There were footsteps and a knock at the door. Everyone inside stood up quickly. The only sound was that of the pot boiling on the stove. There was another knock. No one moved. The footsteps on the other side of the door could be heard moving away.

   - The people inside the room
     (a) Hid behind the stove
     (b) Stood up quickly
     (c) Ran to the door
     (d) Laughed out loud
     (e) Began to cry
   - What was the only sound in the room?
     (a) People talking
     (b) Birds singing
     (c) A pot boiling
     (d) A dog barking
     (e) A man shouting
   - The person who knocked at the door finally
     (a) Walked into the room
     (b) Sat down outside the door
     (c) Shouted for help
     (d) Walked away
     (e) Broke down the door.

   This test was designed to be at the fourth-grade level of reading, and subjects were defined to be literate if they could answer more than half the questions correctly.

   There turned out to be some difference between the performance of males and females on this test: 7% of the males were illiterate, compared to 3% of the females. Is this difference real, or the result of chance variation? You may assume that the investigators took a simple random sample of 1,600 male youths, and an independent simple random sample of 1,600 female youths.

7. Cycle II of the Health Examination Survey used a nationwide probability sample of children age 6 to 11. One object of the survey was to study the relationship between the children's scores on intelligence tests and the family backgrounds.[8] The WISC vocabulary scale was used. This consists of 40 words which the child has to define; 2 points are given for a correct answer, and 1 point for a partially

correct answer. There was some relationship between test scores and the type of community in which the parents lived. For example, big-city children averaged 26 points on the test, and their SD was 10 points. But rural children only averaged 25 points with the same SD of 10 points. Can this difference be explained as a chance variation?

You may assume that the investigators took a simple random sample of 400 big-city children, and an independent simple random sample of 400 rural children.

8. Repeat the previous exercise, if both samples were of size 1,000 instead of 400.

9. Review exercise 12 in chapter 26 described an experiment in which 59 animals were put in treatment (enriched environment), and 59 were in control. The cortex weights for the treatment group averaged 683 milligrams, and the SD was 31 milligrams. The cortex weights for the control group averaged 647 milligrams, and the SD was 29 milligrams. Someone proposes to make a two-sample $z$-test:

$$\text{SE for sum of treatment weights} \approx \sqrt{59} \times 31 \approx 238 \text{ milligrams}$$
$$\text{SE for average of treatment weights} \approx 238/59 \approx 4.0 \text{ milligrams}$$
$$\text{SE for sum of control weights} \approx \sqrt{59} \times 29 \approx 223 \text{ milligrams}$$
$$\text{SE for average of control weights} \approx 223/59 \approx 3.8 \text{ milligrams}$$
$$\text{SE for difference} \approx \sqrt{4.0^2 + 3.8^2} \approx 5.5 \text{ milligrams}$$
$$z = 36/5.5 \approx 6.5, \quad P \approx 0$$

What does statistical theory say?

*The answers to these exercises are on pp. A95–96.*

## 3. EXPERIMENTS

The method of section 2 can also be used to analyze certain kinds of experimental data, where the investigators choose some subjects at random to get treatment "A" and others to get "B." In the Salk vaccine field trial, for instance, treatment A would be the vaccine; treatment B, the placebo given to the control group (chapter 1). We begin with an example to illustrate the mechanics, and then say why the method works.

*Example 4.* There are 200 subjects in a small clinical trial on vitamin C. Half the subjects are assigned at random to treatment (2,000 mg of vitamin C daily) and half to control (2,000 mg of placebo). Over the period of the experiment, the treatment group averaged 2.3 colds, and the SD was 3.1. The controls did a little worse: they averaged 2.6 colds and the SD was 2.9. Is the difference in averages statistically significant?

*Solution.* The difference between the two averages is $-0.3$, and you need to put a standard error on this number. Just pretend that you have two independent samples drawn at random with replacement. The SE for the treatment sum is $\sqrt{100} \times 3.1 = 31$; the SE for the treatment average is $31/100 = 0.31$. Similarly, the SE for the control average is 0.29. The SE for the difference is

$$\sqrt{0.31^2 + 0.29^2} \approx 0.42$$

Suppose the null hypothesis is right: vitamin C has no effect. On this basis, the expected value for the difference is 0.0. The observed difference was −0.3. So

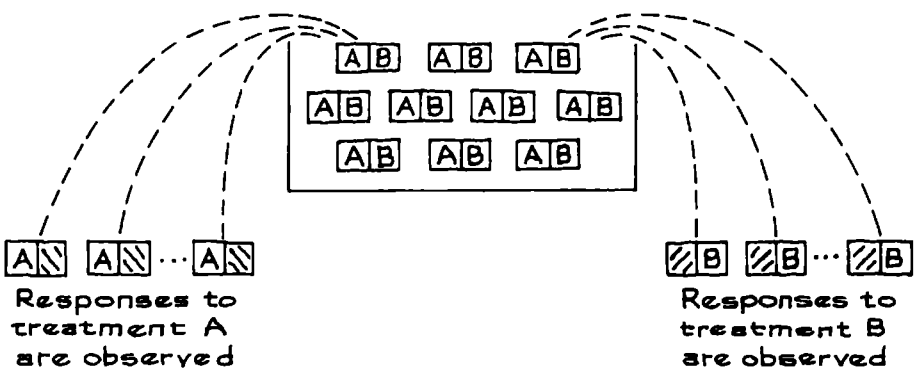
$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}} = \frac{-0.3 - 0.0}{0.42} \approx -0.7$$

The difference could easily be due to chance: a few too many susceptible people were assigned to the control group.[9]

Now, a look behind the scenes. In working the example, you were asked to pretend that the treatment and control samples were drawn independently, at random with replacement, from two boxes. However, the experiment wasn't done that way. There were 200 subjects; 100 were chosen at random—without replacement—to get the vitamin C; the other 100 got the placebo. So the draws are made without replacement. Furthermore, the samples are dependent. For instance, one subject might be quite susceptible to colds. If this subject is in the vitamin C group, he cannot be in the placebo group. The assignment therefore influences both averages.

Why does the SE come out right, despite these problems? The reasoning depends on the box model. The investigators are running an experiment. They choose one group of subjects at random to get treatment A and another group to get treatment B. As usual, the model has a ticket for each subject. But now the ticket has two numbers. One shows what the response would be to treatment A; the other, to treatment B. See figure 1. Only one of the two numbers can be observed, because the subject can be given only one of the two treatments.

Figure 1. A randomized controlled experiment comparing treatments A and B. There is a ticket for each subject. The ticket has two numbers: one shows the subject's response to treatment A; the other, to treatment B. Only one of the two numbers can be observed.




In the model, some tickets are drawn at random without replacement from the box and the responses to treatment A are observed. The data on treatment A are like this first batch of responses. Then, more draws are made at random without replacement from the box and the responses to treatment B are observed. The data on treatment B are like this second batch of responses. In example 4, every one

of the 200 subjects was assigned either to vitamin C or to the placebo. In such a case, the second sample just amounts to the tickets left behind in the box after the first sample has been drawn.

The null hypothesis says that the response is the same for both treatments.[10] To test this hypothesis, investigators usually compare averages (or percents):

average response in group A − average response in group B.

What is the SE for this difference? The solution to example 4 seems to involve the two mistakes mentioned earlier—

- The draws are made without replacement, but the SEs are computed as if drawing with replacement.
- The two averages are dependent, but the SEs are combined as if the averages were independent.

When the number of draws is small relative to the number of tickets in the box, neither mistake is serious. There is little difference between drawing with or without replacement, and the dependence between the averages is small too. There almost are two separate boxes, one for the treatment group and one for the controls. However, the "two-box" model is unrealistic for a randomized controlled experiment—unless the subjects really are chosen as a random sample from a large population. That is unusual, although exercise 8 (p. 520) gives one example.

If the number of draws is large relative to the size of the box—and this is the usual case—then the impact of each mistake by itself can be substantial. For instance, when half the subjects are assigned to each treatment group, as in example 4, the correction factor will be noticeably less than 1 (section 4 of chapter 20). Dependence can also be strong. It is a lucky break that when applied to randomized experiments, the procedure of section 2 is conservative, tending to overestimate the SE by a small amount. That is because the two mistakes offset each other.

- The first mistake inflates the SE.
- The second mistake cuts the SE back down.

There is a box of tickets. Each ticket has two numbers. One shows what the response would be to treatment A; the other, to treatment B. Only one of the numbers can be observed. Some tickets are drawn at random without replacement from the box. In this sample, the responses to treatment A are observed. Then, a second sample is drawn at random without replacement from the remaining tickets. In the second sample, the responses to treatment B are observed.

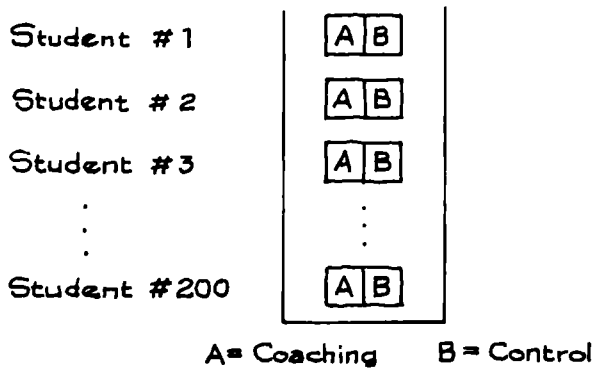
The SE for the difference between the two sample averages can be conservatively estimated as follows:

(i) compute the SEs for the averages as if the draws were made with replacement;
(ii) combine the SEs as if the samples were independent.

To make the mathematics work, the SEs for the two sample averages must be computed on the basis of drawing WITH replacement—even though the draws are made WITHOUT replacement. That is what compensates for the dependence between the two samples.[11] In summary: when the data come from a randomized experiment (like example 4), the procedure of section 2 can be used even though there is dependence.

## Exercise Set C

1. (Hypothetical.) Does coaching for the Math SATs work? A group of 200 high-school seniors volunteer as subjects for an experiment; 100 are selected at random for coaching, and the remaining 100 are controls. After six months, all 200 subjects take the Math SAT. A box model is set up for this experiment, as shown.




   (a) John Doe participated in the experiment—he was student #17. He got assigned to the coaching group. There was a ticket in the box for him. Did this ticket have one number on it, or two numbers?
   (b) His sister Jane Doe participated also (she was student #18). She was assigned to the control group—no coaching. Did her ticket have an A-number on it? If so, what does this number mean? Do the investigators know what this number was?
   (c) The coaching group averaged 486 on the Math SAT; their SD was 98. The control group averaged 477, and had an SD of 103. Did the coaching work? Or was it chance?

2. (Hypothetical.) Is Wheaties a power breakfast? A study is done in an elementary statistics class; 499 students agree to participate. After the midterm, 250 are randomized to the treatment group, and 249 to the control group. The treatment group is fed Wheaties for breakfast 7 days a week. The control group gets Sugar Pops.
   (a) Final scores averaged 66 for the treatment group; the SD was 21. For the control group, the figures were 59 and 20. What do you conclude?
   (b) What aspects of the study could have been done "blind?"

3. This continues exercise 2.
   (a) Midterm scores averaged 61 for the treatment group; the SD was 20. For the control group, the figures were 60 and 19. What do you conclude?

(b) Repeat, if the average midterm score for the treatment group is 68, and the SD is 21; for the control group, the figures are 59 and 18.

4. Suppose the study in example 4 is repeated on 2,000 subjects, with 1,000 assigned to the vitamin C group, and 1,000 to control. Suppose the average number of colds in the vitamin C group is 2.4 and the SD is 2.9; the average in the control group is 2.5 and the SD is 3.0.

   (a) Is the difference in averages statistically significant? What do you conclude?
   (b) Why would the averages change from one study to the next?

5. In the box below, each ticket has a left-hand number and a right-hand number:

| 0 4 | 2 0 | 3 6 | 4 12 | 6 8 |

(For instance, the left-hand number on [0 4] is 0 and the right-hand number is 4.) One hundred draws are made at random with replacement from this box. One investigator computes the average of the left-hand numbers. A second investigator computes the average of the right-hand numbers. True or false, and explain—

   (a) The SE for the first average is 0.2.
   (b) The SE for the second average is 0.4.
   (c) The SE for the difference of the two averages is $\sqrt{0.2^2 + 0.4^2}$.

*The answers to these exercises are on pp. A96–97.*

## 4. MORE ON EXPERIMENTS

The technique described in the previous section can also be used for experiments where the response is qualitative rather than quantitative, so the tickets must show 0's and 1's. This section will give an example; but first, some background material. The standard theory of economic behavior assumes "rational" decision making, according to certain formal (and perhaps unrealistic) rules. In particular, the theory says that decision makers respond to facts, not to the way the facts are presented. Psychologists, on the other hand, tend to think that "framing"—the manner of presentation—counts. Empirical work favors the psychological view.[12]

One study, by Amos Tversky and others, involved presenting information on the effectiveness of surgery or radiation as alternative therapies for lung cancer. The subjects were a group of 167 doctors in a summer course at Harvard.[13] The information was presented in two different ways. Some of the doctors got form A, which reports death rates.

> *Form A)* Of 100 people having surgery, 10 will die during treatment, 32 will have died by one year, and 66 will have died by five years. Of 100 people having radiation therapy, none will die during treatment, 23 will die by one year, and 78 will die by five years.

Other doctors got form B, which reports survival rates.

> *Form B)* Of 100 people having surgery, 90 will survive the treatment, 68 will survive one year or longer, and 34 will survive five years or longer. Of 100 people having radiation therapy, all will survive the treatment, 77 will survive one year or longer, and 22 will survive five years or longer.

Both forms contain exactly the same information. For example, 10 patients out of 100 will die during surgery (form A), so 90 out of 100 will survive (form B). By the fifth year, the outlook for lung cancer patients is quite bleak.

In the experiment, 80 of the 167 doctors were picked at random and given form A. The remaining 87 got form B. After reading the form, each doctor wrote down the therapy he or she would recommend for a lung cancer patient. In response to form A, 40 out of 80 doctors chose surgery (table 1). But in response to form B, 73 out of 87 favored surgery: 40/80 is 50%, and 73/87 is 84%. Style of presentation seems to matter.

Table 1. Results from an experiment on the effect of presentation of data.

| | *Form A* | *Form B* |
|---|---|---|
| Favored surgery | 40 | 73 |
| Favored radiation | 40 | 14 |
| Total | 80 | 87 |
| Percent favoring surgery | 50% | 84% |

An economist who is defending standard decision theory might argue that the difference is just due to chance. Based on the information, which is the same on both forms, some doctors will recommend surgery while others will choose radiation. But the decision can't depend on how the information is presented. By the luck of the draw, the economist might say, too many of the doctors who favor surgery were picked to get form B, and too few of them got form A. After all, there were only 80 people in group A, and 87 in group B. There seems to be a lot of room for chance variation.

To evaluate this argument, a significance test is needed. The difference between the two percentages in table 1 is 34%, and you need to put a standard error on this difference. The method of example 4 can be used. Pretend that you have two independent samples, drawn at random with replacement. The first sample is the doctors who got form A. There were 80 of them, and 40 favored surgery. The SE for the number favoring surgery is

$$\sqrt{80} \times \sqrt{0.50 \times 0.50} \approx 4.5$$

The SE for the percentage favoring surgery is $4.5/80 \times 100\% \approx 5.6\%$. The second sample is the doctors who got form B, and the SE for the percentage favoring surgery in response to form B is 3.9%. The SE for the difference is

$$\sqrt{5.6^2 + 3.9^2} \approx 6.8\%$$

On the null hypothesis, the expected difference between the percentages in the two samples is 0.0%. The observed difference is 34%. The test statistic is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}} = \frac{34\% - 0.0\%}{6.8\%} = 5.0$$

Chance is not a good explanation for the results in table 1.

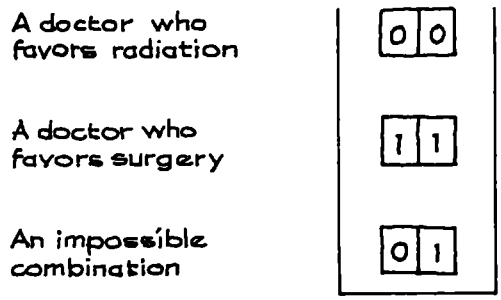
Another look behind the scenes: A box model for this experiment needs a ticket for each doctor. As in the previous section, each ticket shows a pair of

numbers [A|B]. The first number on the ticket codes the response to form A. It is 1 if the doctor would favor surgery when presented with form A, and 0 if she would prefer radiation. Similarly, the second number on the ticket codes the response to form B.

Eighty draws are made at random without replacement from the box, and the responses to form A are observed [A|▧]. The responses to form A in the experiment are like this first batch of 80 draws. The 50% in table 1 is like the percentage of 1's in this batch of draws. The 87 tickets left in the box are the second sample. With this second sample, the responses to form B are observed [▨|B]. The responses to form B are like this second batch of 0's and 1's. The 84% in table 1 is like the percentage of 1's in the second sample.

Now the null hypothesis can be set up in terms of the model. Because both forms convey the same information, the economist thinks that a doctor's response to the two forms must be the same, so both numbers on the ticket are the same (figure 2). The box model can be used to show that our method gives a conservative estimate for the SE.[14]

Figure 2. The null hypothesis for the experiment: deciding between radiation and surgery based on form A or form B with the same information. The first number on the ticket codes the response to form A; the second, to form B. Responses favoring surgery are coded "1."




The experimental design used in this study may seem a bit indirect. Why not give both forms of the questionnaire, one after the other, to all the doctors? The reason is simple. Asking both questions at the same time pushes the subjects to be more consistent: perhaps they see that both forms describe the same data.[15]

## Exercise Set D

1. The study described in the text was replicated on another group of subjects: MBA students at Harvard and Stanford.
   (a) One MBA student would prefer radiation therapy if presented with form A, but surgery if given form B. Fill in his ticket [A|B].
   (b) Another MBA student has the ticket [1|0]. How would she respond to form A? form B?

(c) Which of the three tickets is consistent with the null hypothesis?

(i) $\boxed{1}\boxed{0}$ (ii) $\boxed{0}\boxed{0}$ (iii) $\boxed{0}\boxed{1}$

(d) The results came out as follows.

| | *Form A* | *Form B* |
|---|---|---|
| Favored surgery | 112 | 84 |
| Favored radiation | 84 | 17 |

Can the difference in response to the forms be explained by chance? (Hint: to get started, find how many students got form A; of them, what percentage favored radiation? Then do the same for form B.)

2. In the Salk vaccine field trial, 400,000 children were part of a randomized controlled double-blind experiment. Just about half of them were assigned at random to the vaccine group, and the other half to the placebo.[16] In the vaccine group, there were 57 cases of polio, compared to 142 in the placebo group. Is this difference due to chance? If not, what explains it?

3. (a) In the HIP trial (pp. 22–23), there were 39 deaths from breast cancer in the treatment group, and 63 deaths in the control group. Is the difference statistically significant?

   (b) In the treatment group, there were 837 deaths from all causes, compared to 879 in the control group. Is the difference statistically significant?

4. Many observational studies conclude that low-fat diets protect against cancer and cardiovascular "events" (heart attacks, stroke, and so forth). Experimental results, however, are generally negative. In 2006, the Women's Health Initiative (WHI) published its results.[17] This was a large-scale randomized trial on women who had reached menopause. As one part of the study, 48,835 women were randomized: 19,541 were assigned to the treatment group and put on a low-fat diet. The other 29,294 women were assigned to the control group and ate as they normally would. Subjects were followed for 8 years.

   Among other things, the investigators found that 1,357 women on the low-fat diet experienced at least one cardiovascular event, compared to 2,088 in the control group. Can the difference between the two groups be explained by chance? What do you conclude about the effect of the low-fat diet?

5. A geography test was given to a simple random sample of 250 high-school students in a certain large school district. One question involved an outline map of Europe, with the countries identified only by number. The students were asked to pick out Great Britain and France. As it turned out, 65.6% could find France, compared to 70.4% for Great Britain.[18] Is the difference statistically significant? Or can this be determined from the information given?

6. Some years, the Gallup Poll asks respondents how much confidence they have in various American institutions. You may assume that results are based on a simple random sample of 1,000 persons each year; the samples are independent from year to year.[19]

   (a) In 2005, only 41% of the respondents had "a great deal or quite a lot" of confidence in the Supreme Court, compared to 50% in 2000. Is the difference real? Or can you tell from the information given?

(b) In 2005, only 22% of the respondents had "a great deal or quite a lot" of confidence in Congress, whereas 24% of the respondents had "a great deal or quite a lot" of confidence in organized labor. Is the difference between 24% and 22% real? Or can you tell from the information given?

Discuss briefly.

7. Breast-feeding infants for the first few months after their birth is considered to be better for their health than bottle feeding. According to several observational studies, withholding the bottle in hospital nurseries increases the likelihood that mothers will continue to breast-feed after leaving the hospital. As a result, withholding supplementation has been recommended.

A controlled experiment was done by K. Gray-Donald, M. S. Kramer, and associates at the Royal Victoria Hospital in Montreal.[20] There were two nurseries. In the "traditional" nursery, supplemental bottle-feedings were given as usual—at 2 A.M., and whenever the infant seemed hungry. In the experimental nursery, mothers were awakened at 2 A.M. and asked to breast-feed their babies; bottle-feeding was discouraged.

Over the four-month period of the experiment, 393 mothers and their infants were assigned at random to the traditional nursery, and 388 to the experimental one. The typical stay in the hospital was 4 days, and there was followup for 9 weeks after release from the hospital.

(a) At the end of 9 weeks, 54.7% of the mothers who had been assigned to the traditional nursery were still breast-feeding their infants, compared to 54.1% in the experimental nursery. Is this difference statistically significant? What do you conclude?

(b) It was really up to the mothers whether to breast-feed or bottle-feed. Were their decisions changed by the treatments? To answer that question, the investigators looked at the amounts of bottle-feeding in the two nurseries, expressed as milliliters per day (ml/day). In the traditional nursery, this averaged 36.6 ml/day per infant, and the SD was 44.3. In the experimental nursery, the figures were 15.7 and 43.6. What do you conclude?

(c) Did the different treatments in the two nurseries affect the infants in any way? To answer that question, the investigators looked at the weight lost by each infant during the stay, expressed as a percentage of birth weight. In the traditional nursery, this averaged 5.1% and the SD was 2.0%. In the experimental nursery, the average was 6.0% and the SD was 2.0%. What do you conclude? (It may be surprising, but most newborns lose a bit of weight during the first few days of life.)

(d) Was the randomization successful? To find out, the investigators looked at the birth weights themselves (among other variables). In the traditional nursery, these averaged 3,486 grams and the SD was 438 grams. In the experimental nursery, the average was 3,459 grams and the SD was 434 grams. What do you conclude?

*The answers to these exercises are on pp. A97–98.*

## 5. WHEN DOES THE $z$-TEST APPLY?

The square root law in section 1 was designed for use with two independent simple random samples. Example 1 in section 1 illustrates this application. So do the NAEP results in section 2. The procedure can also be used with a randomized controlled experiment, where each subject has two possible responses but only one is observed. The investigators see the response to treatment for the subjects who are randomly selected into the treatment group. They see the other response for subjects in the control group. Sections 3 and 4 (vitamin C and rational decision making) illustrate this application, which involves a minor miracle—two mistakes that cancel.

You are not expected to derive the formulas, but you should learn when to use them and when not to. The formulas should not be used when two correlated responses are observed for each subject. Exercise 5 on p. 515 (the geography test) is an example of when not to use the formulas. Each subject makes two responses, by answering (i) the question on Great Britain, and (ii) the question on France. Both responses are observed, because each subject answers both questions. And the responses are correlated, because a geography whiz is likely to be able to answer both questions correctly, while someone who does not pay attention to maps is likely to get both of them wrong. By contrast, if you took two independent samples—asking one group about France and the other about Great Britain—the formula would be fine. (That would be an inefficient way to do the study.)

Exercise 9 on p. 508 is another case when you should not use the formulas. This is a bit subtle, because the data were collected in a randomized controlled experiment—but you get two correlated responses for each of the 59 pairs of animals. By contrast, if 59 of the 118 rats had been selected at random and put into treatment, while the remaining 59 were used as controls, our formulas would be fine. (Again, the design used by the investigators turns out to be more efficient.)

> The $z$-test (sections 1 and 2) applies to two independent samples. Generally, the formulas give the wrong answer when applied to dependent samples. There is an exception: the $z$-test can be used to compare the treatment and control groups in a randomized controlled experiment—even though the groups are dependent (sections 3 and 4).

The square root law in section 1 gives the wrong answer with dependent samples because it does not take the dependence into account. Other formulas are beyond our scope. However, it is easy to do the $z$-test on the differences, as in exercise 12 on pp. 498–499.[21] Also see exercise 6 on pp. 258–259, exercise 11 on pp. 262–263, exercise 15 on p. 329, or exercise 11 on p. 488, which all use a technique called "the sign test."

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. Five hundred draws are made at random with replacement from a box of numbered tickets; 276 are positive. Someone tells you that 50% of the tickets in the box show positive numbers. Do you believe it? Answer yes or no, and explain.

2. One hundred draws are made at random with replacement from box A, and 250 are made at random with replacement from box B.
   (a) 50 of the draws from box A are positive, compared to 131 from box B: 50.0% versus 52.4%. Is this difference real, or due to chance?
   (b) The draws from box A average 1.4 and their SD is 15.3; the draws from box B average 6.3 and their SD is 16.1. Is the difference between the averages statistically significant?

3. The Gallup poll asks respondents how they would rate the honesty and ethical standards of people in different fields—very high, high, average, low, or very low.[22] The percentage who rated clergy "very high or high" dropped from 60% in 2000 to 54% in 2005. This may have been due to scandals involving sex abuse; or it may have been a chance variation. (You may assume that in each year, the results are based on independent simple random samples of 1,000 persons in each year.)
   (a) Should you make a one-sample $z$-test or a two-sample $z$-test? Why?
   (b) Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?
   (c) Can the difference between 60% and 54% be explained as a chance variation? Or was it the scandals? Or something else?

4. This continues exercise 3. In 2005, 65% of the respondents gave medical doctors a rating of "very high or high," compared to a 67% rating for druggists. Is the difference real, or a chance variation? Or do you need more information to decide? If the difference is real, how would you explain it? Discuss briefly. You may assume that the results are based on a simple random sample of 1,000 persons taken in 2005; each respondent rated clergy, medical doctors, druggists, and many other professions.[23]

5. One experiment involved 383 students at the University of British Columbia. 200 were chosen at random to get item A, and 92 of them answered "yes." The other 183 got item B, and 161 out of the second group answered "yes."[24]

   *Item A)* Imagine that you have decided to see a play and paid the admission price of $20 per ticket. As you enter the theatre, you discover that you have lost the ticket. The seat was not marked, and the ticket cannot be recovered. Would you pay $20 for another ticket?

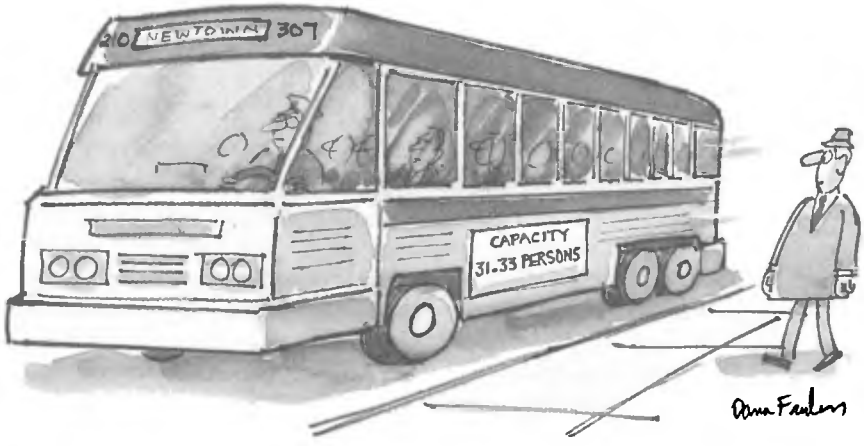
*Item B)* Imagine that you have decided to see a play where admission is $20 per ticket. As you enter the theatre, you discover that you have lost a $20 bill. Would you still pay $20 for a ticket for the play? [In Canada, "theatre" is the right spelling.]

From the standpoint of economic theory, both items present the same facts and call for the same answer; any difference between them must be due to chance. From a psychological point of view, the framing of the question can be expected to influence the answer. What do the data say?

6. An experiment is performed to see whether calculators help students do word problems.[25] The subjects are a group of 500 thirteen-year-olds in a certain school district. All the subjects work the problem below. Half of them are chosen at random and allowed to use calculators; the others do the problem with pencil and paper. In the calculator group, 18 students get the right answer; in the pencil-and-paper group, 59 do. Can this difference be explained by chance? What do you conclude?

   *The problem.* An army bus holds 36 soldiers. If 1,128 soldiers are being bussed to their training site, how many buses are needed?

   *Note.* 1,128/36 = 31.33, so 32 buses are needed. However, 31.33 was a common answer, especially in the calculator group; 31 was another common answer.




7. When convicts are released from prison, they have no money, and there is a high rate of "recidivism:" the released prisoners return to crime and are arrested again. Would providing income support to ex-convicts during the first months after their release from prison reduce recidivism? The Department of Labor ran a randomized controlled experiment to find out.[26] The experiment was done on a selected group of convicts being released from certain prisons in Texas and Georgia. Income support was provided, like unemployment

insurance. There was a control group which received no payment, and four different treatment groups (differing slightly in the amounts paid).

The exercise is on the results for Georgia, and combines the four treatment groups into one. Assume that prisoners were randomized to treatment or control.

(a) 592 prisoners were assigned to the treatment group, and of them 48.3% were rearrested within a year of release. 154 were assigned to the control group, and of them 49.4% were rearrested within a year of release. Did income support reduce recidivism? Answer yes or no, and explain briefly.

(b) In the first year after their release from prison, those assigned to the treatment group averaged 16.8 weeks of paid work; the SD was 15.9 weeks. For those assigned to the control group, the average was 24.3 weeks; the SD was 17.3 weeks. Did income support reduce the amount that the ex-convicts worked? Answer yes or no, and explain briefly.

8. One experiment contrasted responses to "prediction-request" and to "request-only" treatments, in order to answer two research questions.[27]

(i) Can people predict how well they will behave?

(ii) Do their predictions influence their behavior?

In the prediction-request group, subjects were first asked to predict whether they would agree to do some volunteer work. Then they were requested to do the work. In the request-only group, the subjects were requested to do the work; they were not asked to make predictions beforehand. In parts (a-b-c), a two-sample $z$-test may or may not be legitimate. If it is legitimate, make it. If not, why not?

(a) 46 residents of Bloomington, Indiana were chosen at random for the "prediction-request" treatment. They were called and asked to predict "whether they would agree to spend 3 hours collecting for the American Cancer Society if contacted over the telephone with such a request." 22 out of the 46 said that they would. Another 46 residents of that town were chosen at random for the "request-only" treatment. They were requested to spend the 3 hours collecting for the American Cancer Society. Only 2 out of 46 agreed to do it. Can the difference between 22/46 and 2/46 be due to chance? What do the data say about the research questions (i) and (ii)?

(b) Three days later, the prediction-request group was called again, and requested to spend 3 hours collecting for the American Cancer Society: 14 out of 46 agreed to do so. Can the difference between 14/46 and 2/46 be due to chance? What do the data say about the research questions (i) and (ii)?

(c) Can the difference between 22/46 and 14/46 be due to chance? What do the data say about the research questions (i) and (ii)?

9. A researcher wants to see if the editors of journals in the field of social work are biased. He makes up two versions of an article, "in which an asthmatic child was temporarily separated from its parents in an effort to relieve the symptoms of an illness that is often psychosomatic." In one version, the separation has a positive effect; in another, negative.[28] The article is submitted to a group of 107 journals; 53 are chosen at random to get the positive version, and 54 get the negative one. The results are as follows:

| | *Positive* | *Negative* |
|---|---|---|
| Accept | 28 | 8 |
| Reject | 25 | 46 |

The first column of the table says that 28 of the journals getting the positive version accepted it for publication, and 25 rejected it. The second column gives the results for the journals that got the negative version. Is chance a good explanation for the results? If not, what can be concluded about journal publication policy?

10. An investigator wants to show that first-born children score higher on IQ tests than second-borns. He takes a simple random sample of 400 two-child families in a school district, both children being enrolled in elementary school. He gives these children the WISC vocabulary test (described in exercise 7 on pp. 507–508), with the following results.

    - The 400 first-borns average 29 and their SD is 10.
    - The 400 second-borns average 28 and their SD is 10.

(Scores are corrected for age differences.) He makes a two-sample $z$-test:

SE for first-born average $\approx 0.5$

SE for second-born average $\approx 0.5$

SE for difference $= \sqrt{0.5^2 + 0.5^2} \approx 0.7$

$z = 1/0.7 \approx 1.4, \quad P \approx 8\%$

Comment briefly on the use of statistical tests.

11. (Hard.) The logic of the two-sample $z$-test in section 27.2 relies on two mathematical facts: (i) the expected value of a difference equals the difference of the expected values, and (ii) the expected value of the sample average equals the population average. Explain briefly, with reference to the NAEP reading scores.

## 7. SUMMARY

1. The expected value for the difference of two quantities equals the difference of the expected values. (Independence is not required here.)

2. The standard error for the difference of two independent quantities is $\sqrt{a^2 + b^2}$, where

- $a$ is the SE for the first quantity;
- $b$ is the SE for the second quantity.

For dependent quantities, this formula is usually wrong.

3. Suppose that two independent and reasonably large simple random samples are taken from two separate boxes. The null hypothesis is about the difference between the averages of the two boxes. The appropriate test statistic is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}}$$

In the formula, the "difference" is between the averages of the two samples. (If the null hypothesis says that the two boxes have the same average, the expected difference between the sample averages is 0.)

4. Tests based on this statistic are called *two-sample z-tests*.

5. The two-sample $z$-test can handle situations which involve classifying and counting, by putting 0's and 1's in the boxes.

6. The two-sample $z$-test can also be used to compare treatment and control averages or rates in an experiment. Suppose there is a box of tickets. Each ticket has two numbers: one shows what the response would be to treatment A; the other, to treatment B. For each ticket, only one of the two numbers can be observed. Some tickets are drawn at random without replacement from the box, and the responses to treatment A are observed. Then, a second sample is drawn at random without replacement from the remaining tickets. In the second sample, the responses to treatment B are observed. The SE for the difference between the two sample averages can be conservatively estimated as follows:

(i) compute the SEs for the averages as if drawing with replacement;
(ii) combine the SEs as if the two samples were independent.

# 28

# The Chi-Square Test

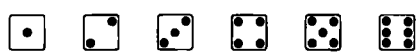
*Don't ask what it means, but rather how it is used.*
—L. WITTGENSTEIN (1889–1951)

## 1. INTRODUCTION

How well does it fit the facts? Sooner or later, this question must be asked about any chance model. And in many cases, it can be settled by the $\chi^2$-*test* (invented in 1900 by Karl Pearson).[1] $\chi$ is a Greek letter, often written as "chi," read like the "ki" in kite, so $\chi^2$ is read as "ki-square." Section 5 of chapter 26 explained how to test a chance model for a parapsychology experiment. There, each guess was classified into one of two categories—right or wrong. According to the model, a guess had 1 chance in 4 to be right, so the number of correct guesses was like the sum of draws from the box

$$\boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{1}$$

In that case, the $z$-test was appropriate, but only two categories were involved. If there are more than two categories, statisticians use the $\chi^2$-test rather than the $z$-test. For instance, you might want to see if a die is fair. Each throw can be classified into one of 6 categories:



1 2 3 4 5 6

The $\chi^2$-test will help to check whether these categories are equally likely, as in the next example.

*Example 1.* A gambler is accused of using a loaded die, but he pleads innocent. A record has been kept of the last 60 throws (table 1). There is disagreement about how to interpret the data and a statistician is called in.

Table 1. Sixty rolls of a die, which may be loaded.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 3 | 3 | 1 | 2 | 3 | 4 | 6 | 5 | 6 |
| 2 | 4 | 1 | 3 | 3 | 5 | 3 | 4 | 3 | 4 |
| 3 | 3 | 4 | 5 | 4 | 5 | 6 | 4 | 5 | 1 |
| 6 | 4 | 4 | 2 | 3 | 3 | 2 | 4 | 4 | 5 |
| 6 | 3 | 6 | 2 | 4 | 6 | 4 | 6 | 3 | 2 |
| 5 | 4 | 6 | 3 | 3 | 3 | 5 | 3 | 1 | 4 |

*Discussion.* If the gambler is innocent, the numbers in table 1 are like the results of drawing 60 times (at random with replacement) from the box

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

According to this model, each number should turn up about 10 times: the *expected frequency* is 10. To find out how the data compare with expectations, you have to count and see how many times each number did in fact turn up. The *observed frequencies* are shown in table 2. A check on the arithmetic: the sum of each frequency column must be 60, the total number of entries in table 1. ("Frequency" is statistical jargon for the number of times something happens.)

Table 2. Observed and expected frequencies for the data in table 1.

| *Value* | *Observed frequency* | *Expected frequency* |
|---|---|---|
| 1 | 4 | 10 |
| 2 | 6 | 10 |
| 3 | 17 | 10 |
| 4 | 16 | 10 |
| 5 | 8 | 10 |
| 6 | 9 | 10 |
| sum | 60 | 60 |

As the table indicates, there are too many 3's. The SE for the number of 3's is $\sqrt{60} \times \sqrt{1/6 \times 5/6} \approx 2.9$, so the observed number is about 2.4 SEs above the expected number. But don't shoot the gambler yet. The statistician won't advise taking the table one line at a time.

- Several lines in the table may look suspicious. For example, in table 2 there are also too many 4's.
- On the other hand, with many lines in the table, there is high probability that at least one of them will look suspicious—even if the die is fair. It's like playing Russian roulette. If you keep on going, sooner or later you're going to lose.

For each line of the table, there is a difference between observed and expected frequencies. The idea is to combine all these differences into one overall measure of the distance between the observed and expected values. What $\chi^2$ does is to square each difference, divide by the corresponding expected frequency, and take the sum:

$$\chi^2 = \text{sum of } \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

There is one term for each line in the table. At first sight, the formula may seem quite arbitrary. However, every statistician uses it because of one very convenient feature, which will be pointed out later.

With the data in table 2, the $\chi^2$-statistic is

$$\frac{(4-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} + \frac{(16-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} = \frac{142}{10} = 14.2$$

When the observed frequency is far from the expected frequency, the corresponding term in the sum is large; when the two are close, this term is small. Large values of $\chi^2$ indicate that observed and expected frequencies are far apart. Small values of $\chi^2$ mean the opposite: observeds are close to expecteds. So $\chi^2$ does give a measure of the distance between observed and expected frequencies.[2]

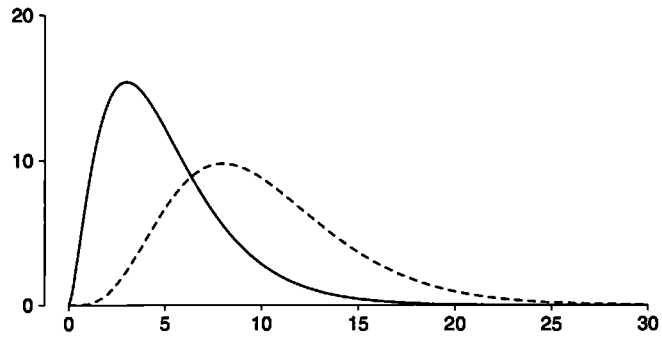
Of course, even if the data in table 1 had been generated by rolling a fair die 60 times, $\chi^2$ could have turned out to be 14.2, or more—the chance variation defense. Is this plausible? To find out, we need to know the chance that when a fair die is rolled 60 times and $\chi^2$ is computed from the observed frequencies, its value turns out to be 14.2 or more.

Why "or more"? The observed value 14.2 may be evidence against the model because it is too big, meaning that the observed frequencies are too far from the expected frequencies. If so, values larger than 14.2 would be even stronger evidence against the model. What is the chance that the model will produce such strong evidence against itself? To find out, we calculate the chance of getting a $\chi^2$-statistic of 14.2 or more.

Calculating this chance looks like a big job, but the computer does it in a flash, and the answer is 1.4%. If the die is fair, there is only a 1.4% chance for it to produce a $\chi^2$-statistic as big as (or bigger than) the observed one. At this point, the statistician has finished. Things do not look good for the gambler.

The 1.4% is called "the observed significance level" and denoted by $P$, as in chapter 26. In Pearson's time, there were no computers to find the chances. So he developed a method for approximating $P$ by hand. This method involved a new curve, called the *$\chi^2$-curve*. More precisely, there is one curve for each number of *degrees of freedom*.[3] The curves for 5 and 10 degrees of freedom are shown in figure 1.

Figure 1. The $\chi^2$-curves for 5 and 10 degrees of freedom. The curves have long right-hand tails. As the degrees of freedom go up, the curves flatten out and move off to the right. (The solid curve is for 5 degrees of freedom; dashed, for 10.)




Sometimes, it is hard to work out the degrees of freedom. However, in example 1, the model was *fully specified*. There were no parameters to estimate from the data, because the model told you what was in the box. When the model is fully specified, computing the degrees of freedom is easy:

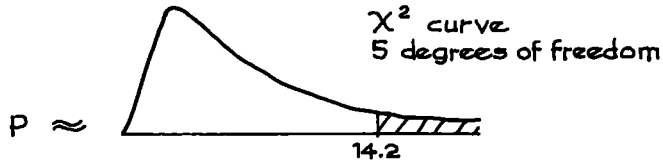
$$\text{degrees of freedom} = \text{number of terms in } \chi^2 - \text{one}.$$

In example 1, there are $6 - 1 = 5$ degrees of freedom. Why? In table 2, the six observed frequencies have to add up to 60. If you know any five of them, you can compute the sixth. Only five of the frequencies can vary freely. (Compare section 6 of chapter 26.)

For the $\chi^2$-test, $P$ is approximately equal to the area to the right of the observed value for the $\chi^2$-statistic, under the $\chi^2$-curve with the appropriate number of degrees of freedom. When the model is fully specified (no parameters to estimate),

degrees of freedom = number of terms in $\chi^2$ − one.

For example 1,




This area can be found using tables or a statistical calculator. In principle, there is one table for each curve but this would be so awkward that a different arrangement is used, as shown in table 3 (extracted from a bigger one on p. A106). Areas, in percent, are listed across the top of the table; degrees of freedom are listed down the left side. For instance, look at the column for 5% and the row for 5 degrees of freedom. In the body of the table there is the entry 11.07, meaning that the area to the right of 11.07 under the curve for 5 degrees of freedom is 5%. The area to the right of 14.2 under the curve for 5 degrees of freedom cannot be

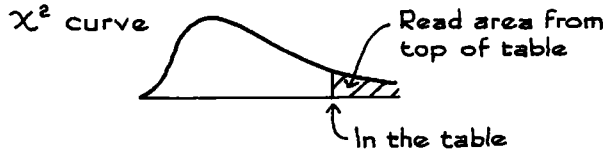
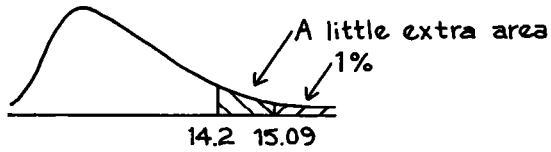



Table 3. A short $\chi^2$ table extracted from the bigger one on p. A106.

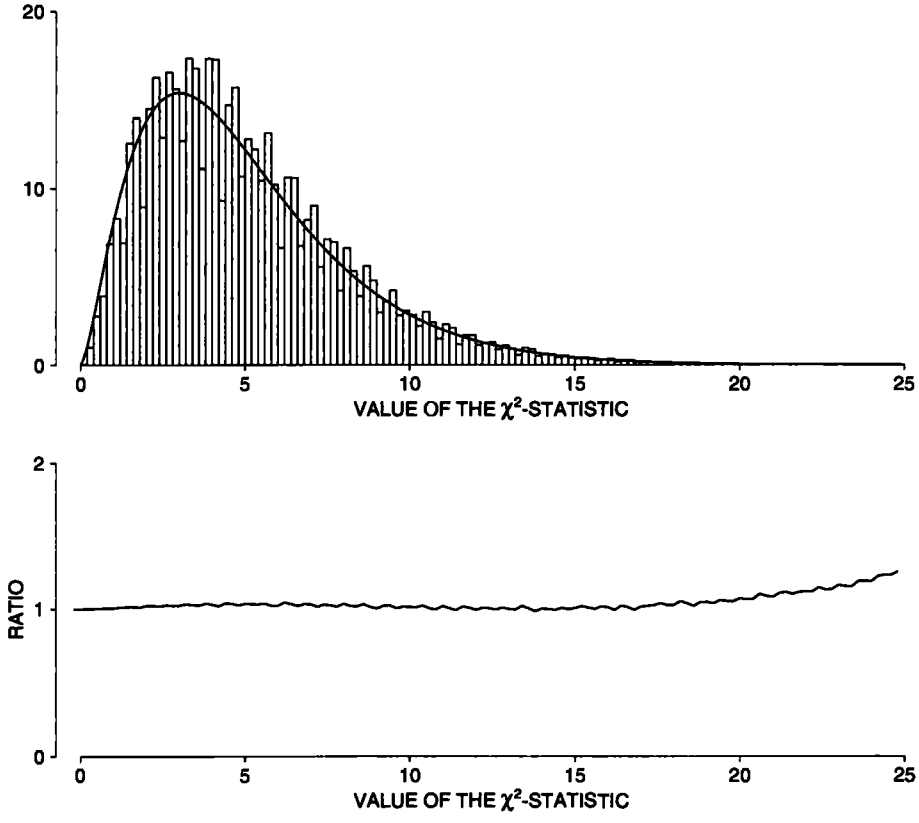
| *Degrees of freedom* | *90%* | *50%* | *10%* | *5%* | *1%* |
|---|---|---|---|---|---|
| 1 | 0.016 | 0.46 | 2.71 | 3.84 | 6.64 |
| 2 | 0.21 | 1.39 | 4.60 | 5.99 | 9.21 |
| 3 | 0.58 | 2.37 | 6.25 | 7.82 | 11.34 |
| 4 | 1.06 | 3.36 | 7.78 | 9.49 | 13.28 |
| 5 | 1.61 | 4.35 | 9.24 | 11.07 | 15.09 |
| 6 | 2.20 | 5.35 | 10.65 | 12.59 | 16.81 |
| 7 | 2.83 | 6.35 | 12.02 | 14.07 | 18.48 |
| 8 | 3.49 | 7.34 | 13.36 | 15.51 | 20.09 |
| 9 | 4.17 | 8.34 | 14.68 | 16.92 | 21.67 |
| 10 | 4.86 | 9.34 | 15.99 | 18.31 | 23.21 |

read from the table, but it is between 5% (the area to the right of 11.07) and 1% (the area to the right of 15.09). It is reasonable to guess that the area under the curve to the right of 14.2 is just a bit more than 1%.




Pearson developed the formulas for the $\chi^2$-statistic and the $\chi^2$-curves in tandem. His objective was to approximate the $P$-values without having to do a computation that was—by the standards of his time—quite formidable. How good is his approximation? Figure 2 shows the probability histogram for the $\chi^2$-statistic with 60 rolls of a fair die. A $\chi^2$-curve with 5 degrees of freedom is plotted too.

Figure 2. Pearson's approximation. The top panel shows the probability histogram for the $\chi^2$-statistic with 60 rolls of a fair die, compared with a $\chi^2$-curve (5 degrees of freedom). The bottom panel shows the ratio of tail areas. For example, take 14.2 on the horizontal axis. The area under the histogram to the right of 14.2 is 1.4382%. The area under the curve is 1.4388%. The ratio $1.4382/1.4388 \approx 0.9996$ is plotted above 14.2. Other ratios are plotted the same way.



The histogram is quite a bit bumpier than the curve, but follows it rather well. The area under the histogram to the right of any particular value is going to be close to the corresponding area under the curve. The ratio of these tail areas is graphed in the bottom panel.

In example 1, the area to the right of 14.2 under the histogram gives the exact value of $P$. This is 1.4382%. The area to the right of 14.2 under the curve gives Pearson's approximate value for $P$. This is 1.4388%. Not bad. When the number of rolls goes up, the approximation gets better, and the histogram gets less bumpy.[4]

As a rule of thumb, the approximation will be good when the expected frequency in each line of the table is 5 or more. In table 2, each expected frequency was 10, and the approximation was excellent. On the other hand, the approximation would not be so good for 100 draws from the box

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\quad 96\ \boxed{4}\text{'s}}$$

In this case, the expected number of $\boxed{1}$'s is only 1; similarly for $\boxed{2}$ and $\boxed{3}$. The expected numbers are too small for the approximation to be reliable.

When should the $\chi^2$-test be used, as opposed to the $z$-test? If it matters how many tickets of each kind are in the box, use the $\chi^2$-test. If it is only the average of the box that matters, use the $z$-test. For instance, suppose you are drawing with replacement from a box of tickets numbered 1 through 6; the percentages of the different kinds of tickets are unknown. To test the hypothesis that each value appears on $16\frac{2}{3}\%$ of the tickets, use the $\chi^2$-test. Basically, there is only one box which satisfies this hypothesis:

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

On the other hand, to test the hypothesis that the average of the box is 3.5, use the $z$-test. Of course, there are many boxes besides $\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$ where the average is 3.5: for instance,

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{3}\,\boxed{4}\,\boxed{4}\,\boxed{5}\,\boxed{6}} \quad \text{or} \quad \boxed{\boxed{1}\,\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}\,\boxed{6}}$$

To sum up:

- The $\chi^2$-test says whether the data are like the result of drawing at random from a box whose contents are given.
- The $z$-test says whether the data are like the result of drawing at random from a box whose average is given.[5]

The balance of this section tells how $\chi^2$ was used on a wheel of fortune.[6] Some winners in the California State Lottery are chosen to appear on a television game show called "The Big Spin." Each contestant spins a heavy cast aluminum wheel, with 100 slots numbered from 1 through 100. A hard rubber ball bounces around inside the wheel and then settles down into one slot or another, determining the prize given to the contestant.

Millions of dollars are at stake, so the wheel has to be tested quite carefully. The State Lottery Commission's statistical consultant Don Ylvisaker had the oper-

ators spin the wheel 800 times and count the number of times the ball landed in each slot. Then he made a $\chi^2$-test of the observed frequencies against the expected frequencies. The $\chi^2$-statistic turned out to be 119. There were $100 - 1 = 99$ degrees of freedom, and $P \approx 8\%$. This seemed marginal.

Slot number 69 came up most often and 19 least often. These two numbers were opposite each other. The wheel was then examined more carefully. A metal weight was found on the back, attached to the rim near slot number 69. Apparently, this had been done to balance the wheel, just as you would balance an automobile tire. The weight was removed, the wheel was rebalanced, and the tests were run again. The first 400 numbers did not look especially random, but things improved from there. As it turned out, the operators had oiled the wheel around spin 400 because it squeaked. The wheel was accepted and works well. (It is oiled regularly.)

"NO! YOU MAY NOT OIL THE WHEEL."

## 2. THE STRUCTURE OF THE $\chi^2$-TEST

Section 1 described the $\chi^2$-test. What are the ingredients?

*(a) The basic data.* This consists of some number of observations, usually denoted $N$. For the die, $N$ was 60 and the basic data were in table 1. For the wheel of fortune, $N$ was 800. The basic data were the 800 numbers generated on the trial spins.

*(b) The chance model.* Only one kind of chance model has been considered so far in this chapter. There is a box of tickets, whose contents are given. Draws are made at random with replacement from the box. According to the model, the data are like the draws. For the die, the box was

| 1 | 2 | 3 | 4 | 5 | 6 |

For the wheel of fortune, the box had 100 tickets, numbered from 1 through 100.

*(c) The frequency table.* For each value, the observed frequency is obtained from the basic data by counting.[7] The expected frequency is obtained from $N$ and the chance model. Table 2 reported the observed and expected frequencies for the die. A frequency table for the wheel would have 100 rows; it is omitted.

*(d) The $\chi^2$-statistic.* This is computed from the formula. For the die, the $\chi^2$-statistic was 14.2; for the wheel, the $\chi^2$-statistic was 119.

*(e) The degrees of freedom.* This is one less than the number of terms in the sum for $\chi^2$ (when the contents of the box are specified by the model). For the die, there were 5 degrees of freedom; for the wheel of fortune, there were 99. The degrees of freedom are computed from the model, not from the data.

*(f) The observed significance level.* This is approximated by the area to the right of the $\chi^2$-statistic, under the $\chi^2$-curve with the appropriate number of degrees of freedom. For the die, $P \approx 1.4\%$; for the wheel, $P \approx 8\%$.

The terminology is complicated because everything starts with "$\chi^2$."

- The $\chi^2$-test involves steps (a–f).
- The $\chi^2$-statistic is calculated from the data each time you make the test.
- Two $\chi^2$-curves are shown in figure 1.
- The $\chi^2$-table is based on the curves and is used to look up $P$-values.

Whatever is in the box, the same $\chi^2$-curves and tables can be used to approximate $P$, provided $N$ is large enough. That is what motivated the formula. With other test statistics, a new curve would be needed for every box.

## Exercisc Sct A

1. Find the area under the $\chi^2$-curve with 5 degrees of freedom to the right of
   (a) 1.61 (b) 9.24 (c) 15.09
2. Find the area to the right of 15.09 under the $\chi^2$-curve with 10 degrees of freedom.
3. Suppose the observed frequencies in table 2 had come out as shown in table 4A below. Compute the value of $\chi^2$, the degrees of freedom, and $P$. What can be inferred?

Table 4A

| *Value* | *Observed frequency* |
|---|---|
| 1 | 5 |
| 2 | 7 |
| 3 | 17 |
| 4 | 16 |
| 5 | 8 |
| 6 | 7 |

Table 4B

| *Value* | *Observed frequency* |
|---|---|
| 1 | 9 |
| 2 | 11 |
| 3 | 10 |
| 4 | 8 |
| 5 | 12 |
| 6 | 10 |

Table 4C

| *Value* | *Observed frequency* |
|---|---|
| 1 | 90 |
| 2 | 110 |
| 3 | 100 |
| 4 | 80 |
| 5 | 120 |
| 6 | 100 |

Table 4D

| *Value* | *Observed frequency* |
|---|---|
| 1 | 10,287 |
| 2 | 10,056 |
| 3 | 9,708 |
| 4 | 10,080 |
| 5 | 9,935 |
| 6 | 9,934 |

4. Suppose the observed frequencies in table 1 had come out as shown in table 4B. Make a $\chi^2$-test of the null hypothesis that the die is fair.

5. Suppose that table 1 had 600 entries instead of 60, with observed frequencies as shown in table 4C. Make a $\chi^2$-test of the null hypothesis that the die is fair.

6. Suppose that table 1 had 60,000 entries, with the observed frequencies as shown in table 4D.
   (a) Compute the percentage of times each value showed up.
   (b) Does the die look fair?
   (c) Make a $\chi^2$-test of the null hypothesis that the die is fair.

7. One study of grand juries in Alameda County, California, compared the demographic characteristics of jurors with the general population, to see if the jury panels were representative.[8] The results for age are shown below. The investigators wanted to know whether these 66 jurors were selected at random from the population of Alameda County. (Only persons 21 and over are considered; the county age distribution is known from Public Health Department data.)
   (a) True or false: to answer the investigators' question, you should make a $z$-test on each line in the table.
   (b) Fill in the blank: the _____-test combines information from all the lines in the table into an overall measure of the difference between the observed frequencies and expected frequencies. Options: $z$, $\chi^2$.
   (c) True or false: the right-hand column in the table gives the observed frequencies.
   (d) Fill in the blank: to make the $\chi^2$-test, you need to compute the _____ frequency in each age group. Options: expected, observed.
   (e) Now answer the investigators' question.

| *Age* | *County-wide percentage* | *Number of jurors* |
|---|---|---|
| 21 to 40 | 42 | 5 |
| 41 to 50 | 23 | 9 |
| 51 to 60 | 16 | 19 |
| 61 and over | 19 | 33 |
| Total | 100 | 66 |

8. Someone tells you to work exercise 7 as follows. (i) Convert each number to a percent: for instance, 5 out of 66 is 7.6%. (ii) Take the difference between the observed and expected percent. (iii) Square the difference. (iv) Divide by the expected percent. (v) Add up to get $\chi^2$. Is this right?

9. Another device tested by the California State Lottery has a set of 10 Ping-Pong balls, numbered from 0 through 9. These balls are mixed in a glass bowl by an air jet, and one is forced out at random. In the trial runs described below, the mixing machine seemed to be working well, but some of the ball sets may not have been behaving themselves. On each run, the machine made 120 draws from the bowl, with replacement.
   (a) Suppose everything is going as it should. In 120 draws from the bowl, each ball is expected to be drawn _____ times.

(b) The table below shows the results of testing 4 sets of balls. Sets A and D seemed marginal and were retested. Set B was rejected outright. Set C was accepted. How do these decisions follow from the data? (The table is read as follows: with ball set A, ball no. 0 was drawn 13 times; ball no. 1 was drawn 11 times; and so forth.)

(c) After retesting, what would you do with sets A and D? Explain briefly.

*FREQUENCIES*

| | *Ball set A* | | *Ball set B* | *Ball set C* | *Ball set D* | |
|---|---|---|---|---|---|---|
| *Ball no.* | *test* | *retest* | *test* | *test* | *test* | *retest* |
| 0 | 13 | 19 | 22 | 12 | 16 | 8 |
| 1 | 11 | 9 | 8 | 10 | 7 | 15 |
| 2 | 16 | 10 | 7 | 14 | 12 | 22 |
| 3 | 11 | 12 | 8 | 10 | 14 | 11 |
| 4 | 5 | 7 | 19 | 11 | 15 | 15 |
| 5 | 12 | 15 | 20 | 10 | 5 | 8 |
| 6 | 12 | 19 | 10 | 20 | 10 | 17 |
| 7 | 19 | 10 | 11 | 12 | 21 | 9 |
| 8 | 5 | 12 | 6 | 12 | 11 | 8 |
| 9 | 16 | 7 | 9 | 9 | 9 | 7 |

10. (a) A statistician wants to test the null hypothesis that his data are like 100 draws made at random with replacement from the box $\boxed{1\,2\,3\,4\,5\,6}$. The alternative hypothesis: the data are like 100 draws made at random with replacement from the box $\boxed{1\,1\,2\,3\,4\,5\,6\,6}$. Can the $\chi^2$-test do the job?

(b) As in (a), but the boxes are

$$\boxed{1\,2\,3\,4\,5\,6}$$

$$\boxed{1\,2\,3\,4\,5\,6\,1\,2\,3\,4\,5\,6}$$

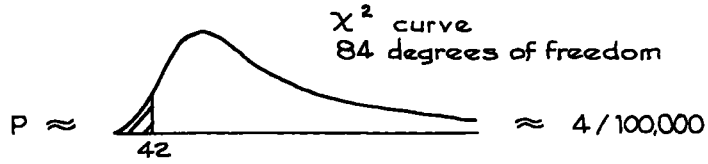
*The answers to these exercises are on pp. A99–100.*

## 3. HOW FISHER USED THE $\chi^2$-TEST

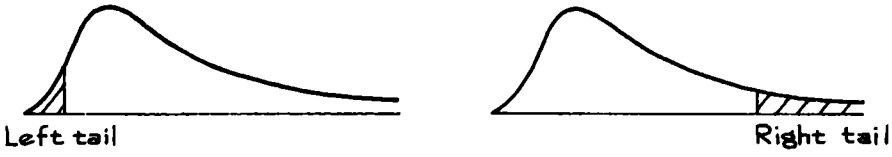
Fisher used the $\chi^2$-statistic to show that Mendel's data (chapter 25) were fudged.[9] For each of Mendel's experiments, Fisher computed the $\chi^2$-statistic. These experiments were all independent, for they involved different sets of plants. Fisher *pooled* the results.

> With independent experiments, the results can be pooled by adding up the separate $\chi^2$-statistics; the degrees of freedom add up too.

For instance, if one experiment gives $\chi^2 = 5.8$ with 5 degrees of freedom, and another independent experiment gives $\chi^2 = 3.1$ with 2 degrees of freedom, the two together have a pooled $\chi^2$ of $5.8 + 3.1 = 8.9$, with $5 + 2 = 7$ degrees of freedom. For Mendel's data, Fisher got a pooled $\chi^2$-value under 42, with 84 degrees of freedom. The area to the left of 42 under the $\chi^2$-curve with 84 degrees of freedom is about 4 in 100,000. The agreement between observed and expected is too good to be true.




At this point, a new principle seems to be involved: $P$ was computed as a left-hand tail area, not a right-hand one. Why?




Here is the reason. Fisher was not testing Mendel's chance model; he took that for granted. Instead, he was comparing two hypotheses—

- The null hypothesis: Mendel's data were gathered honestly.
- The alternative hypothesis: Mendel's data were fudged to make the reported frequencies closer to the expected ones.

Small values of $\chi^2$ say the observed frequencies are closer to the expected ones than chance variation would allow, and argue for the alternative hypothesis. Since it is small values of $\chi^2$ that argue against the null hypothesis, $P$ must be computed as a left-hand tail area. It is straightforward to set up the null hypothesis as a box model (chapter 25). The alternative hypothesis would be more complicated.

## Exercise Set B

1. Suppose the same die had been used to generate the data in tables 4A and 4C (p. 531), rolling it first 60 times for table 4A, and then 600 times for table 4C. Can you pool the results of the two tests? If so, how?

2. Suppose the same die had been used to generate the data in tables 4A and 4C (p. 531), rolling it 600 times in all. The first 60 rolls were used for table 4A; but table 4C reports the results on all 600 rolls. Can you pool the results of the two tests? If so, how?

3. One of Mendel's breeding trials came out as follows.[9] Make a $\chi^2$-test to see whether these data were fudged. Which way does the evidence point? Is it decisive?

| *Type of pea* | *Observed number* | *Expected number* |
|---|---|---|
| Smooth yellow | 315 | 313 |
| Wrinkled yellow | 101 | 104 |
| Smooth green | 108 | 104 |
| Wrinkled green | 32 | 35 |

*The answers to these exercises are on p. A100.*

## 4. TESTING INDEPENDENCE

The $\chi^2$-statistic is also used to test for independence, as will be explained in this section. The method will be indicated by example: Are handedness and sex independent? More precisely, take people age 25–34 in the U.S. The question is whether the distribution of "handedness" (right-handed, left-handed, ambidextrous) among the men in this population differs from the distribution among the women.

If data were available, showing for each man and woman in the population whether they were right-handed, left-handed, or ambidextrous, it would be possible to settle the issue directly, just by computing percentages. Such information is not available. However, HANES (p. 58) took a probability sample of 2,237 Americans 25–34. One of the things they determined for each sample person was handedness. Results are shown in table 5.

Table 5. Handedness by sex.

| | *Men* | *Women* |
|---|---|---|
| Right-handed | 934 | 1,070 |
| Left-handed | 113 | 92 |
| Ambidextrous | 20 | 8 |

This is a "3 × 2 table," because it has 3 rows and 2 columns. In general, when studying the relationship between two variables, of which one has $m$ values and the other has $n$ values, an $m \times n$ table is needed. In Table 5, it is hard to compare the distribution of handedness for men and women, because there are more women than men. Table 6 converts the data to percentages.

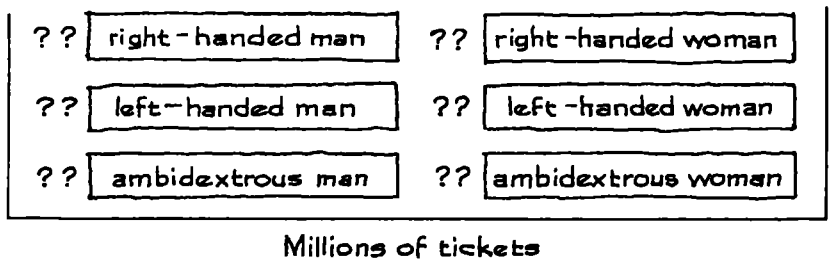
Table 6. Handedness by sex.

| | *Men* | *Women* |
|---|---|---|
| Right-handed | 87.5% | 91.5% |
| Left-handed | 10.6% | 7.9% |
| Ambidextrous | 1.9% | 0.7% |

As you can see, the distributions are different. The women are a bit likelier than men to be right-handed; they are less likely to be left-handed or ambidextrous. According to some neurophysiologists, right-handedness is associated with left-hemisphere dominance in the brain, the rational faculty ruling the emotional.[10] Does the sample show that women are more rational than men? Another interpretation: right-handedness is socially approved, left-handedness is socially deviant. Are women under greater pressure than men to follow the social norm for handedness?

A less dramatic interpretation: it's just chance. Even if handedness is distributed the same way for men and women in the population, the distributions could be different in the sample. Just by the luck of the draw, there could be too few right-handed men in the HANES sample, or too many right-handed women. To decide whether the observed difference is real or just due to chance, a statistical test is needed. That is where the $\chi^2$-test comes in.

The HANES sampling design is too complicated to analyze by means of the $\chi^2$-test. (This issue came up in sampling, where the formula for the SE depended on the design; pp. 388, 403, 424.) To illustrate technique, we are going to pretend that table 5 is based on a simple random sample, with 2,237 people chosen at random without replacement from the population. A box model for the data can be set up on that basis. There is one ticket in the box for each person in the population (Americans age 25–34). Each of these millions of tickets is marked in one of the following ways:




Our model says that the numbers in table 5 were generated by drawing 2,237 tickets at random without replacement from this huge box, and counting to see how many tickets there were for each of the 6 different types. The percentage composition of the box is unknown, so there are 6 parameters in the model.

Now we can formulate the null hypothesis and the alternative in terms of the box. The null hypothesis says that handedness and sex are independent. More explicitly, the percentage of right-handers among all men in the population equals the corresponding percentage among women; similarly for left-handers and the ambidextrous. On the null hypothesis, the differences in the sample percentages (table 6) just reflect chance variation. The alternative hypothesis is dependence—in the population, the distribution of handedness among the men differs from the distribution for women. On the alternative hypothesis, the differences in the sample reflect differences in the population.

To make a $\chi^2$-test of the null hypothesis, we have to compare the observed frequencies (table 5) with the expected frequencies. Getting these expected frequencies is a bit complicated, and the technique is explained later. The expected frequencies themselves are shown in table 7. They are based on the null hypothesis of independence.

Table 7. Observed and expected frequencies.

| | *Observed* | | *Expected* | |
|---|---|---|---|---|
| | *Men* | *Women* | *Men* | *Women* |
| Right-handed | 934 | 1,070 | 956 | 1,048 |
| Left-handed | 113 | 92 | 98 | 107 |
| Ambidextrous | 20 | 8 | 13 | 15 |

The next step is to compute

$$\begin{aligned}\chi^2 &= \text{sum of } \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} \\ &= \frac{(934-956)^2}{956} + \frac{(1{,}070-1{,}048)^2}{1{,}048} \\ &\quad + \frac{(113-98)^2}{98} + \frac{(92-107)^2}{107} \\ &\quad + \frac{(20-13)^2}{13} + \frac{(8-15)^2}{15} \\ &\approx 12\end{aligned}$$

How many degrees of freedom are there?

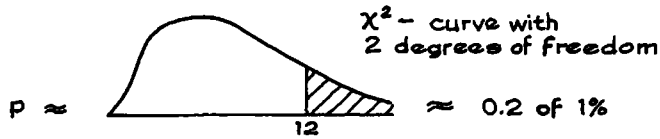
> When testing independence in an $m \times n$ table (with no other constraints on the probabilities), there are $(m-1) \times (n-1)$ degrees of freedom.

There are 6 terms in the sum for $\chi^2$, but there are only $(3-1) \times (2-1) = 2$ degrees of freedom. To see why, look at the differences.

$$\begin{array}{rr} -22 & 22 \\ 15 & -15 \\ 7 & -7 \end{array}$$

(The arithmetic for the first one: $934 - 956 = -22$, see table 7.) The differences add up to 0, horizontally and vertically. So, if you know the $-22$ and the 15, say, you can compute all the rest: only 2 of the differences are free to vary.

Now that we have the $\chi^2$-statistic and its degrees of freedom, $P$ can be worked out on the computer (or looked up in a table):




The observed significance level $P$ is the area to the right of 12 under the $\chi^2$-curve with 2 degrees of freedom, and this is about 0.2 of 1%. (The table will only tell you that the area is quite a bit less than 1%, which is good enough for present purposes.) The null hypothesis should be rejected. There is strong evidence to show that the distribution of handedness among the men in the population is different from the distribution for women. The observed difference in the sample seems to reflect a real difference in the population, rather than chance variation. That is what the $\chi^2$-test says. (A more careful analysis would have to take the design of the sample into account, but the conclusion stays the same.[11])

What is left is to compute the expected frequencies in table 7, and this will take some effort. To get started, you compute the row and column totals for table 5, as shown in table 8.

Table 8. Row and column totals.

| | *Men* | *Women* | *Total* |
|---|---|---|---|
| Right-handed | 934 | 1,070 | 2,004 |
| Left-handed | 113 | 92 | 205 |
| Ambidextrous | 20 | 8 | 28 |
| Total | 1,067 | 1,170 | 2,237 |

How do you get the 956 in table 7? From table 8, the percentage of right-handers in the sample is

$$\frac{2{,}004}{2{,}237} \times 100\% \approx 89.6\%$$

The number of men is 1,067. If handedness and sex are independent, the number of right-handed men in the sample should be

$$89.6\% \text{ of } 1{,}067 \approx 956.$$

The other expected frequencies in table 7 can be worked out the same way.

Expected values ought to be computed directly from the box model. In table 7, however, the "expected frequencies" are estimated from the sample—and the null hypothesis of independence. "Estimated expected frequencies" would be a better phrase, but "expected frequencies" is what statisticians say.[12]

## Exercise Set C

1. The percentage of women in the sample (table 8) is $1,170/2,237 \approx 52.3\%$. Someone wants to work out the expected number of ambidextrous women as 52.3% of 28. Is that OK?

2. (Hypothetical.) In a certain town, there are about one million eligible voters. A simple random sample of size 10,000 was chosen, to study the relationship between sex and participation in the last election. The results:

| | *Men* | *Women* |
|---|---|---|
| Voted | 2,792 | 3,591 |
| Didn't vote | 1,486 | 2,131 |

Make a $\chi^2$-test of the null hypothesis that sex and voting are independent.

*The next few exercises will help you learn which test to use when.*

3. The table below shows the distribution of marital status by sex for persons age 25–34 in Wyoming.[13]

Question: Are the distributions really different for men and women?

You may assume the data are from a simple random sample of 299 persons, of whom 143 were men and 156 were women. To answer the question, you use—

(i) the one-sample $z$-test.
(ii) the two-sample $z$-test.
(iii) the $\chi^2$-test, with a null hypothesis that tells you the contents of the box (section 1).
(iv) the $\chi^2$-test for independence (section 4).

Now answer the question. If the distributions are different, who are the women marrying?

| | *Men* | *Women* |
|---|---|---|
| Never married | 31.5% | 19.2% |
| Married | 60.1% | 67.3% |
| Widowed, divorced, separated | 8.4% | 13.5% |

4. Suppose all the numbers in exercise 3 had come from the Current Population Survey for March 2005, by extracting the data for people age 25–34 in Wyoming. Would that affect your answers? Explain briefly.

5. A study is made of incomes among full-time workers age 25–54 in a certain town. A simple random sample is taken, of 250 people with high school degrees: the sample average income is \$30,000 and the SD is \$25,000. Another simple random sample is taken, of 250 people with college degrees: the sample average income is \$50,000 and the SD is \$40,000.

Question: Is the difference in averages real, or due to chance?

To answer this question, you use—

(i) the one-sample $z$-test.
(ii) the two-sample $z$-test.
(iii) the $\chi^2$-test, with a null hypothesis that tells you the contents of the box (section 1).
(iv) the $\chi^2$-test for independence (section 4).

Now answer the question.

6. Demographers think that about 55% of newborns are male. In a certain hospital, 568 out of 1,000 consecutive births are male.

Question: Are the data consistent with the theory?

To answer this question, you use—

(i) the one-sample $z$-test.
(ii) the two-sample $z$-test.
(iii) the $\chi^2$-test, with a null hypothesis that tells you the contents of the box (section 1).
(iv) the $\chi^2$-test for independence (section 4).

Now answer the question.

7. To test whether a die is fair, someone rolls it 600 times. On each roll, he just records whether the result was even or odd, and large (4, 5, 6) or small (1, 2, 3). The observed frequencies turn out as follows:

| | *Large* | *Small* |
|---|---|---|
| Even | 183 | 113 |
| Odd | 88 | 216 |

Question: Is the die fair?

To answer this question, you use—

(i) the one-sample $z$-test.
(ii) the two-sample $z$-test.
(iii) the $\chi^2$-test, with a null hypothesis that tells you the contents of the box (section 1).
(iv) the $\chi^2$-test for independence (section 4).

Now answer the question.

*The answers to these exercises are on pp. A100–101.*

## 5. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. You are drawing 100 times at random with replacement from a box. Fill in the blanks, using the options below.

(a) To test the null hypothesis that the average of the box is 2, you would use _____.
(b) To test the null hypothesis that the box is $\boxed{1}\boxed{2}\boxed{3}$, you would use _____.

Options (some may not be used):

(i) the one-sample $z$-test.
(ii) the two-sample $z$-test.
(iii) the $\chi^2$-test, with a null hypothesis that tells you the contents of the box (section 1).
(iv) the $\chi^2$-test for independence (section 4).

2. As part of a study on the selection of grand juries in Alameda county, the educational level of grand jurors was compared with the county distribution:[14]

| *Educational level* | *County* | *Number of jurors* |
|---|---|---|
| Elementary | 28.4% | 1 |
| Secondary | 48.5% | 10 |
| Some college | 11.9% | 16 |
| College degree | 11.2% | 35 |
| Total | 100.0% | 62 |

Could a simple random sample of 62 people from the county show a distribution of educational level so different from the county-wide one? Choose one option and explain.

(i) This is absolutely impossible.
(ii) This is possible, but fantastically unlikely.
(iii) This is possible but unlikely—the chance is around 1% or so.
(iv) This is quite possible—the chance is around 10% or so.
(v) This is nearly certain.

3. Each respondent in the Current Population Survey of March 2005 was classified as employed, unemployed, or outside the labor force. The results for men in California age 35–44 can be cross-tabulated by marital status, as follows:[15]

| | *Married* | *Widowed, divorced, or separated* | *Never married* |
|---|---|---|---|
| Employed | 790 | 98 | 209 |
| Unemployed | 56 | 11 | 27 |
| Not in labor force | 21 | 7 | 13 |

Men of different marital status seem to have different distributions of labor force status. Or is this just chance variation? (You may assume the data come from a simple random sample.)

4. (a) Does the histogram in figure 2 represent data, or chance?
(b) There is a block over the interval from 5 to 5.2. What does the area of this block represent? (Ranges include the left endpoint, but not the right.)
(c) Which chance is larger for 60 throws of a die? Or can this be determined from figure 2?
   (i) The chance that the $\chi^2$-statistic is in the range from 4.8 to 5.0.
   (ii) The chance that the $\chi^2$-statistic is in the range from 5.0 to 5.2.

(d) If $\chi^2 = 10$, then $P$ is about

1% 10% 25% 50% cannot be determined from the figure

5. An investigator makes a $\chi^2$-test, to see whether the observed frequencies are too far from the expected frequencies.

   (a) If $\chi^2 = 15$, the $P$-value will be bigger with _____ degrees of freedom than with _____ degrees of freedom. Options: 5, 10.

   (b) If there are 10 degrees of freedom, the $P$-value will be bigger with $\chi^2 =$ _____ than with $\chi^2 =$ _____. Options: 15, 20.

   No calculations are needed, just look at figure 1.

6. Someone claims to be rolling a pair of fair dice. To test his claim, you make him roll the dice 360 times, and you count up the number of times each sum appears. The results are shown below. (For your convenience, the chance of throwing each sum with a pair of fair dice is shown too.) Should you play craps with this individual? Or are the observed frequencies too close to the expected frequencies?

| *Sum* | *Chance* | *Frequency* |
|---|---|---|
| 2 | 1/36 | 11 |
| 3 | 2/36 | 18 |
| 4 | 3/36 | 33 |
| 5 | 4/36 | 41 |
| 6 | 5/36 | 47 |
| 7 | 6/36 | 61 |
| 8 | 5/36 | 52 |
| 9 | 4/36 | 43 |
| 10 | 3/36 | 29 |
| 11 | 2/36 | 17 |
| 12 | 1/36 | 8 |

7. The International Rice Research Institute in the Philippines develops new lines of rice which combine high yields with resistance to disease and insects. The technique involves crossing different lines to get a new line which has the most advantageous combination of genes. Detailed genetic modeling is required. One project involved breeding new lines to resist the "brown plant hopper" (an insect): 374 lines were raised, with the results shown below.[16]

| | *Number of lines* |
|---|---|
| All plants resistant | 97 |
| Mixed: some plants resistant, some susceptible | 184 |
| All plants susceptible | 93 |

According to the IRRI model, the lines are independent: each line has a 25% chance to be resistant, a 50% chance to be mixed, and a 25% chance to be susceptible. Are the facts consistent with this model?

8. Two people are trying to decide whether a die is fair. They roll it 100 times, with the results shown at the top of the next page. One person wants to make

a $z$-test, the other wants to make a $\chi^2$-test. Who is right? Explain briefly.

21 ⚀'s 15 ⚁'s 13 ⚂'s 17 ⚃'s 19 ⚄'s 15 ⚅'s

Average of numbers rolled $\approx 3.43$, SD $\approx 1.76$

9. Each respondent in the Current Population Survey of March 2005 can be classified by age and marital status. The table below shows results for women age 20–29 in Montana.

   Question A. Women of different ages seem to have different distributions of marital status. Or is this just chance variation?

   Question B. If the difference is real, what accounts for it?

   (a) Can you answer these questions with the information given? If so, answer them. If not, why not?
   (b) Can you answer these questions if the data in the table resulted from a simple random sample of women age 20–29 in Montana? If so, answer them. If not, why not?

| | *AGE* | |
|---|---|---|
| | *20–24* | *25–29* |
| Never married | 46 | 21 |
| Married | 17 | 32 |
| Widowed, divorced, separated | 1 | 6 |

10. The U.S. has bilateral extradition treaties with many countries. (A person charged with a crime in his home country may escape to the U.S.; if he is captured in the U.S., authorities in his home country may request that he be "extradited," that is, turned over for prosecution under their laws.)

    The Senate attached a special rider to the treaty governing extradition to Northern Ireland: fugitives cannot be returned if they will be discriminated against on the basis of religion. In a leading case, the defense tried to establish discrimination in Northern Ireland's criminal justice system.

    One argument was based on 1991 acquittal rates for persons charged with terrorist offenses.[17] According to a defense expert, these rates were significantly different for Protestants and Catholics: $\chi^2 \approx 6.2$ on 1 degree of freedom, $P \approx 1\%$. The data are shown below: 8 Protestants out of 15 were acquitted, compared to 27 Catholics out of 65.

    (a) Is the calculation of $\chi^2$ correct? If not, can you guess what the mistake was? (That might be quite difficult.)
    (b) What box model did the defense have in mind? Comment briefly on the model.

| | *Protestant* | *Catholic* |
|---|---|---|
| Acquitted | 8 | 27 |
| Convicted | 7 | 38 |

## 6. SUMMARY

1. The $\chi^2$-statistic can be used to test the hypothesis that data were generated according to a particular chance model.

2. $$\chi^2 = \text{sum of } \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

3. When the model is fully specified (no parameters to estimate from the data),

$$\text{degrees of freedom} = \text{number of terms} - \text{one}.$$

4. The observed significance level $P$ can be approximated as the area under the $\chi^2$-curve to the right of the observed value for $\chi^2$. The significance level gives the chance of the model producing observed frequencies as far from the expected frequencies as those at hand, or even further, distance being measured by $\chi^2$.

5. Sometimes the model can be taken as true, and the problem is to decide whether the data have been fudged to make the observed frequencies closer to the expected ones. Then $P$ would be computed as the area to the left of the observed value for $\chi^2$.

6. If experiments are independent, the $\chi^2$-statistics can be pooled by addition. The degrees of freedom are just added too.

7. The $\chi^2$-statistic can also be used to test for independence. This is legitimate when the data have been obtained from a simple random sample, and an inference about the population is wanted. With an $m \times n$ table (and no extra constraints on the probabilities) there are $(m - 1) \times (n - 1)$ degrees of freedom.

# 29

# A Closer Look at Tests of Significance

*One of the misfortunes of the law [is that] ideas become encysted in phrases and thereafter for a long time cease to provoke further analysis.*

—OLIVER WENDELL HOLMES, JR. (UNITED STATES, 1841–1935)[1]

## 1. WAS THE RESULT SIGNIFICANT?

How small does $P$ have to get before you reject the null hypothesis? As reported in section 4 of chapter 26, many investigators draw lines at 5% and 1%. If $P$ is less than 5%, the result is "statistically significant," and the "null hypothesis is rejected at the 5% level." If P is less than 1%, the result is "highly significant." However, the question is almost like asking how cold it has to get before you are entitled to say, "It's cold." A temperature of 70°F is balmy, −20°F is cold indeed, and there is no sharp dividing line. Logically, it is the same with testing. There is no sharp dividing line between probable and improbable results.

A $P$-value of 5.1% means just about the same thing as 4.9%. However, these two $P$-values can be treated quite differently, because many journals will only publish results which are "statistically significant"—the 5% line. Some of the more prestigious journals will only publish results which are "highly significant"—the 1% line.[2] These arbitrary lines are taken so seriously that many

investigators only report their results as "statistically significant" or "highly significant." They don't even bother telling you the value of $P$, let alone what test they used.

> Investigators should summarize the data, say what test was used, and report the $P$-value instead of just comparing $P$ to 5% or 1%.

*Historical note.* Where do the 5% and 1% lines come from? To find out, we have to look at the way statistical tables are laid out. The $t$-table is a good example (section 6 of chapter 26). Part of it is reproduced below as table 1.

Table 1. A short $t$-table.

| *Degrees of freedom* | *10%* | *5%* | *1%* |
|---|---|---|---|
| 1 | 3.08 | 6.31 | 31.82 |
| 2 | 1.89 | 2.92 | 6.96 |
| 3 | 1.64 | 2.35 | 4.54 |
| 4 | 1.53 | 2.13 | 3.75 |
| 5 | 1.48 | 2.02 | 3.36 |

How is this table used in testing? Suppose investigators are making a $t$-test with 3 degrees of freedom. They are using the 5% line, and want to know how big the $t$-statistic has to be in order to achieve "statistical significance"—a $P$-value below 5%. The table is laid out to make this easy. They look across the row for 3 degrees of freedom and down the column for 5%, finding the entry 2.35 in the body of the table. The area to the right of 2.35 under the curve for 3 degrees of freedom is 5%. So the result is "statistically significant" as soon as $t$ is more than 2.35. In other words, the table gives the cutoff for "statistical significance." Similarly, it gives the cutoff for the 1% line, or for any other significance level listed across the top.

R. A. Fisher was one of the first to publish such tables, and it seems to have been his idea to lay them out that way. There is a limited amount of room on the page. Once the number of levels was limited, 5% and 1% stood out as nice round numbers, and they soon acquired a magical life of their own. With computers everywhere, this kind of table is almost obsolete. So are the 5% and 1% levels.[3]

## Exercise Set A

1. True or false, and explain:
   (a) If $P = 1.1\%$, the result is "significant" but not "highly significant."
   (b) If $P = 0.9$ of 1%, the result is "highly significant."
2. True or false, and explain:
   (a) The $P$-value of a test is the chance that the null hypothesis is true.

(b) If a result is statistically significant, there are only 5 chances in 100 for it to be due to chance, and 95 chances in 100 for it to be real.

*The answers to these exercises are on p. A101.*

## 2. DATA SNOOPING

The point of testing is to help distinguish between real differences and chance variation. People sometimes jump to the conclusion that a result which is statistically significant cannot be explained as chance variation. This is false. Once in a while, the average of the draws will be 2 SEs above the average of the box, just by chance. More specifically, even if the null hypothesis is right, there is a 5% chance of getting a difference which the test will call "statistically significant." This 5% chance could happen to you—an unlikely event, but not impossible. Similarly, on the null hypothesis, there is 1% chance to get a difference which is highly significant but just a fluke.

Put another way, an investigator who makes 100 tests can expect to get five results which are "statistically significant" and one which is "highly significant" even if the null hypothesis is right in every case—so that each difference is just due to chance. (See exercise 5 on p. 483.) You cannot determine, for sure, whether a difference is real or just coincidence.

To make bad enough worse, investigators often decide which hypotheses to test only after they have seen the data. Statisticians call this *data snooping*. Investigators really ought to say how many tests they ran before statistically significant differences turned up. And to cut down the chance of being fooled by "statistically significant" flukes, they ought to test their conclusions on an independent batch of data—for instance by replicating the experiment.[4] This good advice is seldom followed.

| Data-snooping makes $P$-values hard to interpret. |
|---|

*Example 1. Clusters.* Liver cancer is a rare disease, often thought to be caused by environmental agents. The chance of having 2 or more cases in a given year in a town with 10,000 inhabitants is small—perhaps 1/2 of 1%. A cluster of liver cancer cases (several cases in a small community) prompts a search for causes, like the contamination of the water supply by synthetic chemicals.[5]

*Discussion.* With (say) 100 towns of this size and a 10-year time period, it is likely that several clusters will turn up, just by chance. There are $100 \times 10 = 1{,}000$ combinations of towns and years; and $0.005 \times 1{,}000 = 5$. If you keep on testing null hypotheses, sooner or later you will get significant differences.

One form of data snooping is looking to see whether your sample average is too big or too small—before you make the statistical test. To guard against this kind of data snooping, many statisticians recommend using *two-tailed* rather than

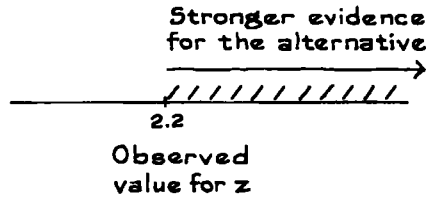
*one-tailed* tests. The point is easiest to see in a hypothetical example. Someone wants to test whether a coin is fair: does it land heads with probability 50%? The coin is tossed 100 times, and it lands heads on 61 of the tosses. If the coin is fair, the expected number of heads is 50, so the difference between 61 and 50 just represents chance variation. To test this null hypothesis, a box model is needed. The model consists of 100 draws from the box

$$\boxed{?? \ \boxed{0}\text{'s} \quad ?? \ \boxed{1}\text{'s}} \qquad 0 = \text{tails}, \quad 1 = \text{heads}.$$

The fraction of 1's in this box is an unknown parameter, representing the probability of heads. The null hypothesis says that the fraction of 1's in the box is 1/2. The test statistic is

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{61 - 50}{5} = 2.2$$

One investigator might formulate the alternative hypothesis that the coin is biased toward heads: in other words, that the fraction of 1's in the box is bigger than 1/2. On this basis, large positive values of $z$ favor the alternative hypothesis, but negative values of $z$ do not. Therefore, values of $z$ bigger than 2.2 favor the alternative hypothesis even more than the observed value does.



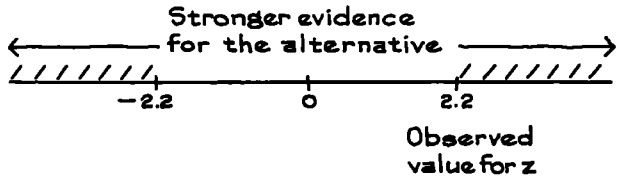

So $P$ is figured as the area to the right of 2.2 under the normal curve:

Another investigator might formulate a different alternative hypothesis: that the probability of heads differs from 50%, in either direction. In other words, the fraction of 1's in the box differs from 1/2, and may be bigger or smaller. On this basis, large positive values of $z$ favor the alternative, but so do large negative values. If the number of heads is 2.2 SEs above the expected value of 50, that is bad for the null hypothesis. And if the number of heads is 2.2 SEs below the expected value, that is just as bad. The $z$-values more extreme than the observed 2.2 are:

- 2.2 or more

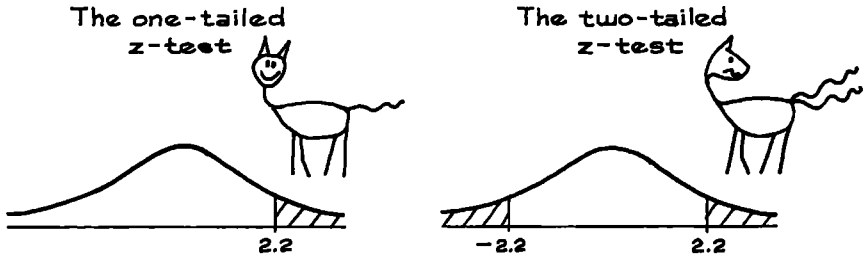
or

- −2.2 or less.

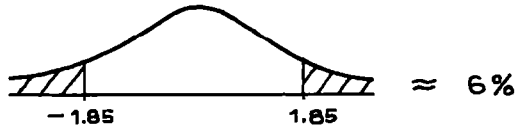



Now $P$ is figured differently:

The first way of figuring $P$ is the *one-tailed z-test*; the second is *two-tailed*. Which should be used? That depends on the precise form of the alternative hypothesis. It is a matter of seeing which $z$-values argue more strongly for the alternative hypothesis than the one computed from the data. The one-tailed test is appropriate when the alternative hypothesis says that the average of the box is bigger than a given value. The two-tailed test is appropriate when the alternative hypothesis says that the average of the box differs from the given value—bigger or smaller.



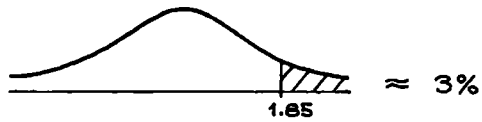

In principle, it doesn't matter very much whether investigators make one-tailed or two-tailed tests, as long as they say what they did. For instance, if they made a one-tailed test, and you think it should have been two-tailed, just double the $P$-value.[6] To see why such a fuss is made over this issue, suppose a group of investigators makes a two-tailed $z$-test. They get $z = 1.85$, so $P \approx 6\%$.




Naturally, they want to publish. But as it stands, most journals won't touch the report—the result is not "statistically significant."

What can they do? They could refine the experimental technique, gather more data, use sharper analytical methods. This is hard. The other possibility is

simpler: do a one-tailed test. It is the arbitrary lines at 5% and 1% which make the distinction between two-tailed and one-tailed tests loom so large.

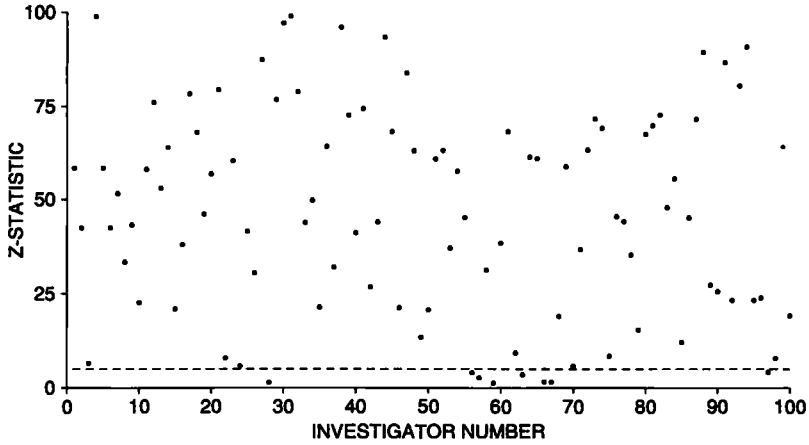



*Example 2. Cholesterol.* A randomized controlled double-blind experiment was performed to demonstrate the efficacy of a drug called "cholestyramine" in reducing blood cholesterol levels and preventing heart attacks. There were 3,806 subjects, who were all middle-aged men at high risk of heart attack; 1,906 were chosen at random for the treatment group and the remaining 1,900 were assigned to the control group. The subjects were followed for 7 years. The drug did reduce the cholesterol level in the treatment group (by about 8%). Furthermore, there were 155 heart attacks in the treatment group, and 187 in the control group: 8.1% versus 9.8%, $z \approx -1.8$, $P \approx 3.5\%$ (one-tailed). This was called "strong evidence" that cholesterol helps cause heart attacks.[7]

*Discussion.* With a two-tailed test, $P \approx 7\%$ and the difference is not significant. (The article was published in the *Journal of the American Medical Association*, whose editors are quite strict about the 5% line.) The investigators are overstating their results, and the emphasis on "statistical significance" encourages them to do so.

## Exercise Set B

1. One hundred investigators each set out to test a different null hypothesis. Unknown to them, all the null hypotheses happen to be true. Investigator #1 gets a $P$-value of 58%, plotted in the graph below as the point (1, 58). Investigator #2 gets a $P$-value of 42%, plotted as (2, 42). And so forth. The 5%-line is shown.
   (a) How many investigators should get a statistically significant result?
   (b) How many do?
   (c) How many should get a result which is highly significant?



2. In "Ganzfeld" experiments on ESP, there are two subjects, a sender and a receiver, located in separate rooms.[8] There is a standard set of patterns, arranged in 25 sets of 4. The experimenter goes through the 25 sets in order. From each set, one pattern is chosen at random, and shown to the sender (but not to the receiver). The sender tries to convey a mental image of the pattern to the receiver. The receiver is shown the 4 patterns, and ranks them from 1 = most likely to 4 = least likely. After going through all 25 sets of patterns, the experimenter makes a statistical analysis to see if the receiver did better than the chance level. Three test statistics are used.

   - The number of "hits." A receiver scores a hit by assigning rank 1 to the pattern that was in fact chosen. The number of hits ranges from 0 to 25. (If the number of hits is large, that is evidence for ESP.)
   - The number of "high ranks." A receiver scores a high rank by assigning rank 1 or rank 2 to the pattern that was chosen. The number of high ranks ranges from 0 to 25. (If the number of high ranks is large, that is evidence for ESP.)
   - The sum of the ranks assigned to the 25 chosen patterns. This sum ranges from 25 to 100. (If the sum is small, that is evidence for ESP.)

   Suppose there is no ESP, no cheating, and the choice of patterns is totally random.

   (a) The number of hits is like the sum of _____ draws from $\boxed{1\ 0\ 0\ 0}$. Fill in the blank and explain.

   (b) The number of high ranks is like the sum of 25 draws from the box _____. Fill in the blank, and explain.

   (c) Make a box model for the sum of the ranks.

For use in exercise 3, you are given the following information. Suppose 25 tickets are drawn at random with replacement from the box $\boxed{1\ 2\ 3\ 4}$.

- There is about a 3% chance of getting 11 or more tickets marked 1.
- There is about a 5% chance of getting 17 or more tickets marked 1 or 2.
- There is about a 5% chance that the sum of the draws will be 53 or less.

3. (This continues exercise 2.) Suppose there is no ESP, no cheating, and the choice of patterns is totally random.

   (a) One hundred investigators do Ganzfeld experiments. They will publish "significant" evidence for ESP if the number of hits is 11 or more. About how many of them will get significant evidence?

   (b) Repeat, if the definition of significant evidence is changed to "the number of high ranks is 17 or more."

   (c) Repeat, if the definition of significant evidence is changed to "the sum of the ranks is 53 or less."

4. (This continues exercises 2 and 3.) Suppose there is no ESP, no cheating, and the choice of patterns is totally random. One hundred investigators do Ganzfeld experiments. They will decide on a statistical test after seeing the data.

   - If the number of hits is 11 or more, they will base the test on the number of hits.
   - If not, but the number of high ranks is 17 or more, they will base the test on the number of high ranks.
   - If not, but the sum of the ranks is 53 or less, they will base the test on the sum of the ranks.

The number of these investigators who get "significant" evidence of ESP will be _____ 5. Fill in the blank, using one of the options below, and explain briefly.

just about  somewhat more than  somewhat less than

5. New chemicals are screened to see if they cause cancer in lab mice. A "bioassay" can be done with 500 mice: 250 are chosen at random and given the test chemical in their food, the other 250 get a normal lab diet. After 33 months, cancer rates in the two groups are compared, using the two-sample $z$-test.[9]

   Investigators look at cancer rates in about 25 organs and organ systems—lungs, liver, circulatory system, etc. With one chemical, $z \approx -1.8$ for the lungs, $z \approx 2.4$ for the liver, $z \approx -2.1$ for leukemia, and there are another 22 values of $z$ that range from $-1.6$ to $+1.5$. The investigators conclude that the chemical causes liver cancer ($z \approx 2.4$, $P \approx 1\%$, one-tailed). Comment briefly.

6. One hundred draws are made at random from box X. The average of the draws is 51.8, and their SD is 9. The null hypothesis says that the average of the box equals 50, while the alternative hypothesis says that the average of the box differs from 50. Is a one-tailed or a two-tailed $z$-test more appropriate?

7. One hundred draws are made at random from box Y. The average of the draws is 51.8, and their SD is 9. The null hypothesis says that the average of the box equals 50, while the alternative hypothesis says that the average of the box is bigger than 50. Is a one-tailed or a two-tailed $z$-test more appropriate?

8. An investigator has independent samples from box A and from box B. Her null hypothesis says that the two boxes have the same average. She looks at the difference

   average of sample from A − average of sample from B.

   The two-sample $z$-test gives $z \approx 1.79$. Is the difference statistically significant—

   (a) if the alternative hypothesis says that the average of box A is bigger than the average of box B?
   (b) if the alternative hypothesis says that the average of box A is smaller than the average of box B?
   (c) if the alternative hypothesis says that the average of box A is different from the average of box B?

9. (Hard.) Transfusion of contaminated blood creates a risk of infection. (AIDS is a case in point.) A physician must balance the gain from the transfusion against the risk, and accurate data are important. In a survey of the published medical literature on serum hepatitis resulting from transfusions, Chalmers and associates found that the larger studies had lower fatality rates.[10] How can this be explained?

*The answers to these exercises are on pp. A101–102.*

## 3. WAS THE RESULT IMPORTANT?

If a difference is statistically significant, then it is hard to explain away as a chance variation. But in this technical phrase, "significant" does not mean "important." Statistical significance and practical significance are two different ideas.[11]

The point is easiest to understand in the context of a hypothetical example (based on exercise 7, pp. 507–508). Suppose that investigators want to compare WISC vocabulary scores for big-city and rural children, age 6 to 9. They take a simple random sample of 2,500 big-city children, and an independent simple random sample of 2,500 rural children. The big-city children average 26 on the test, and their SD is 10 points. The rural children only average 25, with the same SD of 10 points. What does this one-point difference mean? To find out, the investigators make a two-sample $z$-test. The SE for the difference can be estimated as 0.3, so

$$z \approx 1/0.3 \approx 3.3, \qquad P \approx 5/10{,}000.$$

The difference between big-city children and rural children is highly significant, rural children are lagging behind in the development of language skills, and the investigators launch a crusade to pour money into rural schools.

The commonsense reaction must be, slow down. The $z$-test is only telling us that the one-point difference between the sample averages is almost impossible to explain as a chance variation. To focus the issue, suppose that the samples are a perfect image of the population, so that all the big-city children in the U.S. (not just the ones in the sample) would average 26 points on the WISC vocabulary scale, while the average for all the rural children in the U.S. would be 25 points. Then what? There is no more chance variation to worry about, so a test of significance cannot help. All the facts are in, and the problem is to find out what the difference means.

To do that, it is necessary to look at the WISC vocabulary scale itself. There are forty words which the child has to define. Two points are given for a correct definition, and one point for a partially correct definition. So the one-point difference between big-city and rural children only amounts to a partial understanding of one word out of forty. This is not a solid basis for a crusade. Quite the opposite: the investigators have proved there is almost no difference between big-city and rural children on the WISC vocabulary scale.[12]

Of course, the sample does not reflect the population perfectly, so a standard error should be attached to the estimate for the difference. Based on the two samples of 2,500 children, the difference in average scores between all the big-city and rural children in the U.S. would be estimated as 1 point, give or take 0.3 points or so. The $z$-statistic is impressive because 1 is a lot, relative to 0.3.

A big sample is good because it enables the investigator to measure a difference quite accurately—with a small SE. But the $z$-test compares the difference to the SE. Therefore, with a large sample even a small difference can lead to an impressive value for $z$. The $z$-test can be too sensitive for its own good.

> The $P$-value of a test depends on the sample size. With a large sample, even a small difference can be "statistically significant," that is, hard to explain by the luck of the draw. This doesn't necessarily make it important. Conversely, an important difference may not be statistically significant if the sample is too small.

*Example 3.* As reported in section 2 of chapter 27, reading test scores declined from 290 in 1990 to 285 in 2004. These were averages based on nationwide samples; $z \approx -2.8$ and $P \approx 1/4$ of 1% (one-tailed). Does the increase matter?

*Solution.* The $P$-value says the increase is hard to explain away as chance error. The $P$-value does not say whether the increase matters. More detailed analysis of the data suggests that each extra year of schooling is associated with about a 6-point increase in average test scores. On this basis, a 5-point decline is worrisome. Other measures of school performance, however, are more reassuring.[13]

## Exercise Set C

1. True or false, and explain:
   (a) A difference which is highly significant must be very important.
   (b) Big samples are bad because small differences will look significant.

2. A large university wants to compare the performance of male and female undergraduates on a standardized reading test, but can only afford to do this on a sample basis. An investigator chooses 100 male undergraduates at random, and independently 100 females. The men average 49 on the test, and their SD is 10 points. The women average 51 on the test, with the same SD of 10 points. Is the difference in the average scores real, or a chance variation? Or does the question make sense?

3. Repeat exercise 2, keeping the averages and SDs the same, but increasing the sample sizes from 100 to 400.

4. Someone explains the point of a test of significance as follows.[14] "If the null hypothesis is rejected, the difference isn't trivial. It is bigger than what would occur just by chance." Comment briefly.

5. Other things being equal, which is stronger evidence for the null hypothesis: $P = 3\%$ or $P = 27\%$?

6. Before publication in a scholarly journal, papers are reviewed. Is this process fair? To find out, a psychologist makes up two versions of a paper.[15] Both versions describe a study on the effect of rewarding children for classroom performance. The versions are identical, except for the data. One data set shows that rewards help motivate learning; the other, that rewards don't help. Some reviewers were chosen at random to get each version. All the reviewers were associated with a journal whose position was "behaviorist:" rewards for learning should work. As it turned out, both versions of the paper contained a minor inconsistency in the description of the study. The investigator did a two-sample $z$-test, concluding that—

   > Of the individuals who got the positive version, only 25% found the mistake. Of those who got the negative version, 71.5% found the mistake. By the two-sample $z$-test, this difference must be considered substantial, $P \approx 2\%$, one-tailed.

   (a) Why is the two-sample $z$-test legitimate? Or is it?
   (b) The standard error for the difference was about _____ percentage points.
   (c) Is the difference between 71.5% and 25% substantial? Answer yes or no, and discuss briefly.

(d) What do the results of the $z$-test add to the argument?
(e) What do the data say about the fairness of the review process?

7. An economist makes a study of how CALTRANS chose freeway routes in San Francisco and Los Angeles.[16] He develops a chance model in order to assess the effect of different variables on the decisions. "External political and public variables" include the views of other state agencies, school boards, businesses, large property owners, and property owners' associations. To find out how much influence these variables have on the freeway decisions, the economist makes a test of significance. The null hypothesis says that the external political and public variables make no difference in the decisions. The observed significance level is about 3%.

   Since the result is statistically significant but not highly significant, the economist concludes "these factors do influence the freeway decisions, but their impact is relatively weak." Does the conclusion follow from the statistical test?

8. An economist estimates the price elasticity for refined oil products as $-6$. (An elasticity of $-6$ means that a 1% increase in prices leads to a 6% drop in sales; an elasticity of $-4.3$ means that a 1% increase in prices leads to a 4.3% drop in sales, and so forth.)

   The standard error on the estimated elasticity is 2.5. The economist tests the null hypothesis that the elasticity is 0, and gets $z = -6/2.5 = -2.4$, so $P \approx 1\%$ (one-tailed). The conclusion: he is "99% confident of the estimate."[17]

   The economist seems to have assumed something like the Gauss model—

   $$\text{estimated elasticity} = \text{true elasticity} + \text{error}.$$

   The error has an expected value of 0, an SE of 2.5, and follows the normal curve. You can use these assumptions, but comment on them when you answer part (d).

   (a) Find a 99% confidence interval for the "true" elasticity.
   (b) What does the $P$-value of 1% mean?
   (c) Was the economist using statistical tests in an appropriate way?
   (d) What do you think of the "Gauss model for elasticity?"

*The answers to these exercises are on pp. A102–103.*

## 4. THE ROLE OF THE MODEL

To review briefly, a test of significance answers the question, "Is the difference due to chance?" But the test can't do its job until the word "chance" has been given a precise definition. That is where the box model comes in.[18]

| To make sense out of a test of significance, a box model is needed. |
|---|

This idea may be a little surprising, because the arithmetic of the test does not use the box model. Instead, the test seems to generate the chances directly from the data. That is an illusion. It is the box model which defines the chances. The formulas for the expected values and standard errors make a tacit assumption:

that the data are like draws from a box. So do the statistical tables—normal, $t$, and $\chi^2$. If the box model is wrong, the formulas and the tables do not apply, and may give silly results. This section discusses some examples.

*Example 4.* Census data show that in 1980, there were 227 million people in the U.S., of whom 11.3% were 65 or older. In 2000, there were 281 million people, of whom 12.3% were 65 or older.[19] Is the difference in the percentages statistically significant?

*Discussion.* The arithmetic of a two-sample $z$-test is easy enough to do, but the result is close to meaningless. We have Census data on the whole population. There is no sampling variability to worry about. Census data are subject to many small errors, but these are not like draws from a box. The aging of the population is real. It makes a difference to the health care and social security systems. However, the concept of statistical significance does not apply. The $P$-value would not help us to interpret the data.

> If a test of significance is based on data for the whole population, watch out.

*Example 5.* The Graduate Division at the University of California, Berkeley compares admission rates for men and women. For one year and one graduate major, this came out as follows: 825 men applied, and 61.7% were admitted; 108 women applied, and 82.4% were admitted. Is the difference between admission rates for men and women statistically significant?

*Discussion.* Again, there is nothing to stop you from doing a two-sample $z$-test. However, to make sense out of the results, a box model would be needed, and there doesn't seem to be one in the neighborhood. It is almost impossible to identify the pool of potential applicants. Even if you could, the actual applicants were not drawn from this pool by any probability method. Nor do departments admit candidates by drawing names from a hat (although that might not be such a bad idea). The concept of statistical significance does not apply.

Statisticians distinguish between samples drawn by probability methods and samples of convenience (section 4 of chapter 23). A sample of convenience consists of whoever is handy—students in a psychology class, the first hundred people you bump into, or all the applicants to a given department in a given year. With a sample of convenience, the concept of chance becomes quite slippery, the phrase "the difference is due to chance" is hard to interpret, and so are $P$-values. Example 5 was based on a sample of convenience.[20]

> If a test of significance is based on a sample of convenience, watch out.

*Example 6.* Academic gains were made by minority children in the Head-start preschool program, but tended to evaporate when the children went on to regular schools. As a result, Congress established Project Follow Through to provide continued support for minority children in regular schools. Seven sponsors were given contracts to run project classrooms according to different educational philosophies, and certain other classrooms were used as controls. SRI (a consulting firm based in Stanford) was hired to evaluate the project, for the Department of Health, Education, and Welfare.[21] One important question was whether the project classrooms really were different from the control classrooms.

To see whether or not there were real differences, SRI devised an implementation score to compare project classrooms with control classrooms. This score involved observing the classrooms to determine, for instance, the amount of time children spent playing, working independently, asking questions of the teacher, and so on. The results for one sponsor, Far West Laboratory, are shown in table 2.

Table 2. SRI Implementation Scores for 20 Far West Laboratory classrooms. Scores are between 0 and 100.

| *Site* | *Classroom scores* | | | |
|---|---|---|---|---|
| Berkeley | 73 | 79 | 76 | 72 |
| Duluth | 76 | 84 | 81 | 80 |
| Lebanon | 82 | 76 | 84 | 81 |
| Salt Lake City | 81 | 86 | 76 | 80 |
| Tacoma | 78 | 72 | 78 | 71 |

The average of these 20 scores is about 78; their SD is about 4.2. The average score for the control classrooms was about 60, so the difference is 18 points. As far as the SRI implementation score is concerned, the Far West classrooms are very different from the control classrooms. So far, so good. However, SRI was not satisfied. They wished to make a $z$-test,

> to test whether the average implementation score for Follow Through was significantly greater than the average for Non-Follow Through.

The computation is as follows.[22] The SE for the sum of the scores is estimated as $\sqrt{20} \times 4.2 \approx 19$. The SE for their average is $19/20 \approx 1$ and $z \approx (78-60)/1 = 18$. Now

The inference is:

> the overall Far West classroom average is significantly different from the Non-Follow Through classroom average of 60.

*Discussion.* The arithmetic is all in order, and the procedure may seem reasonable at first. But there is a real problem, because SRI did not have a chance model for the data. It is hard to invent a plausible one. SRI might be thinking of the 20 treatment classrooms as a sample from the population of all classrooms. But they didn't choose their 20 classrooms by simple random sampling, or even by some more complicated probability method. In fact, no clear procedure for choosing the classrooms was described in the report. This was a sample of convenience, pure and simple.

SRI might be thinking of measurement error. Is there some "exact value" for Far West, which may or may not be different from the one for controls? If so, is this a single number? Or does it depend on the site? on the classroom? the teacher? the students? the year? Or are these part of the error box? If so, isn't the error box different from classroom to classroom, or site to site? Why are the errors independent?

The report covers 500 pages, and there isn't a single one which touches on these problems. It is taken as self-evident that a test of significance can be used to compare the average of any sample, no matter where it comes from, with an external standard. The whole argument to show that the project classrooms differ from the controls rests on these tests, and the tests rest on nothing. SRI does not have a simple random sample of size 20, or 20 repeated measurements on the same quantity. It has 20 numbers. These numbers have chance components, but almost nothing is understood about the mechanism which generated them. Under these conditions, a test of significance is an act of intellectual desperation.

We went down to SRI to discuss these issues with the investigators. They insisted that they had taken very good statistical advice when designing their study, and were only doing what everybody else did. We pressed our arguments. The discussion went on for several hours. Eventually, the senior investigator said:

> Look. When we designed this study, one of our consultants explained that some day, someone would arrive out of the blue and say that none of our statistics made any sense. So you see, everything was very carefully considered.

## Exercise Set D

1. One term, there were 600 students who took the final in Statistics 2 at the University of California, Berkeley. The average score was 65, and the SD was 20 points. At the beginning of the next academic year, the 25 teaching assistants assigned to the course took exactly the same test. The TAs averaged 72, and their SD was 20 points too.[23] Did the TAs do significantly better than the students? If appropriate, make a two-sample $z$-test. If this isn't appropriate, explain why not.

2. The five planets known to the ancient world may be divided into two groups: the *inner planets* (Mercury and Venus), which are closer to the Sun than the Earth; and the *outer planets* (Mars, Jupiter, and Saturn), which are farther from the Sun. The densities of these planets are shown below; the density of the Earth is taken as 1.

| Mercury | Venus | Mars | Jupiter | Saturn |
|---|---|---|---|---|
| 0.68 | 0.94 | 0.71 | 0.24 | 0.12 |

The two inner planets have an average density of 0.81, while the average density for the three outer planets is 0.36. Is this difference statistically significant?[24] Or does the question make sense?

3. Two researchers studied the relationship between infant mortality and environmental conditions in Dauphin County, Pennsylvania. As a part of the study, the researchers recorded, for each baby born in Dauphin County during a six-month period, in what season the baby was born, and whether or not the baby died before reaching one year of age.[25] If appropriate, test to see whether infant mortality depends on season of birth. If a test is not appropriate, explain why not.

| | *Season of birth* | |
|---|---|---|
| | *July–Aug.–Sept.* | *Oct.–Nov.–Dec.* |
| Died before one year | 35 | 7 |
| Lived one year | 958 | 990 |

4. In the WISC block design test, subjects are given colored blocks and asked to assemble them to make different patterns shown in pictures. As part of Cycle II of the Health Examination Survey, this test was given to a nationwide sample of children age 6 to 9, drawn by probability methods. Basically, this was a multistage cluster sample of the kind used by the Current Population Survey (chapter 22). There were 1,652 children in the sample with family incomes in the range \$5,000 to \$7,000 a year: these children averaged 14 points on the test, and the SD was 8 points. There were 813 children in the sample with family incomes in the range \$10,000 to \$15,000 a year: these children averaged 17 points on the test, and the SD was 12 points. (The study was done in 1963–65, which explains the dollars.[26]) Someone asks whether the difference between the averages can be explained as chance variation.

   (a) Does this question make sense?
   (b) Can it be answered on the basis of the information given?

   Explain briefly.

5. Political analysts think that states matter: different states have different political cultures, which shape voters' attitudes.[27] After controlling for certain demographic variables, investigators estimate the effect of state of residence on party affiliation (Republican or Democratic). The data base consists of 55,145 persons surveyed by CBS/*New York Times* over a six-year period in the U.S. The null hypothesis—no difference among states—is rejected ($P \approx 0$, adjusted for multiple comparisons across states). True or false, and explain briefly: since $P$ is tiny, there are big differences in state political cultures.

6. An investigator asked whether political repression of left-wing views during the McCarthy era was due to "mass opinion or elite opinion."[28] He measured the effect of mass and elite opinion on the passage of repressive laws. (Effects were measured on a standardized scale going from $-1$ to $+1$.) Opinions were measured by surveys of—

   > .... a sample of the mass public and the political elites.... The elites selected were in no sense a random sample of the state elites.... Instead, the elite samples represent only themselves.... The [effect of] mass opinion is $-0.06$; for elite opinion it is $-0.35$ (significant beyond .01). Thus political repression

occurred in states with relatively intolerant elites. Beyond the intolerance of elites, the preferences of the mass public seemed to matter little.

Comment briefly on the use of statistical tests.

*The answers to these exercises are on p. A103.*

## 5. DOES THE DIFFERENCE PROVE THE POINT?

Usually, an investigator collects data to prove a point. If the results can be explained by chance, the data may not prove anything. So the investigator makes a test of significance to show that the difference was real. However, the test has to be told what "chance" means. That is what the box model does, and if the investigator gets the box model wrong, the results of the test may be quite misleading. Section 4 made this point, and the discussion continues here.

For example, take an ESP experiment in which a die is rolled, and the subject tries to make it land showing six spots.[29] This is repeated 720 times, and the die lands six in 143 of these trials. If the die is fair, and the subject's efforts have no effect, the die has 1 chance in 6 to land six. So in 720 trials, the expected number of sixes is 120. There is a surplus of $143 - 120 = 23$ sixes.

Is the difference real, or a chance variation? That is where a test of significance comes in. The null hypothesis can be set up as a box model: the number of sixes is like the sum of 720 draws from the box $\boxed{0\,0\,0\,0\,0\,1}$. The SE for the sum of the draws is

$$\sqrt{720} \times \sqrt{1/6 \times 5/6} = 10.$$

So $z = (143 - 120)/10 = 2.3$, and $P \approx 1\%$. The difference looks real.

Does the difference prove that ESP exists? As it turned out, in another part of the experiment the subject tried to make the die land showing aces, and got too many sixes. In fact, whatever number the subject tried for, there were too many sixes. This is not good evidence for ESP.

Did the $z$-test lead us astray? It did not. The test was asked whether there were too many sixes to explain by chance. It answered, correctly, that there were. But the test was told what "chance" meant: rolling a fair die. This assumption was used to compute the expected value and SE in the formula for $z$. The test proves that the die was biased, not that ESP exists.

A test of significance can only tell you that a difference is there. It cannot tell you the cause of the difference. The difference could be there because the investigator got the box model wrong, or made some other mistake in designing the study.

| A test of significance does not check the design of the study. |
|---|

Tests of significance have to be told what chances to use. If the investigator gets the box model wrong, as in the ESP example, do not blame the test.

*Example 7.* Tart's experiment on ESP was discussed in section 5 of chapter 26. A machine called the "Aquarius" picked one of 4 targets at random, and subjects tried to guess which one. The subjects scored 2,006 correct guesses in 7,500 tries, compared to the chance level of $1/4 \times 7{,}500 = 1{,}875$. The difference was $2{,}006 - 1{,}875 = 131$, $z \approx 3.5$, and $P \approx 2/10{,}000$ (one-tailed). What does this prove?

*Discussion.* The difference is hard to explain as a chance variation. That is what the $z$-test shows. But was it ESP? To rule out other explanations, we have to look at the design of the study. Eventually, statisticians got around to checking Tart's random number generators. These generators had a flaw: they seldom picked the same target twice in a row. In the experiment, the Aquarius lit up the target after each guess. Subjects who noticed the pattern, or picked new targets each time for some other reason, may have improved their chances due to the flaw in the random number generator. Tart's box model—which defined the chances for the test—did not correspond to what the random number generator was really doing.

Tart began by denying that the non-randomness in the numbers made any difference. Eventually, he replicated the experiment. He used better random number generators, and tightened up the design in other ways too. In the replication, subjects guessed at about the chance level. There was no ESP. The subjects in both experiments were students at the University of California, Davis. Tart's main explanation for the failure to replicate—"a dramatic change" in student attitudes between experiments.[30]

> In the last year or two, students have become more serious, competitive and achievement-oriented than they were at the time of the first experiment. Such "uptight" attitudes are less compatible with strong interest and motivation to explore and develop a "useless" talent such as ESP. Indeed, we noticed that quite a few of our participants in the present experiment did not seem to really "get into" the experiment and were anxious to "get it over with."

## Exercise Set E

1. Exercise 7 on p. 482 discussed an experiment where flex-time was introduced at a plant, for a sample of 100 employees. For these employees, on average, absenteeism dropped from 6.3 to 5.5 days off work. A test indicated that this difference was real. Is it fair to conclude that flex-time made the difference? If not, what are some other possible explanations for the drop in absenteeism?

2. Chapter 1 discussed the Salk vaccine field trial, where there were many fewer polio cases in the vaccine group than in the control group. A test of significance showed that the difference was real (exercise 2 on p. 515). Is it fair to conclude that the vaccine protected the children against polio? If not, what are some other possible explanations?

3. Saccharin is used as an artificial low-calorie sweetener in diet soft drinks. There is some concern that it may cause cancer. Investigators did a bioassay on rats. (Bioassays are discussed in exercise 5 on p. 552.) In the treatment group, the animals got

2% of their daily food intake in the form of saccharin. The treatment group had a higher rate of bladder cancer than the control group, and the difference was highly significant. The investigators concluded that saccharin probably causes cancer in humans. Is this a good way to interpret the $P$-value?

4. A company has 7 male employees and 16 female. However, the men earn more than the women, and the company is charged with sex discrimination in setting salaries. One expert reasons as follows:

   There are $7 \times 16 = 112$ pairs of employees, where one is male and the second female. In 68 of these pairs, the man earns more. If there was no sex discrimination, the man would have only a 50–50 chance to earn more. That's like coin tossing. In 112 tosses of a coin, the expected number of heads is 56, with an SE of about 5.3. So

$$z = \frac{\text{obs} - \text{exp}}{\text{SE}} \approx \frac{68 - 56}{5.3} \approx 2.3$$

   And $P \approx 1\%$. That's sex discrimination if I ever saw it.

   Do you agree? Answer yes or no, and explain.

*The answers to these exercises are on p. A103.*

## 6. CONCLUSION

When a client is going to be cross-examined, lawyers often give the following advice:

> Listen to the question, and answer the question. Don't answer the question they should have asked, or the one you wanted them to ask. Just answer the question they really asked.

Tests of significance follow a completely different strategy. Whatever you ask, they answer one and only one question:

> How easy is it to explain the difference between the data and what is expected on the null hypothesis, on the basis of chance variation alone?

Chance variation is defined by a box model. This model is specified (explicitly or implicitly) by the investigator. The test will not check to see whether this model is relevant or plausible. The test will not measure the size of a difference, or its importance. And it will not identify the cause of the difference.

Often, tests of significance turn out to answer the wrong question. Therefore, many problems should be addressed not by testing but by estimation. That involves making a chance model for the data, defining the parameter you want to estimate in terms of the model, estimating the parameter from the data, and attaching a standard error to the estimate.

Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory. Another reason is that many investigators just cannot be bothered to set up chance models. The language of testing makes it easy to bypass the model, and talk about "statistically significant" results. This sounds so impressive, and there is so much

mathematical machinery clanking around in the background, that tests seem truly scientific—even when they are complete nonsense. St. Exupéry understood this kind of problem very well:

> When a mystery is too overpowering, one dare not disobey.
> —*The Little Prince*[31]

## 7. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. True or false and explain briefly.
   (a) A difference which is highly significant can still be due to chance.
   (b) A statistically significant number is big and important.
   (c) A $P$-value of 4.7% means something quite different from a $P$-value of 5.2%.

2. Which of the following questions does a test of significance deal with?
   (i) Is the difference due to chance?
   (ii) Is the difference important?
   (iii) What does the difference prove?
   (iv) Was the experiment properly designed?

   Explain briefly.

3. Two investigators are testing the same null hypothesis about box X, that its average equals 50. They agree on the alternative hypothesis, that the average differs from 50. They also agree to use a two-tailed $z$-test. The first investigator takes 100 tickets at random from the box, with replacement. The second investigator takes 900 tickets at random, also with replacement. Both investigators get the same SD of 10. True or false: the investigator whose average is further from 50 will get the smaller $P$-value. Explain briefly.[32]

4. In employment discrimination cases, courts have held that there is proof of discrimination when the percentage of blacks among a firm's employees is lower than the percentage of blacks in the surrounding geographical region, provided the difference is "statistically significant" by the $z$-test. Suppose that in one city, 10% of the people are black. Suppose too that every firm in the city hires employees by a process which, as far as race is concerned, is equivalent to simple random sampling. Would any of these firms ever be found guilty of discrimination by the $z$-test? Explain briefly.

5. The inner planets (Mercury, Venus) are the ones closer to the sun than the Earth. The outer planets are farther away. The masses of the planets are shown below, with the mass of the Earth taken as 1.

| Mercury | Venus | Mars | Jupiter | Saturn | Uranus | Neptune | Pluto |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.81 | 0.11 | 318 | 95 | 15 | 17 | 0.8 |

   The masses of the inner planets average 0.43, while the masses of the outer planets average 74. Is this difference statistically significant?[33] Or does the question make sense? Explain briefly.

6. Using election data, investigators make a study of the various factors influencing voting behavior. They estimate that the issue of inflation contributed about 7 percentage points to the Republican vote in a certain election. However, the standard error for this estimate is about 5 percentage points. Therefore, the increase is not statistically significant. The investigators conclude that "in fact, and contrary to widely held views, inflation has no impact on voting behavior."[34] Does the conclusion follow from the statistical test? Answer yes or no, and explain briefly.

7. According to Census data, in 1950 the population of the U.S amounted to 151.3 million persons, and 13.4% of them were living in the West. In 2000, the population was 281.4 million, and 22.5% of them were living in the West.[35] Is the difference in percentages practically significant? statistically significant? Or do these questions make sense? Explain briefly.

8. According to Current Population Survey data for 1985, 50% of the women age 16 and over in the United States were employed. By 2005, the percentage had increased to 59%.[36] Is the difference in percentages statistically significant?
   (a) Does the question make sense?
   (b) Can you answer it based on the information given?
   (c) Can you answer it if you assume the Current Population Survey was based on independent simple random samples in each year of 50,000 women age 16 and over?

9. In 1970, 36% of first-year college students thought that "being very well off financially is very important or essential." By 2000, the percentage had increased to 74%.[37] These percentages are based on nationwide multistage cluster samples.
   (a) Is the difference important? Or does the question make sense?
   (b) Does it make sense to ask if the difference is statistically significant? Can you answer on the basis of the information given?
   (c) Repeat (b), assuming the percentages are based on independent simple random samples of 1,000 first-year college students drawn each year.

10. R. E. Just and W. S. Chern claimed that the buyers of California canning tomatoes exercised market power to fix prices. As proof, the investigators estimated the price elasticity of demand for tomatoes in two periods—before and after the introduction of mechanical harvesters. (An elasticity of $-5$, for instance, means that a 1% increase in prices causes a 5% drop in demand.) They put standard errors on the estimates.

    In a competitive market, the harvester should make no difference in demand elasticity. However, the difference between the two estimated elasticities—pre-harvester and post-harvester—was almost statistically significant ($z \approx 1.56$, $P \approx 5.9\%$, one-tailed). The investigators tried several ways of estimating the price elasticity before settling on the final version.[38] Comment briefly on the use of statistical tests.

11. A market research company interviews a simple random sample of 3,600 persons in a certain town, and asks what they did with their leisure time last

year: 39.8% of the respondents read at least one book, whereas 39.3% of them entertained friends or relatives at home.[39] A reporter wants to know whether the difference between the two percentages is statistically significant. Does the question make sense? Can you answer it with the information given?

12. There have been arguments about the validity of identification based on DNA matching in criminal cases. One problem is that different subgroups may have different frequencies of "alleles," that is, variants of a gene. What is rare in one group may be common in another. Some empirical work has been done, to measure differences among subgroups. According to one geneticist,[40]

    > Statistical significance is an objective, unambiguous, universally accepted standard of scientific proof. When differences in allele frequencies among ethnic groups are statistically significant, it means that they are real—the hypothesis that genetic differences among ethnic groups are negligible cannot be supported.

    Comment briefly on this interpretation of statistical significance.

## 8. SPECIAL REVIEW EXERCISES

1. Oregon has an experimental boot camp program to rehabilitate prisoners before their release. The object is to reduce the "recidivism rate"—the percentage who will be back in prison within three years. Prisoners volunteer for the program, which lasts several months. Some prisoners drop out before completing the program.

   To evaluate the program, investigators compared prisoners who completed the program with prisoners who dropped out. The recidivism rate for those who completed the program was 29%. For the dropouts, the recidivism rate was 74%. On this basis, the investigators argued that the program worked. However, a second group of investigators was skeptical.

   (a) Was this an observational study or a controlled experiment?

   (b) What was the treatment group? the control group?

   (c) Why might the second group of investigators have been skeptical?

   (d) The second group of investigators combined the graduates of the program with the dropouts. For the combined group, the recidivism rate was 36%. By comparison, among prisoners who did not volunteer for treatment, the recidivism rate was 37%.

      (i) What were the treatment and control groups used by the second group of investigators?

      (ii) Do their data support the idea that treatment works? or that the program has no effect? Explain your answer.

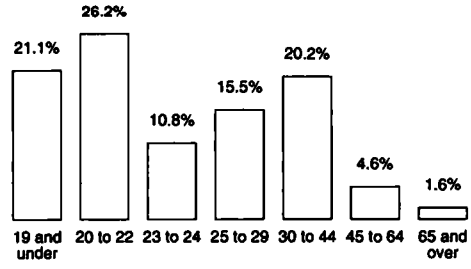
   (e) The description of the studies and the data is paraphrased from the *New York Times*.[41] Assume the numbers are correct. Did most of the prisoners who volunteered for the program complete it, or did most drop out? Explain briefly.

2. A study of baseball players shows that left-handed players have a higher death rate than right-handers. One observer explained this as "due to confounding: baseball players are more likely to be left-handed than the general population, and the players have higher death rates too." Is that a good explanation for the data? Answer yes or no, and explain briefly.

3. Schools in Northern Ireland are run on the English system. "Grammar Schools" and "Secondary Intermediate Schools" are both roughly equivalent to U.S. high schools, but students who plan on attending college generally go to Grammar Schools. Before graduation, students in both types of schools take standardized proficiency examinations.

   At Grammar Schools, Catholic students do a little better on the proficiency exams than Protestant students. At the Secondary Intermediate Schools too, Catholic students do a little better.[42] True or false, and explain: if you combine the results from both kinds of schools, the Catholic students must do a little better on the proficiency exams than the Protestants.

4. The City University of New York has about 200,000 students on 21 campuses. The figure below (adapted from the *New York Times*) shows the distribution of these students by age. For example, 21.1% of them were age 19 and under. The percentages start high, rise a little, then drop, climb, and finally drop again. How can this pattern be explained?




Note: Figure redrawn from original, copyright 1991 by the *New York Times*; reproduced by permission.

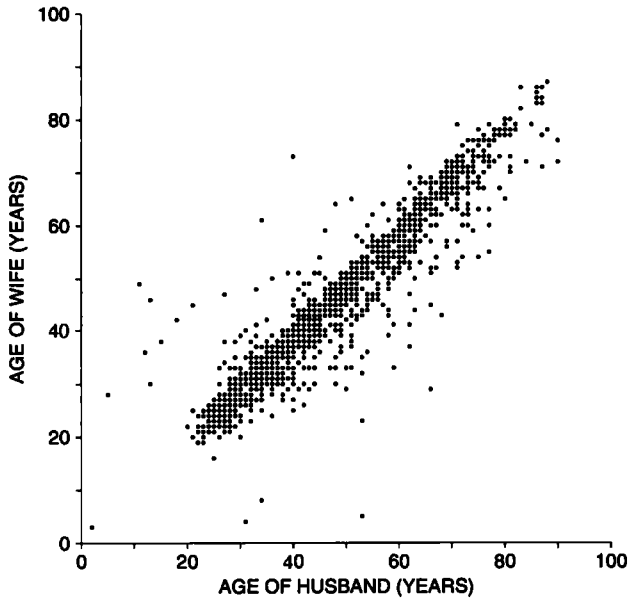
5. Data from one month of the National Health Interview Survey are shown below. (The survey is done monthly by the Census Bureau for the National Center for Health Statistics.) For example, 70% of the people age 18–64 ate breakfast every day, compared to 90% of the people age 65 and over. True or false: the data show that as people age, they adopt more healthful lifestyles. Explain your answer. If false, how do you account for the pattern in the data?

| *Age* | *Eats breakfast* | *Current drinker* | *Current smoker* |
|---|---|---|---|
| 18–64 | 70% | 40% | 35% |
| 65 and over | 90% | 10% | 15% |

Note: Percents are rounded. Source: *Statistical Abstract*, 1988, Table 178.

6. The U.S. Department of Justice made a study of 12,000 civil jury cases that were decided one year in state courts in the nation's 75 largest counties.[43] Juries gave money damages to plaintiffs in 55% of the cases. The median amount was $30,000, and the average was $600,000. Percentiles were computed for this distribution. Investigator A looks at the difference between the 90th percentile and the 50th percentile. Investigator B looks at the difference between the 50th percentile and the 10th percentile. Which difference is bigger? Or are they about the same? Explain briefly.

7. The scatter diagram below shows ages of husbands and wives in Ohio. Data were extracted from the March Current Population Survey. Or did something go wrong? Explain your answer.




8. For the data set below, find the regression equation for predicting $y$ from $x$.

| $x$ | $y$ |
|---|---|
| 1 | 1 |
| 8 | 4 |
| 10 | 6 |
| 10 | 12 |
| 14 | 12 |
| 17 | 7 |

9. Investigators are studying the relationship between income and education, for women age 25–54 who are working.

   (a) Investigator A computes the correlation between income and education for all these women. Investigator B computes the correlation only for

women who have professional, technical, or managerial jobs. Who gets the higher correlation? Or should the correlations be about the same? Explain.

(b) Investigator C computes the correlation between income and education for all the women. Investigator D looks at each state separately, computes the average income and average education for that state—and then computes the correlation coefficient for the 50 pairs of state averages. Which investigator gets the higher correlation? Or should the correlations be about the same? Explain.

10. Data on the heights of fathers and sons can be summarized as follows:

average height of fathers $\approx$ 68 inches, SD $\approx$ 2.7 inches
average height of sons $\approx$ 69 inches, SD $\approx$ 2.7 inches, $r \approx 0.5$

The scatter diagram is football-shaped. On average, the sons of the 72-inch fathers are _____ 6 inches taller than the sons of the 66-inch fathers. Fill in the blank, using one of the options below, and explain your reasoning.

(i) just about
(ii) somewhat more than
(iii) somewhat less than

11. A university made a study of all students who completed the first two years of undergraduate work. The average first-year GPA was 3.1, and the SD was 0.4. The correlation between first-year and second-year GPA was 0.4. The scatter diagram was football shaped.

Sally Davis was in the study. She had a first-year GPA of 3.5, and her second-year GPA was just about average—among those who had a first-year GPA of 3.5. What was her percentile rank on the second-year GPA, relative to all the students in the study? If this cannot be determined from the information given, say what else you need to know, and why.

12. A report by the Environmental Defense Association is discussing the relationship between air pollution and annual death rates for a sample of 47 major cities in the U.S. The average death rate is reported as 9/1,000 and the SD is 3/1,000. The r.m.s. error of the regression line for predicting death rates from air pollution is reported as 4/1,000. Is there anything wrong with the numbers? Or do you need more information to decide? Explain briefly.

13. For women age 25–54 working full time in the U.S. in 2005, the relationship between income and education (years of schooling completed) can be summarized as follows:

average education $\approx$ 14.0 years, SD $\approx$ 2.5 years
average income $\approx$ \$38,000, SD $\approx$ \$35,000, $r \approx 0.34$

If you take the women with 8 years of education, the SD of their incomes will be _____ $\sqrt{1 - .34^2} \times \$35{,}000$. Fill in the blank, and explain briefly. Options: less than, just about, more than.

14. About 1.5 million high-school students took the SATs in 2005. The regression equation for predicting the Math SAT score from the Verbal SAT score is

$$\text{predicted M-SAT} = 0.6 \times \text{V-SAT} + 220$$

The r.m.s. error of the regression line is 80 points. (The scatter diagram is football-shaped; numbers have been simplified a little.) About 50,000 students scored 500 points on the V-SAT. Of these students, about how many scored better than 500 on the M-SAT? Or do you need more information?

15. Three cards are dealt off the top of a well-shuffled deck. Find the chance that—

    (a) You only get kings.
    (b) You get no kings.
    (c) You get no face cards.
    (d) You get at least one face card.

    *Reminder.* A deck has 52 cards. There are 4 suits—clubs, diamonds, hearts, and spades. In each suit, there are 4 face cards—jack, queen, king, ace—and 9 cards numbered 2 through 10.

16. A die is rolled 6 times. Find the chance of getting 3 aces and 3 sixes.
    *Reminder.* A die has 6 faces, showing 1 through 6 spots. An ace is ⚀. Each face is equally likely to come up.

17. According to *Esquire Magazine*,

    > If you want to play roulette, do it in Atlantic City, where the house lets you "surrender" on the results of 0 and 00—that is, it returns half your wager.

    A gambler in Atlantic City plays roulette 100 times, staking $1 on red each time. Find the chance that he comes out ahead of the game.

    *Reminder.* The roulette wheel has 38 pockets, numbered 0, 00, and 1 through 36 (figure 3 on p. 282). The green numbers are 0 and 00. Of the other numbers, half are red and half are black. If you bet $1 on red and a red number comes up, you win $1. If a black number comes up, you lose $1. But if 0 or 00 comes up, you only lose $0.50—because of the "surrender."

18. A nationwide telephone survey used random digit dialing. Out of 1,507 respondents, 3% said they had been homeless at some point in the last five years. Is there selection bias in this 3% estimate? Which way does the bias go? Discuss briefly.

19. R. C. Lewontin wrote a critical review of *The Social Organization of Sexuality* by E. O. Laumann and others. Laumann was using data from a sample survey, in which respondents answered questions about their sexual behavior, including the number of partners in the previous five-year period. On average, among heterosexuals, men reported having about twice as many partners as women. Lewontin thought this was a serious inconsistency, showing that respondents "are telling themselves and others enormous lies." Laumann

replied that you should not use averages to summarize such skewed and long-tailed distributions.[44]

(a) Why is it inconsistent for men to report having twice as many partners as women?
(b) Evaluate Laumann's response.
(c) One objective of Laumann's study was to get baseline data on the epidemiology of AIDS. However, about 3% of the population (including homeless people and people in jail) were deliberately excluded from the sample. Lewontin considered this to be a serious flaw in the design of the study. Do you agree or disagree? Why?
(d) The non-response rate was about 20%. Does this matter? Explain your answer.

20. A certain town has 25,000 families. These families own 1.6 cars, on the average; the SD is 0.90. And 10% of them have no cars at all. As part of an opinion survey, a simple random sample of 1,500 families is chosen. What is the chance that between 9% and 11% of the sample families will not own cars? Show work.

21. The Census Bureau is planning to take samples in several cities, in order to estimate the percentage of the population in those areas with incomes below the poverty level. They will interview 1,000 people in each city that they study. Other things being equal:

(i) The accuracy in New York (population 8,000,000) will be about the same as the accuracy in Buffalo (population 300,000).
(ii) The accuracy in New York will be quite a bit higher than in Buffalo.
(iii) The accuracy in New York will be quite a bit lower than in Buffalo.

Choose one option, and explain briefly.

22. A market research company knows that out of all car owners in a certain large town, 80% have cell phones. The company takes a simple random sample of 500 car owners. What is the chance that exactly 400 of the car owners in the sample will have cell phones?

23. (a) What's wrong with quota samples?
(b) What's the difference between a cluster sample and a sample of convenience?
(c) What are the advantages and disadvantages of a cluster sample compared to a simple random sample?

24. (Hypothetical.) The Plaintiff's Bar Association estimates that 10% of its members favor no-fault auto insurance. This estimate is based on 2,500 questionnaires filled out by members attending a convention. True or false, and explain: the SE for this estimate is 0.6 of 1%, because

$$\sqrt{2{,}500} \times \sqrt{0.1 \times 0.9} = 15, \quad \frac{15}{2{,}500} = 0.6 \text{ of } 1\%.$$

25. A cable company takes a simple random sample of 350 households from a city with 37,000 households. In all, the 350 sample households had 637 TV sets. Fill in the blanks, using the options below.

    (a) The observed value of the _____ is 637.
    (b) The observed value of the _____ is 1.82.
    (c) The expected value of the _____ is equal to the _____.

    Options:

    (i) total number of TV sets in the sample households
    (ii) average number of TV sets per household in the sample
    (iii) average number of TV sets per household in the city

26. An airline does a market research survey on travel patterns. It takes a simple random sample of 225 people aged 18 and over in a certain city, and works out the 95%-confidence interval for the average distance they travelled on vacations in the previous year. This was 488 to 592 miles. Say whether each statement below is true or false; give reasons. If there is not enough information to decide, explain what else you need to know.

    (a) The average of the 225 distances is about 540 miles.
    (b) The SD of the 225 distances is about 390 miles.
    (c) The histogram for the 225 distances follows the normal curve.
    (d) The probability histogram for the sample average is close to the normal curve.
    (e) The probability histogram for the population average is close to the normal curve.
    (f) A 95%-confidence interval based on a sample of 450 people will be about half as wide as one based on a sample of 225 people.

27. The National Assessment of Educational Progress (NAEP) tests nationwide samples of students in school.[45] Here is an item from one of the mathematics tests.

    One plan for a state income tax requires those persons with income of $10,000 or less to pay no tax and those persons with income greater than $10,000 to pay a tax of 6 percent only on the part of their income that exceeds $10,000. A person's effective tax rate is defined as the percent of total income that is paid in tax. Based on this definition, could any person's effective tax rate be 5 percent? Could it be 6 percent?
    [Answer: People with incomes of $60,000 pay 5%, nobody pays 6%.]

    Of the grade 12 students in the sample, only 3% could answer this question correctly. The likely size of the chance error in the 3% is about _____.

    (a) Can you fill in the blank if a cluster sample of 1,000 students was tested? If so, what is the answer? If not, why not?
    (b) Can you fill in the blank if a simple random sample of 1,000 students was tested? If so, what is the answer? If not, why not?

28. Courts have ruled that standard errors and confidence intervals take bias into account.[46] Do you agree? Answer yes or no, and explain briefly.

29. One month, the Current Population Survey interviewed 54,000 households, and estimated that 94.2% of all households in the U.S. had telephones. Choose one option, and explain.

    (i) The standard error on the 94.2% can be computed as follows:

    $$\sqrt{54{,}000} \times \sqrt{0.942 \times 0.058} \approx 54, \quad \frac{54}{54{,}000} \times 100\% \approx 0.1 \text{ of } 1\%$$

    (ii) The standard error on the 94.2% can be computed some other way.
    (iii) Neither of the above

30. You may assume the Gauss model with no bias. Say whether each assertion is true or false, and why. If (c) is true, say how to do the calculations.

    (a) If all you have is one measurement, you can't estimate the likely size of the chance error in it—you'd have to take another measurement, and see how much it changes.
    (b) If all you have is one hundred measurements, you can't estimate the likely size of the chance error in their average—you'd have to take another hundred measurements, and see how much the average changes.
    (c) If all you have is one hundred measurements, you can estimate (i) the likely size of the chance error in a single measurement, and (ii) the likely size of the chance error in the average of all one hundred measurements.

31. A laboratory makes 25 repeated measurements on the molecular weight of a protein (in "kilo-Daltons"). The average is 119, and the SD is 15. The lab now wants to estimate the likely size of certain chance errors. Fill in the blanks, using the options below; some options will be left over. You may assume the Gauss model, with no bias. Explain your answers.

    (a) The chance error in one measurement is about _____.
    (b) The chance error in the average of the measurements is about _____.

    Options:

    $$15/25 \qquad 15/\sqrt{25} \qquad 15 \qquad 15 \times \sqrt{25} \qquad 15 \times 25$$

32. Feather color in Leghorn chickens is controlled by one gene pair with variants $C$ and $c$. The variant $C$ is dominant and makes colored feathers; $c$ is recessive and makes white feathers. A geneticist mates a $C/c$ rooster with some $C/c$ hens and gets 24 chicks. Find the chance that half the chicks have colored feathers.[47]

33. In the U.S., there are two sources of national statistics on crime rates: (i) the FBI's Uniform Crime Reporting Program, which publishes summaries on all crimes reported to police agencies in jurisdictions covering virtually 100%

of the population; (ii) the National Crime Survey, based on interviews with a nationwide probability sample of households.[48]

In 2001, 3% of the households in the sample told the interviewers they had experienced at least one burglary within the past 12 months. The same year, the FBI reported a burglary rate of 20 per 1,000 households, or 2%. Can this difference be explained as chance error? If not, how would you explain it? You may assume that the Survey is based on a simple random sample of 50,000 households out of 100 million households.

34. A statistician tosses a coin 100 times and gets 60 heads. His null hypothesis says that the coin is fair; the alternative, that the coin is biased—the probability of landing heads is more than 50%. True or false, and explain:
    (a) If the coin is fair, the chance of getting 60 or more heads is about 3%.
    (b) Given that it lands heads 60 times, there is only about a 3% chance for the coin to be fair.
    (c) Given that it lands heads 60 times, there is about a 97% chance for the coin to be biased.

35. The Multiple Risk Factor Intervention Trial tested the effect of an intervention to reduce three risk factors for coronary heart disease—serum cholesterol, blood pressure, and smoking. The subjects were 12,866 men age 35–57, at high risk for heart disease. 6,428 were randomized to the intervention group and 6,438 to control. The intervention included counseling on diet and smoking, and in some cases therapy to reduce blood pressure. Subjects were followed for a minimum of 6 years.[49]
    (a) On entry to the study, the diastolic blood pressure of the intervention group averaged 91.0 mm Hg; their SD was 7.6 mm Hg. For the control group, the figures were 90.9 and 7.7. What do you conclude? (Blood pressure is measured in millimeters of mercury, or mm Hg.)
    (b) After 6 years, the diastolic blood pressure of the intervention group averaged 80.5 mm Hg; their SD was 7.9 mm Hg. For the control group, the figures were 83.6 and 9.2. What do you conclude?
    (c) On entry to the study, the serum cholesterol level of the intervention group averaged 253.8 mg/dl; their SD was 36.4 mg/dl. For the control group, the figures were 253.5 and 36.8. What do you conclude? (mg/dl is milligrams per deciliter.)
    (d) After 6 years, the serum cholesterol level of the intervention group averaged 235.5 mg/dl; their SD was 38.3 mg/dl. For the control group, the figures were 240.3 and 39.9. What do you conclude?
    (e) On entry to the study, 59.3% of the intervention group were smoking, compared to 59.0% for the control group. What do you conclude?
    (f) After 6 years, the percentage of smokers was 32.3% in the intervention group and 45.6% in the control group. What do you conclude?
    (g) In the treatment group, 211 men had died after 6 years, compared to 219 in the control group. What do you conclude?

36. The Gallup Poll asks respondents how they would rate the honesty and ethical standards of people in different fields—very high, high, average, low, or very low. In 2005, only 8% of the respondents gave car salesmen a rating of "very high or high," while 7% rated telemarketers as "very high or high." Is the difference between 8% and 7% real, or a chance variation? Or do you need more information? Discuss briefly. You may assume that the results are based on a simple random sample of 1,000 persons taken in 2005; each respondent rated car salesmen, telemarketers, and many other professions.[50]

37. Each respondent in the Current Population Survey of March 2005 can be classified by education and occupation. The table below shows the observed frequencies for civilian women age 25–29 in Virginia.

    (i) Women with different educational levels seem to have different occupations. Or is this just chance variation?
    (ii) If the difference is real, what accounts for it?

    (a) Can you answer these questions with the information given? If so, answer them. If not, why not?
    (b) Can you answer these questions if the data in the table resulted from a simple random sample of women age 25–29 in Virginia? If so, answer them. If not, why not?

| | *Educational level* | |
|---|---|---|
| | *High school or less* | *More than high school* |
| Professional, managerial, technical | 12 | 34 |
| Other white collar | 15 | 17 |
| Blue collar | 5 | 2 |
| Not in labor force | 31 | 14 |

Notes: "Other white collar" includes sales and clerical. "Blue collar" includes hotel and restaurant service, factory work, and so forth, as well as unemployed workers with no civilian experience.
Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

38. Defining statistical significance, a court writes: "Social scientists consider a finding of two SEs significant, meaning there is about one chance in 20 that the explanation for the difference could be random." Do you agree or disagree with this interpretation of significance? Explain briefly.[51]

39. M. S. Kanarek and associates studied the relationship between cancer rates and levels of asbestos in the drinking water, in 722 Census tracts around San Francisco Bay.[52] After adjusting for age and various demographic variables, but not smoking, they found a "strong relationship" between the rate of lung cancer among white males and the concentration of asbestos fibers in the drinking water: $P < 1/1{,}000$.

    Multiplying the concentration of asbestos by a factor of 100 was associated with an increase in the level of lung cancer by a factor of about 1.05, on average. (If tract B has 100 times the concentration of asbestos fibers in the

water as tract A, and the lung cancer rate for white males in tract A is 1 per 1,000 persons per year, a rate of 1.05 per 1,000 persons per year is predicted in tract B.)

The investigators tested over 200 relationships—different types of cancer, different demographic groups, different ways of adjusting for possible confounding variables. The $P$-value for lung cancer in white males was by far the smallest one they got.

Does asbestos in the drinking water cause lung cancer? Is the effect a strong one? Discuss briefly.

40. Belmont and Marolla conducted a study on the relationship between birth order, family size, and intelligence.[53] The subjects consisted of all Dutch men who reached the age of 19 between 1963 and 1966. These men were required by law to take the Dutch army induction tests, including Raven's intelligence test. The results showed that for any particular birth order, intelligence decreased with family size. For example, first-borns in two-child families did better than first-borns in three-child families. Results remained true even after controlling for the social class of the parents. Moreover, for each family size, measured intelligence decreased with birth order: first-borns did better than second-borns, second-borns did better than third-borns, and so on. For instance, with two-child families:

    - the first-borns averaged 2.575 on the test;
    - the second-borns averaged 2.678 on the test.

    (Raven test scores range from 1 to 6, with 1 being best and 6 worst.) The difference is small, but it could have interesting implications.

    To show that the difference was real, Belmont and Marolla made a two-sample $z$-test. The SD for the test scores was around 1 point, both for the first-borns and the second-borns, and there were 30,000 of each, so

    $$\text{SE for sum} \approx \sqrt{30{,}000} \times 1 \text{ point} \approx 173 \text{ points}$$

    $$\text{SE for average} \approx 173/30{,}000 \approx 0.006 \text{ points}$$

    $$\text{SE for difference} \approx \sqrt{(0.006)^2 + (0.006)^2} \approx 0.008 \text{ points.}$$

    Therefore, $z \approx (2.575 - 2.678)/0.008 \approx -13$, and $P$ is astonishingly small. Belmont and Marolla concluded:

    > Thus the observed difference was highly significant.... a high level of statistical confidence can be placed in each average because of the large number of cases.

    (a) What was the population? the sample? What parameters were estimated from the sample?
    (b) Was the two-sample $z$-test appropriate? Answer yes or no, and explain.

## 9. SUMMARY AND OVERVIEW

1. A result is "statistically significant" if $P$ is less than 5%, and "highly significant" if $P$ is less than 1%. However, these lines are somewhat arbitrary because there is no sharp dividing line between probable and improbable results.

2. Investigators should summarize the data, say what test was used, and report the $P$-value instead of just comparing $P$ to 5% or 1%.

3. Even if the result is statistically significant, it can still be due to chance. Data-snooping makes $P$-values hard to interpret.

4. A $z$-test can be done either one-tailed or two-tailed, depending on the form of the alternative hypothesis.

5. The $P$-value of a test depends on the sample size. With a large sample, even a small difference can be "statistically significant," that is, hard to explain as a chance variation. That doesn't necessarily make it important. Conversely, an important difference may be statistically insignificant if the sample is too small.

6. To decide whether a difference observed in the sample is important, pretend it applies to the whole population and see what it means in practical terms. This is a "test of real significance."

7. A test of significance deals with the question of whether a difference is real or due to chance variation. It does not say how important the difference is or what caused it. Nor does the test check the design of the study.

8. Usually, a test of significance does not make sense when data are available for the whole population, because there is no chance variation to screen out.

9. To test whether a difference is due to chance, you have to define the chances. That is what the box model does. A test of significance makes little sense unless there is a chance model for the data. In particular, if the test is applied to a sample of convenience, the $P$-value may be hard to interpret.

10. Part VIII of the book introduced two tests, the $z$-test and the $\chi^2$-test. The $z$-test can be used to compare the average of a sample with an external standard (chapter 26). The test can also be used to compare the averages of two samples (chapter 27). The $\chi^2$-test compares observed and expected frequencies (chapter 28).

11. Tests ask whether a difference is too large to explain by chance. Procedures are based on the descriptive statistics of part II, and the mathematical theory of part V—including the square root law (chapter 17) and the central limit theorem (chapter 18).

12. There are many pitfalls in testing, as chapter 29 indicates.

NOTES

ANSWERS TO EXERCISES

TABLES

INDEX

# Notes

## Part I. Design of Experiments

### Chapter 1. Controlled Experiments

1. The method of comparison was used in the early nineteenth century, to show that bleeding was not such an effective treatment for pneumonia. See Pierre Charles-Alexandre Louis, *Recherches sur les effets de la saignée dans quelques maladies inflammatoires: et sur l'action de l'émetique et des vésicatoires dans la pneumonie* (J. B. Baillière, Paris, 1835; English translation, 1836; reprinted by The Classics of Medicine Library, Birmingham, Alabama, 1986). For discussion, see R. H. Shryock, *The Development of Modern Medicine* (University of Pennsylvania Press, 1936, p. 163). Lind's trial on vitamin C for scurvy should also be mentioned: see K. J. Carpenter, *The History of Scurvy and Vitamin C* (Cambridge University Press, 1986).
2. Thomas Francis, Jr. et al., "An evaluation of the 1954 poliomyelitis vaccine trials—summary report," *American Journal of Public Health* vol. 45 (1955) pp. 1–63. Also see the article by P. Meier, "The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine," in J. M. Tanur et al., *Statistics: A Guide to the Unknown*, 3rd ed. (Wadsworth, 1989). There is a less-formal account in Jane S. Smith, *Patenting the Sun* (Anchor, 1990).
3. One example: anti-arrhythmic drugs probably killed substantial numbers of people. See Thomas J. Moore, *Deadly Medicine* (Simon & Schuster, 1995). For a survey of drug trials, see N. Freemantle et al., "Composite outcomes in randomized trials," *Journal of the American Medical Association* vol. 289 (2003) pp. 2554–59.
4. "Control what you can and randomize the rest" is the advice often given by statisticians. Matching or blocking will reduce variance, at the expense of complicating the analysis. Also see note 12 to chapter 19, and note 16 to chapter 27.
5. H. K. Beecher, *Measurement of Subjective Responses* (Oxford University Press, 1959, pp. 66–67). Also see Berton Roueché, *The Medical Detectives* (Washington Square Press, New York, 1984, vol. II, chapter 9). More recent references include K. B. Thomas, "General practice consultations: Is there any point in being positive?" *British Medical Journal* vol. 294 (1987) pp. 1200–2 and J. A. Turner et al., "The importance of placebo effects in pain treatment and research," *Journal of the American Medical Association* vol. 271 (1994) pp. 1609–14.
6. N. D. Grace, H. Muench and T. C. Chalmers, "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* vol. 50 (1966) pp. 684–91. We found this example in J. P. Gilbert, R. J. Light and F. Mosteller, "Assessing social innovations: An empirical guide for policy," *Benefit Cost and Policy Analysis Annual* (1974). For a review of more recent therapies, see A. J. Stanley and P. C. Hayes, "Portal hypertension and variceal haemorrhage," *Lancet* vol. 350 (1997) pp. 1235–39; there does not seem to be any survival advantage for current surgical therapies (including "TIPS," see p. 1238). But see A. J. Sanyal et al., "The North American study for the treatment of refractory ascites," *Gastroenterology* vol. 124 (2003) pp. 634–41.
7. The definition of "randomized controlled trial" is not strict. The original table included data on anticoagulants after myocardial infarct. Even in the 1980s, there was some controversy about the interpretation of clinical trials on anticoagulants. Since then, thrombolytic therapies have changed considerably, and there are many new experiments. For reviews, see—

   *Coronary Artery Disease* vol. 5 no. 4 (1994).

   "ACC/AHA guidelines for the management of patients with ST-elevation myocardial infarction: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *Circulation* vol. 110 (2004) pp. 588–636.

   J. D. Talley, "Review of thrombolytic intervention for acute myocardial infarction—is it valuable?" *Journal of the Arkansas Medical Society* 91 (1994) pp.70–79.

   C. H. Hennekens, "Thrombolytic therapy: Pre- and post-GISSI-2, ISIS-3, and GUSTO-1," *Clinical Cardiology* vol. 17 suppl. I (1994) pp. 115–7.

   R. Collins, R. Peto, S. Parish and P. Sleight, "ISIS-3 and GISSI-2: No survival advantage with tissue plasminogen activator over streptokinase, but a significant excess of strokes with tissue plasminogen activator in both trials," *American Journal of Cardiology* vol. 71 (1993) pp. 1127–30.

   M. J. Stampfer et al., "Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials," *New England Journal of Medicine* vol. 307 (1982) pp. 1180–82.

8. T. C. Chalmers, "The impact of controlled trials on the practice of medicine," *Mount Sinai Journal of Medicine* vol. 41 (1974) pp. 753–59.

## Chapter 2. Observational Studies

1. "Epidemiology" is the statistical study of disease. An excellent introductory text is Leon Gordis, *Epidemiology*, 3rd ed. (Elsevier Saunders, Philadelphia, 2004). Some references on the epidemiology of smoking—

   J. Berkson, "The statistical study of association between smoking and lung cancer," *Proceedings of the Mayo Clinic* vol. 30 (1955) pp. 319–48.

   R. A. Fisher, *Smoking: The Cancer Controversy* (Oliver and Boyd, 1959)

   J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin and E. L. Wynder, "Smoking and lung cancer: Recent evidence and a discussion of some questions," *Journal of the National Cancer Institute* vol. 22 (1959) pp. 173–203.

   U.S. Public Health Service, *Smoking and Health: Report of the Advisory Committee to the Surgeon General* (Washington, D.C., 1964).

   International Agency for Research on Cancer, *Tobacco Smoking*. Monograph 38 (Lyon, France, 1986).

   U.S. Public Health Service, *The Health Benefits of Smoking Cessation. A Report of the Surgeon General* (Washington, D.C., 1990).

   For evidence on mechanism, see—

   M. F. Denissenko et al., "Preferential formation of Benzo[a]pyrene adducts at lung cancer mutational hotspots in p53," *Science* vol. 274 (1996) pp. 430–2.

2. The Coronary Drug Project Research Group, "Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project," *New England Journal of Medicine* vol. 303 (1980) pp. 1038–41. The other drugs were estrogens (at two dose levels), dextrothyroxine, and nicotinic acid. Enrollment ran from March 1966 to October 1969. The estrogens and dextrothyroxine were discontinued, due to adverse side effects. Clofibrate and nicotinic acid lowered the cholesterol level, but did not reduce mortality. See "Clofibrate and niacin in coronary heart disease," *Journal of the American Medical Association* vol. 231 (1975) pp. 360–81.

   *The Lancet* (March 4, 1989) pp. 473–74 suggests a positive effect from nicotinic acid in late followup. However, there is evidence of harm from such drugs. See Toronto Working Group on Cholesterol Policy, *Journal of Clinical Epidemiology* vol. 43 (1990) no. 10 pp. 1021ff. Other papers are cited in note 7, chapter 29.

   An interesting sidelight: about 40 risk factors were measured at baseline. As a group, the non-adherers did seem to be in worse shape when the study began. But it was impossible to adjust out the difference between adherers and non-adherers by regression using the measured risk factors; adherence to protocol was not well related to the covariates. This suggests caution in using regression models to control for confounding in observational studies. Also see R. T. Tsuyuki et al., "A meta-analysis of the association between adherence to drug therapy and mortality," *British Medical Journal* vol. 333 (2006) pp. 15–19.

3. The quote is from D. A. Roe, *A Plague of Corn* (Cornell University Press, 1973). Another excellent reference, which reprints many of the original papers with commentary, is K. J. Carpenter, *Pellagra* (Academic Press, 1981). Also see M. Terris, editor, *Goldberger on Pellagra* (Louisiana State University Press, 1964)

   Corn was brought to Europe from America. The Indians treated corn with alkali before cooking it, which released the niacin; pellagra was not a problem for them. The pellagra epidemic in the U.S. seems to have begun when millers started extracting the germ from corn; the germ contains much of the available niacin or tryptophan. (Tryptophan is an amino acid which can be converted to niacin in the body.) In the U.S., pellagra deaths peaked around 1930, and declined fairly steadily from then on, presumably because of a general improvement in economic conditions and diet; enriching the flour came too late to have much impact. Pellagra is still endemic in parts of Africa and India. In the late 1980s, research in South Africa suggested a possible etiologic role for mycotoxins; so the infection theory may have a little truth to it after all. We would like to thank K. J. Carpenter (U.C. Berkeley) for his help with this example.

4. References—

   E. L. Wynder, J. Cornfield, P. D. Schroff and K. R. Doraiswami, "A study of environmental factors in carcinoma of the cervix," *American Journal of Obstetrics and Gynecology* vol. 68 (1954) pp. 1016–52.

   C. Buck et al., editors, *The Challenge of Epidemiology: Issues and Selected Readings*, Scientific Publication No. 505 (World Health Organization, Geneva, 1989).

   N. Muñoz, F. X. Bosch, K. V. Shah and A. Meheus, editors, *The Epidemiology of Cervical*

*Cancer and Human Papillomavirus*. International Agency for Research on Cancer, Scientific Publication no. 119 (1992).

A. S. Evans, *Causation and Disease: A Chronological Journey* (Plenum, 1993).

S. A. Cannistra and J. M. Niloff, "Cancer of the uterine cervix," *New England Journal of Medicine* vol. 334 (1996) pp. 1030–38.

A. Storey et al., "Role of a p53 polymorphism in the development of human papillomavirus-associated cancer," *Nature* vol. 393 (1998) pp. 229–34.

X. Castellsague et al., "Male circumcision, penile human papillomavirus infection, and cervical cancer in female partners," *New England Journal of Medicine* vol. 346 (2002) pp. 1105–12.

Wynder et al. found that circumcision was protective. The history is discussed by Evans, and some of the key papers are reprinted in Buck et al. Castellsague et al. conclude that circumcision is protective if the man is highly active sexually. The death rate from cervical cancer has been declining for some time. Smoking is a risk factor for this disease, so the decline in smoking may explain the decline in death rates, and screening is protective. A vaccine against papilloma virus is now available. The example was suggested by Michael Kramer (Montreal).

5. R. M. Moore et al., "The relationship of birthweight and intrauterine diagnostic ultrasound exposure," *Journal of Obstetrics and Gynecology* vol. 71 (1988) pp. 513–17. The confounding variables: race, registration status (public or private), smoking status, delivery status (full-term or premature), spontaneous abortion history, alcohol history, amniocentesis status, fetal monitoring, method of delivery, education, weeks pregnant at registration, number of prenatal visits, maternal weight and weight gain, gestational age at delivery.

   The clinical trial is U. Waldenstrom et al., "Effects of routine one-stage ultrasound screening in pregnancy: A randomized controlled trial," *Lancet* (Sept. 10, 1988) pp. 585–88. Babies exposed to ultrasound had higher weights, on average, than the controls. In the treatment group, women watched ultrasound images of the babies they were carrying. Many of them gave up smoking as a result, and smoking does cause low birthweight. The change in smoking behavior may account for the protective effect.

6. "Suicide and the Samaritans," *Lancet* (Oct. 7, 1978) pp. 772–73 (editorial). The original investigator was C. Bagley (*Social Science and Medicine*, 1968). He did not match the towns by type of gas used, and these data do not seem to be available now. The replication was by B. Barraclough et al. (*Lancet*, 1977; *Psychological Medicine*, 1978). We found this example in D. C. Hoaglin, R. J. Light, B. McPeek, F. Mosteller and M. A. Stoto, *Data for Decisions* (University Press of America, 1982, p. 133).

7. The paradox in the Berkeley data was noticed by Eugene Hammel, then associate dean of the graduate division. He resolved it with the help of two colleagues, P. Bickel and J. W. O'Connell. We are following their report, "Is there a sex bias in graduate admissions?" *Science* vol. 187 (1975) pp. 398–404. The admissions data are from fall, 1973.

8. For a review, see Myra Samuels, "Simpson's Paradox and related phenomena," *Journal of the American Statistical Association* vol. 88 (1993) pp. 81–88.

9. Some typical examples:

   (i) The confounder may be a common cause of exposure and disease.
   (ii) The confounder may be associated with exposure and cause disease.
   (iii) The confounder may be associated with disease and cause exposure.

   A common effect of exposure and disease will generally not explain the association. Paradoxically, selecting on a common effect may create a negative correlation: see the Berkson paper cited in note 1.

10. *Statistical Abstract*, 2003, table 108.

11. See note 2, chapter 1.

12. References—

    L. M. Friedman, C. D. Furberg and D. L. DeMets, *Fundamentals of Clinical Trials*, 3rd corr. ed. (Springer, 2006, p. 83).

    T. L. Lewis, T. R. Karlowski, A. Z. Kapikian, J. M. Lynch, G. W. Shaffer, D. A. George and T. C. Chalmers, "A controlled clinical trial of ascorbic acid for the common cold," *Annals of the New York Academy of Science* vol. 258 (1975) pp. 505–12.

    T. R. Karlowski, T. C. Chalmers, L. D. Frenkel, A. Z. Kapikian, T. L. Lewis and J. M. Lynch, "Ascorbic acid for the common cold," *Journal of the American Medical Association* vol. 231 (1975) pp. 1038–42.

    K. J. Carpenter, *The History of Scurvy and Vitamin C* (Cambridge University Press, 1986).

13. "Nicotinic acid" is the technical term for niacin, the pellagra-preventive factor. Apparently, the term "niacin" was introduced because "nicotinic acid" looked too ominous on flour labels. Nicotinic acid was tried in the Coronary Drug Project and had no effect.

14. The savings in lives persist over many years, and other trials give quite similar results. Screening speeds up detection by a year or so, and that seems to be enough to matter. Unpublished data were kindly provided by the late Sam Shapiro, professor of epidemiology, Johns Hopkins. In the HIP trial, there was an initial screening examination and three annual rescreenings, each including breast examination by a doctor and mammography.

    The risk of breast cancer is modulated by hormone balance, and pregnancy is protective; early first pregnancy has a marked effect. Presumably, that accounts for the gradient with income. On social gradients in disease risk, with further references, see

    J. N. Morris et al., "Levels of mortality, education, and social conditions in the 107 local education authority areas of England," *Journal of Epidemiology and Community Health* vol. 50 (1996) pp. 15–17.

    J. Pekkanen et al., "Social class, health behavior, and mortality among men and women in eastern Finland," *British Medical Journal* vol. 311 (1995) pp. 589–93.

    M. G. Marmot et al., "Contribution of job control and other risk factors to social variations in coronary heart disease," *Lancet* vol. 350 (1997) pp. 235–9.

    The key reference on the HIP trial is S. Shapiro, W. Venet, P. Strax, and L. Venet, *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986* (Hopkins, 1988). In 2000, questions were raised again about the value of screening, but the critics seem to have misinterpreted much of the evidence. For a review and further references, see D. A. Freedman, D. B. Petitti, and J. M. Robins, "On the efficacy of screening for breast cancer," *International Journal of Epidemiology* vol. 33 (2004) pp. 43–73, 1404–6.
15. For references, see note 4.
16. This example was suggested by Shanna Swan (Rochester), based on data from an observational study done at Kaiser Permanente in Walnut Creek, California.
17. *Statistical Abstract*, 2003, tables 17, 307.
18. *Federal Register*, vol. 69, no. 169, Sept. 1, 2004, pp. 53354–59. Technically, it is not sales figures that are reported, but vehicles "manufactured for [model year] 2002, as reported to the Environmental Protection Agency."
19. *Statistical Abstract*, 1971, table 118. The study was done in 1964; the same effect turns up in many other studies. If you quit smoking and survive more than a few years, your risk will drop relative to continuing smokers. See U.S. Public Health Service, *The Health Benefits of Smoking Cessation. A Report of the Surgeon General* (Washington, D.C., 1990).
20. We found the example in Friedman et al., cited in note 12 above. References—

    P. J. Schechter, W. T. Friedewald, D. A. Bronzert, M. S. Raff and R. I. Henkin, "Idiopathic hypoguesia: a description of the syndrome and a single-blind study with zinc sulfate," *International Review of Neurobiology* (1972) Supplement 1 pp. 125–39.

    R. I. Henkin, P. J. Schechter, W. T. Friedewald, D. L. DeMets and M. S. Raff, "A double blind study of the effects of zinc sulfate on taste and smell dysfunction," *American Journal of the Medical Sciences* vol. 272 (1976) pp. 285–99.
21. This example was suggested by Shanna Swan. See E. Peritz et al., "The incidence of cervical cancer and duration of oral contraceptive use," *American Journal of Epidemiology* vol. 106 (1977) pp. 462–69. Adjustments were also made for religion, smoking (a risk factor for cervical cancer), number of Pap smears before entry, and "selected infections." For additional references, see note 4.
22. Quoted by Herb Caen in the *San Francisco Chronicle*, Wednesday, August 9, 1995.
23. References—

    E. R. Greenberg et al., "A clinical trial of antioxidant vitamins to prevent colorectal adenoma," *New England Journal of Medicine* vol. 331 (1994) pp. 141–47.

    O. P. Heinonen et al., "Effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers," *New England Journal of Medicine* vol. 330 (1994) pp. 1029–35.

    For other trials and additional discussion, see—

    C. H. Hennekens et al., "Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease," *New England Journal of Medicine* vol. 334 (1996) pp. 1145–9.

    J. Virtamo, P. Pietinen, J. K. Huttunen et al., "Incidence of cancer and mortality following alpha-tocopherol and beta-carotene supplementation: A postintervention follow-up," *Journal of the American Medical Association* 290 (2003) pp. 476–85.

    D. A. Lawlor, G. D. Smith, K. R. Bruckdorfer et al., "Those confounded vitamins: What can we learn from the differences between observational vs randomised trial evidence," *Lancet* 363 (2004) pp. 1724–27.

G. S. Omenn et al., "Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease," *New England Journal of Medicine* vol. 334 (1996) pp. 1150–5.

24. The story ran November 9, 1994. The source was S. L. Johnson and L. L. Birch, "Parents' and children's adiposity and eating style," *Pediatrics* vol. 94 (1994) pp. 653–61. "Mothers who were more controlling of their children's food intake had children who showed less ability to self-regulate energy intake ($r = -.67$, $P < .0001$)."
25. This exercise is based on a story in the *San Francisco Chronicle*, January 19, 1993. The quote is edited to simplify the study design. Generally, prisoners are offered early parole as an inducement to volunteer.

# Part II. Descriptive Statistics

## Chapter 3. The Histogram

1. By Antoine de St. Exupéry. Reproduced by permission of the publisher, Harcourt Brace Jovanovich.
2. *Money Income in 1973 of Families and Persons in the United States*, Current Population Reports, Series P-60, No. 97 (January, 1975). U.S. Department of Commerce.
3. For the 1973 data, see note 2. The 2004 data are from the March 2005 Current Population Survey; a CD-ROM was supplied by the Bureau of the Census. Generally, "family income" means the total income of primary family members in a household, and includes the income of related sub-family members. Income from self-employment is net, and can be quite negative if the business lost money for a year.

   The March survey over-samples large families to address questions about availability of health insurance. We use sample weights for issues that involve the number of children in the family; see, e.g., the histogram for family size shown in figure 6, or exercise 2 in set D. We do not use weights for other questions, e.g., the correlation between education and income: the weights make surprisingly little difference.

   In many cases, unequal class intervals are forced by the design of the questionnaire. For instance, a respondent may only be asked to specify which of several ranges best describes his income; the ranges—class intervals—will be of unequal length.

   Adjusting for inflation by price indices may be problematic, because quality improvements are hard to account for. See W. D. Nordhaus, "Do real output and real wage measures capture reality? The history of lighting suggests not," in T. F. Bresnahan and R. J. Gordon, editors, *The Economics of New Goods* (Chicago University Press, 1997).
4. See note 2.
5. Among other things, data histograms in chapter 3 prepare the way for probability histograms in chapter 18. The latter can be approximated by probability densities; in particular, a histogram can "follow" the normal curve, in the sense that the curve is a good approximation to the histogram. If $f$ is a probability density on the line, areas under $f$ represent probability; the height of $f$ at $x$ represents probability per unit length, rather than probability: indeed, the probability of $x$ is 0, not $f(x)$.

   Generally, mathematicians do not bother with physical units—inches, pounds, or dollars. For statistical purposes, such units matter. The units for $f$ are inverse to the units for $x$. Thus, we present the density scale for an income histogram as "percent per thousand dollars"; the density scale for a weight histogram would be "percent per pound." For non-mathematicians, "2 percent per pound" may be more palatable than .02 $\text{lb}^{-1}$.
6. This is exact for class intervals, approximate for other intervals.
7. *Statistical Abstract*, 1971, table 118.
8. Many variables can be classified either way, depending on how you view them. Incomes, for instance, can never differ by less than a penny. Nevertheless, it is convenient to treat income as continuous—because the range is so much larger than the minimum difference.
9. With narrow class intervals, the histogram may be so ragged that its shape is impossible to make out. With wider class intervals, the shape of the histogram may be easier to see, even though some information is lost. For discussion, see P. Diaconis and D. Freedman, "On the histogram as a density estimator: $L_2$ theory," *Z. Wahrscheinlichkeitstheorie* vol. 57 (1981) pp. 453–76.
10. See I. R. Fisch, S. H. Freedman and A. V. Myatt, "Oral contraceptives, pregnancy, and blood pressure," *Journal of the American Medical Association* vol. 222 (1972) pp. 1507–10. Our discussion follows this paper. We are grateful to Shanna Swan and Michael Grossman for technical advice.

    Blood pressure is taken in two phases, *systolic* and *diastolic*. We are looking at the systolic phase. Results on the diastolic phase are quite similar. In the Contraceptive Drug Study, blood

pressures were measured by a machine. The study excluded about 3,500 women who were pregnant, post-partum, or taking hormonal medication other than the pill; these factors affect blood pressure. The Drug Study found that four age groups were enough: 17–24, 25–34, 35–44, and 45–58. The age distributions of users or non-users within each of these age groups were quite similar.

11. R. C. Tryon, "Genetic differences in maze-learning techniques in rats," 39th yearbook, *National Society for the Study of Education* part I (1940) pp. 111–19. This article is reprinted in a very nice book of readings: Anne Anastasi, *Individual Differences* (John Wiley & Sons, 1965). Tryon uses a non-linear scale for his histograms, so they look quite different from our sketches.
12. *1970 Census of Population*. See vol. 1, part 1, section 2, appendix, p. 14. U.S. Department of Commerce. Only persons age 23–99 are counted in the column for 1880; only persons age 23–82 are counted in the column for 1970.
13. K. Bemesderfer and J. May, *Social and Political Inquiry* (Belmont, California: Duxbury Press, 1972, p. 6).
14. References—

    R. A. Baron and V. M. Ransberger, "Ambient temperature and the occurrence of collective violence: The 'long, hot summer' revisited," *Journal of Personality and Social Psychology* vol. 36 (1978) pp. 351–60. The quote is edited slightly.

    J. M. Carlsmith and C. A. Anderson, "Ambient temperature and the occurrence of collective violence: A new analysis," *Journal of Personality and Social Psychology* vol. 37 (1979) pp. 337–44.

The figure is redrawn from Baron and Ransberger, by permission of the authors and copyright holder (the American Psychological Association).

## Chapter 4. The Average and the Standard Deviation

1. *Natural Inheritance* (Macmillan, London, 1889; reprinted by the American Mathematical Society Press, 1973).
2. The point where the histogram is highest, the *mode*, is sometimes used to indicate the center. This is not recommended, as minor changes in the data can cause major shifts in the mode.
3. Tom Alexander, "A revolution called plate tectonics," *Smithsonian Magazine* vol. 5, no. 10 (1975). A. Hallam, "Alfred Wegener and the hypothesis of continental drift," *Scientific American* vol. 232, no. 2 (1975). Ursula Marvin, *Continental Drift* (Smithsonian Press, 1973).
4. The Public Health Service and the National Center for Health Statistics (NCHS) are in the Department of Health and and Human Services. Data on HANES2 are from series 11 of the Vital and Health Statistics publications, and from tapes supplied by the National Center for Health Statistics and by the Inter-University Consortium for Political and Social Research. Data on HANES3 were kindly supplied on a CD-ROM by NCHS. These data, and data for HANES5, are now available on the internet, at

   http://www.cdc.gov/nchs/about/major/nhanes/datalink.htm

   (URLs cited here were alive in February 2006, but time will doubtless take its toll.) We are responsible for the interpretation of the data. For help with earlier editions of the book, we thank Dale Hitchcock, Arthur McDowell, and Bob Murphy at NCHS, as well as Dorothy Rice (UCSF). For the fourth edition, we thank Wim van Veen (Health Council of the Netherlands).

   The histograms in figures 4, 8, 9 are based on sample counts, unweighted, ages 18+; likewise for the scatter diagrams discussed in part III. Summary statistics are heavily rounded. Sample weights made little difference in HANES2, but have more noticeable effects in HANES5. The table below compares HANES5 to HANES2 (ages 18–74).

HANES2: 1976–80

| | *Men 18–74 unweighted* | *Men 18–74 weighted* | *Women 18–74 unweighted* | *Women 18–74 weighted* |
|---|---|---|---|---|
| Height | 68.78 ± 2.83 | 69.11 ± 2.82 | 63.46 ± 2.62 | 63.71 ± 2.60 |
| Weight | 170.92 ± 30.13 | 172.19 ± 29.75 | 145.71 ± 32.65 | 144.18 ± 32.27 |

HANES5: 2003–04

| | *Men 18–74 unweighted* | *Men 18–74 weighted* | *Women 18–74 unweighted* | *Women 18–74 weighted* |
|---|---|---|---|---|
| Height | 69.11 ± 3.10 | 69.61 ± 2.97 | 63.67 ± 2.76 | 64.09 ± 2.65 |
| Weight | 188.92 ± 42.95 | 193.94 ± 41.95 | 165.84 ± 43.76 | 165.32 ± 44.19 |

5. The groups in figure 3: 18–24, 25–34, 35–44, 45–54, 55–64, 65–74. On HANES2, see *Anthropometric Reference Data and Prevalence of Overweight: United States, 1976–80*; data are from

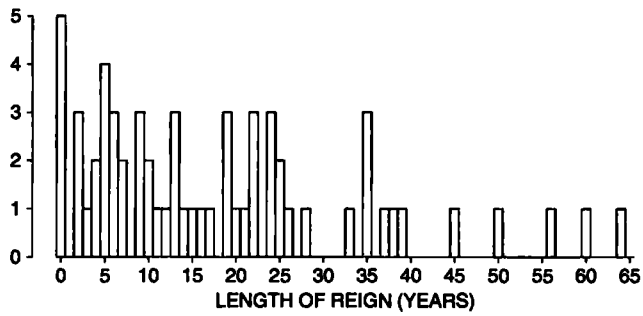
the National Health Survey, series 11, no. 238, U.S. Department of Health and Human Services, Washington, D.C. In the 1970s, the secular trend was estimated at about 0.4 inch per decade; and, over the 20-year period 1960–80, Americans did become 0.8 inches taller, on average. Furthermore, people seem to lose 0.5–1.5 inch of height as they age from 50 to 75. (One possible explanation: about 2 inches of height is made up of air spaces between the bones in the body; the body settles in on itself with age, so these air spaces get smaller and smaller.) The secular trend and the shrinking would suggest a total drop of 2.5–3.5 inches from age 20 to age 70. The observed drop in HANES2 was 2.3 inches for the men and 2.1 inches for the women, so there may have been other factors at work. We would like to thank Reubin Andres (NIH) and Stanley Garn (University of Michigan) for their help. See R. Floud, K. Wachter and A. Gregory, *Height, Health, and History* (Cambridge University Press, 1991) for a discussion of trends in height as indicators of social change. Also see Gina Kolata, *The New York Times*, July 30, 2006, p. 1.

6. See note 4 above for the data source. Cases with missing or implausible values (for instance, diastolic pressure below 30 mm) were excluded. The good news is that blood pressures have dropped by 5–10 mm since HANES2. Some of the decline may be due to increased use of anti-hypertensive medications.
7. This is exact for integer data and class intervals centered at the integers; more generally, if the mean over each class interval is the midpoint of the interval. Otherwise, it is only an approximation.
8. Data from the Current Population Survey, March 2005 (note 3 to chapter 3). See section 5.4 for discussion.
9. The basic reason is called *orthogonality* by statisticians. When errors in some situation arise from several independent sources, there is a simple and exact formula for getting the r.m.s. size of the total error: the r.m.s. errors combine like the sides of a right-angled triangle. With two orthogonal sources of error,

$$c = \sqrt{a^2 + b^2}$$

where $a$ is the r.m.s. size of the errors coming from one source, $b$ is the r.m.s. size of the errors coming from another source, and $c$ is the r.m.s. size of the total error. This fact will be used several times in the book: in regression (part III), in computing the standard error for a sum (part V), and in computing the standard error for a difference (part VIII). No such formulas are possible for the average absolute value.

10. The 68%–95% rule works quite well even for many data sets which do not follow the normal curve. Take, for example, the lengths of the reigns of the 61 English monarchs through George VI. These average 18.1 years, with an SD of 15.5 years. Their histogram is shown below, and it is nothing like the normal curve. Still, 42 out of 61, or 69%, were within 1 SD of average. And 57 out of 61, or 93%, were within 2 SDs of average. (By definition, the length of a reign is the difference between its first and last years, as reported on pp. 274–75 of the 1988 *Information Please Almanac*; this example was contributed by David Lane, Modena, Italy.)




11. The square of the SD is called the *variance*. This is often used as a measure of spread, but we do not recommend it as a descriptive statistic. For instance, the SD of weight for American men is about 40 pounds: individual men are roughly 40 pounds away from average weight. The variance of weight is

$$(40 \text{ pounds})^2 = 1600 \text{ square pounds.}$$

12. However, this formula may be vulnerable to roundoff error.
13. See note 4 above for the data source. Cases with missing or implausible values (for instance, diastolic pressure below 30 mm) were excluded.
14. See note 4 above for the data source.
15. Patricia Ruggles, *Drawing the Line* (Urban Institute Press, Washington, D.C., 1990). The description of the underclass is paraphrased from p. 105. The book discusses the impact of definitions; see also chapter 5 on the time dimension. More recent data are available from SIPP (Survey of

Income and Program Participation). See *Dynamics of Economic Well-Being: Poverty 1996–1999* (P70-91), http://www.sipp.census.gov/sipp/. Also see *Statistical Abstract*, 2003, table 700. But see Ann Huff Stevens, "Climbing out of poverty, falling back in: Measuring the persistence of poverty over multiple spells," *Journal of Human Resources* (Summer 1999).

## Chapter 5. The Normal Approximation for Data

1. Also called standard scores, $z$-scores, sigma-scores.
2. Here, we use "estimate" in its ordinary sense, "to compute approximately." "Estimate" also has a technical meaning in statistics, to be discussed in part VI.
3. See note 3 to chapter 3. The histogram, the SD, and the percentiles are computed from the CD-ROM, without weights, for primary families. The mean and SD are about right when incomes above $200,000 are censored. The percentiles are for the full distribution of primary-family incomes, the corresponding mean and SD are both about $70,000. (Note that high incomes have been top-coded by the Bureau to protect respondent confidentiality.)
4. The data are from a College Board press release, September 20, 1988; *1994 Profile of SAT and Achievement Test Takers* (Educational Testing Service, Princeton, N.J.); *Statistical Abstract*, 1993, table 265. Data for the fourth edition are from *Statistical Abstract*, 2003, table 264 and from *2005 College-Bound Seniors: Total Group Profile Report*, available at http://www.collegeboard.com. "College bound seniors" are students in a high-school graduating class who took the SAT at some point in their high-school years.

   The SD seems to drift upwards over time; it was 109 in 1975 and 113 in 1994 on the Verbal SAT. In 1994, the SAT was re-normed to a mean of 500 and an SD of 100; changes took effect in 1995. By 2005, the SD was 115. On the Math SAT (exercise 4) and many other such tests, the men have higher SDs than the women—120 vs 110; that is one reason why there are more men at the extremes of the distributions. See L. V. Hedges and A. Nowell, "Sex differences in mental test scores, variability, and numbers of high-scoring individuals," *Science* vol. 269 (1995) pp. 41–45.

   For a discussion of the decline in SAT scores, see Willard Wirtz et al., *On Further Examination. Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline* (College Entrance Examination Board, New York, 1977). We thank Susan Bryce (ETS), Paul Holland (ETS), and Howard Wainer (National Board of Medical Examiners) for their help with the third edition.
5. March 2005 Current Population Survey; data from a CD-ROM supplied by the Census Bureau.
6. *2005 College-Bound Seniors: Total Group Profile Report*.

## Chapter 6. Measurement Error

1. NBS started out as the Bureau of Weights of Measures, under the leadership of Ferdinand Hassler (born in 1770 in Switzerland, Director of the Bureau from 1830 to 1843). Hassler's great achievement was to bring the weights and measures used all over the U.S. into a high degree of consistency. The Bureau is now called NIST—The National Institute of Standards and Technology. We would like to thank H. H. Ku of the Bureau for his help with earlier editions.
2. Weight is used here instead of the more technical word *mass*. In 1983, the General Conference on Weights and Measures superseded the Treaty of the Meter, but weight was still defined by reference to a standard object. For an update on The Kilogram, see *Science* (May 12, 1995) p. 804.
3. Two major sources of chance error in the precision weighing at the Bureau are thought to be:
   - minute amount of play in the balance mechanism, especially at the knife edge;
   - slight variations in the position of the weights on the balance pans.
4. P. E. Pontius, "Measurement philosophy of the pilot program for mass calibration," *NBS Technical Note No. 288* (1966). The Bureau rejects outliers only "for cause, such as door-slam or equipment malfunction." Also see H. H. Ku, editor, *Precision Measurement and Calibration*, NBS Special Publication no. 300, vol. 1 (Washington, D.C., 1969).
5. Data were extracted from tapes supplied by the National Center for Health Statistics and by the Inter-University Consortium for Political and Social Research. Also see note 4 to chapter 4.
6. The data are from the March 2005 Current Population Survey; a CD-ROM was supplied by the Bureau of the Census. Primary families only; includes the income of related sub-family members.
7. J. N. Morris and J. A. Hardy, "Physique of London busmen," *Lancet* (1956) pp. 569–570. This reference was supplied by Eric Peritz (Jerusalem).
8. See note 14 in chapter 2.
9. The observational data are discussed in H. Zeisel, H. Kalven, Jr., and B. Buchholz, *Delay in the Court*, 4th ed. (Little, Brown & Co., 1959). The experiment was reported in Maurice Rosenberg, *The Pretrial Conference and Effective Justice* (Columbia University Press, 1964). The definitions are on p. 19, the numbers on p. 20, and the data in tables 8 and 9, pp. 48 and 52. For another discussion, see H. Zeisel. *Say It with Figures*, 6th ed. (Harper & Row, 1985, p. 141).

# Part III. Correlation and Regression

## Chapter 8. Correlation

1. There are methods for dealing with more than two variables, but these are quite complicated. Some matrix algebra is needed to follow the discussion. References—

   M. L. Eaton, *Multivariate Statistics: A Vector Space Approach* (John Wiley & Sons, 1983).
   D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, 2005).
   C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. (John Wiley & Sons, 1973).
   J. A. Rice, *Mathematical Statistics and Data Analysis*, 3d ed. (Duxbury Press, 2005).
   H. Scheffé, *The Analysis of Variance* (John Wiley & Sons, 1961).
   G. A. F. Seber and Alan J. Lee, *Linear Regression Analysis* (John Wiley & Sons, 2003).

   There is a brief discussion of multiple regression in section 3 of chapter 12.
2. K. Pearson and A. Lee, "On the laws of inheritance in man," *Biometrika* vol. II (1903) pp. 357–462. Their table xxii gives the joint distribution, with heights rounded to the nearest inch. We added uniform noise to get continuous data. Due to independent randomization, the data set here differs slightly from the one in our first edition.
3. "Point of averages" is not a standard term.
4. H. N. Newman, F. N. Freeman, and K. J. Holzinger, *Twins: A Study of Heredity and Environment* (University of Chicago Press, 1937). In twin studies, the convention is to plot each twin pair twice; once as $(x, y)$, and once as $(y, x)$.
5. Data from the March 2005 Current Population Survey; a CD-ROM was supplied by the Bureau of the Census. Income is from 2004, and was censored at $150,000: this reduces the mean and the SD, but increases the correlation by a little. Starting in 1992, the Current Population Survey (CPS) reports educational attainment in categories, for instance, "1st–4th grade" or "some college, no degree." See *Monthly Labor Review*, September 1993, pp. 34–38. Single years of education were imputed from grouped data. The correlation between the imputed educational level and the ordered categorical variable used in the CPS is about 0.97.
6. See note 5. "Number of children" is number of own, never-married children under the age of 18. (The data are for women age 25–39, and the correlation depends to some extent on the age range that is used.) Weights must be used here, because the March CPS over-samples large families (note 3 to chapter 3).
7. When the correlation is 0, either slope can be used. "SD line" is not a standard term.
8. However, this formula can be vulnerable to roundoff error.
9. *Consumption Patterns of Household Vehicles 1985*. Residential Transportation Energy Consumption Survey. Energy Information Administration, Washington, D.C.
10. Marjorie Honzik (Institute of Human Development, Berkeley) was kind enough to supply the data.

## Chapter 9. More about Correlation

1. The New York temperatures are measured at JFK; Boston, at Logan Airport. Data are from the Weather Undergound,

   http://www.wunderground.com
2. R. Doll, "Etiology of lung cancer," *Advances in Cancer Research* vol. 3 (1955) pp. 1–50. Report of the U.S. Surgeon General, *Smoking and Health* (Washington, D.C., 1964).
3. The idea goes back to W. S. Robinson, "Ecological correlations and the behavior of individuals," *American Sociological Review* vol. 15 (1950) pp. 351–57. Robinson gives the example of literacy and race, based on 1930 Census data. Our example is a replication; see note 5 to chapter 8 on the data source.

   If each cluster is bivariate normal, with a common regression line, then the slope and intercept can be estimated from the averages. Also see L. Goodman, "Ecological regression and the behavior of individuals," *American Sociological Review* vol. 18 (1953) pp. 663–64. For more discussion, see S. P. Klein and D. A. Freedman, "Ecological regression in voting rights cases," *Chance* vol. 6 (1993) pp. 38–43. Also see D. A. Freedman, "Ecological inference and the ecological fallacy," in N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences* (Elsevier, 2001, vol. 6, pp. 4027–30).
4. E. Durkheim, *Suicide* (Macmillan, 1951, p. 164). We computed the correlation. Durkheim looked at averages of clusters of provinces, for which the correlation was 0.9. His conclusion was "Public instruction and suicide are identically distributed."
5. Multiple regression is some help, but may raise more questions than it answers (section 3, chapter 12). Also see note 11 to chapter 12.

6. For more discussion, see H. Zeisel, *Say It With Figures*, 6th ed. (Harper & Row, 1985, pp. 152ff.)
7. Data supplied by M. Russell from table 1 in D. Jablonski, "Larval ecology and macroevolution in marine invertebrates," *Bulletin of Marine Science* vol. 39 part 2 (1986) pp. 565–87. Also see *Science* vol. 240 (1988) p. 969.
8. References—

   R. Doll and R. Peto, *The Causes of Cancer* (Oxford University Press, 1981).

   B. E. Henderson, R. K. Ross and M. C. Pike, "Toward the primary prevention of cancer," *Science* vol. 254 (1991) pp. 1131–38.

   B. N. Ames, L. S. Gold and W. C. Willett, "The causes and prevention of cancer," *Proceedings of the National Academy of Science U.S.A.* vol. 92 (1995) pp. 5258–65.

   B. S. Hulka and A. T. Stark, "Breast cancer: Cause and prevention," *Lancet* vol. 346 (September 30, 1995) pp. 883–887.

   Figure 8 controls for age, but number of children would seem to be an important confounder (note 14 to chapter 2). Diet in the 1950s and 1960s would be at issue in the figure. There is strong evidence from epidemiology—and animal experiments—to show that over-eating is carcinogenic. The impact of fat (in isocaloric diets) is less clear. Two prospective studies support the ecological analysis: A. Schatzkin et al., "Serum cholesterol and cancer in the NHANES I epidemiologic followup study," *Lancet* ii (1987) pp. 298–301; W. C. Willett et al., "Relation of meat, fat, and fiber intake to the risk of colon cancer in a prospective study among women," *New England Journal of Medicine*, December 13, 1990, pp. 1664–71. But see D. Hunter et al., "Cohort studies of fat intake and the risk of breast cancer—a pooled analysis," *New England Journal of Medicine* vol. 334 (1996) pp. 356–61. Recent experimental evidence contradicts the hypothesis that low-fat diets are protective against cancer.

   A Schatzkin et al., "Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas," *New England Journal of Medicine* vol. 342 (2000) pp. 1149–55.

   R. L. Prentice et al., "Low-fat dietary pattern and risk of invasive breast cancer: The Women's Health Initiative randomized controlled dietary modification trial," *Journal of the American Medical Association* vol. 295 (2006) 629–642.

   S. A. Beresford et al., "Low-fat dietary pattern and risk of colorectal cancer: The Women's Health Initiative randomized controlled dietary modification trial," *Journal of the American Medical Association* vol. 295 (2006) 643–54.
9. National Assessment of Educational Progress, *The Reading Report Card* (ETS/NAEP, Princeton, N.J., 1985, p. 53). There is also a negative correlation with scores on standardized knowledge tests. See Lee R. Jones et al., *The 1990 Science Report Card: NAEP's Assessment of 4th, 8th, and 12th Graders* (U.S. Department of Education, Office of Educational Research and Improvement, Washington, D.C., 1992).
10. T. R. Dawber et al., "Coffee and cardiovascular disease: Observations from the Framingham study," *New England Journal of Medicine* vol. 291 (1974) pp. 871–74.
11. M. P. Rogin and J. L. Shover, *Political Change in California* (Greenwood Press, Westport, Connecticut, 1970, p. xvii).
12. See note 5 to chapter 8.
13. This replicates a study by M. and B. Rodin, "Student evaluations of teachers," *Science* vol. 177 (1972) pp. 1164–66. At the individual level, the correlations would be weaker; however, it is the sign which is interesting. More recent papers include the following—

    L. D. Barnett, "Are teaching questionnaires valid?" *Journal of Collective Negotiations in the Public Sector* vol. 25 (1996) pp. 335–49.

    A. G. Greenwald and J. M. Gillmore, "No pain, no gain? The importance of measuring course workload in student ratings of instruction," *Journal of Educational Psychology* vol. 89 (1997) pp. 743–51.

    M. Scriven, "A unified theory approach to teacher evaluation," *Studies in Educational Evaluation* vol. 21 (1995) pp. 111–29.
14. http://www.collegeboard.com/about/news_info/cbsenior/yr2005/links.html, table 3. In Connecticut, 86% of the seniors took the test. In Iowa, only 5% took the test. The reason: in Iowa and neighboring states, most seniors take the ACT—only those planning to attend elite schools take the SAT. The data are quite non-linear, but Connecticut and Iowa seem close to average, after adjustment for participation rate.

## Chapter 10. Regression

1. These figures are rounded. The exact figures (unweighted):

| | | |
|---|---|---|
| average height = 69.6 inches | SD = 3.19 inches | |
| average weight = 177 pounds | SD = 46.8 pounds | $r = 0.414$ |

The data source is http://www.cdc.gov/nchs/about/major/nhanes/datalink.htm

2. See note 5 to chapter 8.
3. The term "graph of averages" is not standard. In principle, the graph depends on how finely the $x$-values are subdivided.
4. Surprisingly, the "graph of medians" is no more regular.
5. The average height of the fathers was 67.7 inches, with an SD of 2.74 inches; the average height of the sons was 68.7 inches, with an SD of 2.81 inches; $r$ was 0.501. From the original cross-tab, the SDs were 2.72 and 2.75 inches, $r \approx 0.514$; the averages were barely affected by unrounding. See note 2 to chapter 8.
6. D. Kahneman and A. Tversky, "On the psychology of prediction," *Psychological Review* vol. 80 (1973) pp. 237–51.
7. Marjorie Honzik supplied the data. For comparison, in the March 2005 Current Population Survey, the correlation between educational level of husbands and wives was 0.63.

## Chapter 11. The R.M.S. Error for Regression

1. *Statistical Methods for Research Workers* (Oliver and Boyd, 1958, p.182).
2. In multiple regression, the residuals can be plotted against the dependent variable, each independent variable, the fitted values, and omitted variables.
3. There were 60 families where the father was 64 inches tall (to the nearest inch); the sons averaged 66.7 inches in height, with an SD of 2.29 inches. There were 50 families where the father was 72 inches tall (to the nearest inch); the sons averaged 70.7 inches in height, with an SD of 2.30 inches.
4. A "homoscedastic" scatter diagram has more or less the same SD in any narrow vertical strip, but may have a non-linear trend. A "football-shaped" scatter diagram is homoscedastic, and the trend is linear; these diagrams look like sample data from a bivariate normal distribution.
5. In this part of the book, the focus is mainly descriptive. There is a finite data set, and each point is given equal weight. The regression line is viewed as smoothing $E\{Y \mid X\}$ in this finite population. Likewise, $Y$ is predicted from $X$ for a random element in the given population. The $\sqrt{1-r^2}$ formula for prediction error is correct in this setting.

   If a regression line is fitted to a training sample and then used to make forecasts, there is a component of variance around the regression line as discussed above. The sample regression line also has a component of variance around the population line. This latter depends on $X$, of course, and is beyond the scope of this book. With a large training sample, however, the second component of variance may be rather small. Take example 1 in section 5, with a training sample of size 100. The first component of variance (around the regression line) is about 64. For a student who is $z$ SDs away from average on the LSAT, the second component of variance—due to sampling error in the regression line—is about $0.64(1+z^2)$. A fairly extreme case is $z = 3$. The total r.m.s. prediction error is about 8.4; ignoring the second component of variance gives 8.0.

   Of course, there is no free lunch. If you get far enough away from the center of the scatter diagram, the regression estimates become quite untrustworthy. For one thing, linearity may break down. For another thing, sampling error takes its toll. Here is one way to visualize the latter issue. The regression line has to go through the point of averages, and that point is quite solidly anchored—at least with a large-enough sample. However, there is some uncertainty about the slope. The further you go from the point of averages, the more impact the slope has: its uncertainty gets magnified.
6. See note 5 to chapter 8. Summary statistics are unweighted and simplified. Each dot may represent several women. Incomes were censored above at $250,000.
7. Ages censored above 79. In public-use files, to protect respondent confidentiality, the Bureau reports ages 80–84 as 80, and ages 85+ as 85.
8. Station 3 in Lake Mead; 53 monthly geometric averages of available data; year-round; sampling period 1976–86; data provided by the late Jerome Horowitz in connection with a hearing on water quality standards.
9. The data were supplied by Marjorie Honzik.
10. Pitchers are excluded; however, sophomore slump can also be observed on the earned run average for pitchers. In some years, there are two "Rookies of the Year" in the same league. The problem was originally suggested by David Lane. We thank Sam Buttrey and Oren Tversky for expert advice on baseball statistics. Data for the third edition were provided by STATS, Skokie, Illinois. For the fourth edition, we used

    http://www.baseball-almanac.com

    In present format, awards for Rookies of the Year go back to 1949. Summary statistics depend on number of times at bat, and therefore whether averages are weighted by number of at-bats. (The good batters go to bat more frequently, but the effect diminishes as number of at-bats increases.)

There are some minor differences between the two leagues and between years; but on the whole, the results are fairly stable over time.

The following data are for the 1992–1993 seasons. Both leagues are pooled and simple averages are used. There were 588 men who played in both seasons; 438 had at least 25 at-bats in both seasons. The summary statistics for the 438 pairs of batting averages—

1992 average = 241 SD = 55
1993 average = 250 SD = 55
year-to-year correlation = 0.52.

There were 298 players who had at least 100 at-bats in both seasons. The summary statistics—

1992 average = 260 SD = 30
1993 average = 269 SD = 35
year-to-year correlation = 0.26.

The correlation may be attenuated due to restriction of range: many players with 25 to 100 at-bats had batting averages below 200, few players with over 100 at-bats do that poorly. Measurement error plays some role, too. There were 186 players who had at least 250 at-bats in both seasons. The summary statistics—

1992 average = 268 SD = 27
1993 average = 276 SD = 31
year-to-year correlation = 0.40.

11. HANES5 only has categorical data on education (less than high school, high school, more than high school); years are imputed from the Current Population Survey.

## Chapter 12. The Regression Line

1. *Abhandlungen zur Methode der kleinsten Quadrate* (Berlin, 1887, p. 6). We follow the translation by L. Le Cam and J. Neyman, *Bayes-Bernoulli-Laplace Seminar* (Springer-Verlag, 1965, p. viii).
2. See note 5 to chapter 8. Each dot may represent several men. Incomes were censored above at $250,000, and summary statistics were rounded. The exact figures (unweighted):

   average educational level = 12.62 years SD = 3.31 years
   average income = $30,161 SD = $24,007 $r = 0.2713$

3. Data are from the Current Population Survey of March 2005. See note 5 to chapter 8.
4. Returns to education are discussed in—

   Orley Ashenfelter, Colm Harmon, and Hessel Oosterbeek,"A review of estimates of the schooling/earnings relationship, with tests for publication bias," *Journal of Labor Economics* vol. 6 (1999) pp. 453–70.

   David Card, "The causal effect of education on earnings," in Orley Ashenfelter and David Card, editors, *The Handbook of Labor Economics* (Elsevier, Amsterdam, 1999).

   David Card, "Estimating the return to schooling: Progress on some persistent econometric problems," *Econometrica* vol. 69 (2001) pp. 1127–60.

5. Suppose, for instance, that $y = a + bx + cx^2$. If subjects are randomized to different levels of $x$, and a regression line is fitted to the data, the slope does not predict the response of $y$ to changes in $x$, except in some overall, average sense—averaged across subjects and their values of $x$. Of course, it may be possible to estimate the correct functional form from the data. In example 1, the relationship between education and income is non-linear. The value of a college degree relative to a high school degree—measured by the difference in average earnings—is substantially more than the slope would suggest; the relative value of a post-graduate degree is substantially less, probably because the women with advanced degrees are just starting out on their careers. For comparing high school and middle school, the slope is fine.
6. This equation is based on rounded values of the summary statistics supplied by IRRI.
7. Paraphrase of testimony by Franklin Fisher (MIT) in *Cuomo v. Baldrige* (80 civ. 4550 S.D.N.Y. 1987) on regression models for adjusting the Census, transcript pp. 2149ff.
8. Carried out by the late William Fretter, former professor of physics, University of California, as a demonstration in his elementary course; details are simplified.
9. Regression is appropriate here, in Berkson's case of the errors-in-variable model. The nominal values of the weights are fixed by the investigator, the actual value is subject to error; it is the nominal value which goes into the regression. When the value of the weight is measured subject to error, and the measured value goes into the regression, then the usual regression estimates are biased. A reference is G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 6th ed. (Iowa State University Press, 1973). In effect, we are summarizing the data in table 1 by the two averages,

the two SDs, and $r$. A more natural summary would be based on

(i) the slope, intercept, and r.m.s. error of the regression line, and
(ii) the average and SD of the weights.

The statistics in (i) would be more or less invariant (up to sampling error) across experiments with different weights. Perhaps that is why many statisticians find it more natural to begin with the regression line and do correlation later. From our perspective, however, there are not so many examples with the kind of invariance shown by Hooke's law. With examples of a different texture—say, educational levels of husbands and wives—the correlation coefficient seems to offer the more natural summary. Furthermore, starting from the line makes it very difficult (at least in our experience) to explain $r$. That is why we start with $r$—and hope that devotees of the other approach will bear with us.

10. The sample size is 1,036, so the slope is real, not a fluke of sampling. (Cases with missing data on income or education are excluded; summary statistics are rounded.) Also see T. W. Teasdale et al., "Fall in association of height with intelligence and educational level," *British Medical Journal* vol. 298 (1989) pp. 1292–93. In data from HANES3 (1988–91), the slope of height on education for men age 25–34 was about 0.20 inches per year of schooling completed; there were 763 sample persons. See note 4 to chapter 4 for data sources.
11. For more discussion of regression models in the social sciences, see the Summer 1987 issue of *Journal of Educational Statistics*; *Sociological Methodology 1991*; *Foundations of Science* vol. 1, no. 1 (1995). Also see D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, 2005).
12. See note 4 to chapter 4. HANES5 reports incomes by category only; we used the correlation and summary statistics on heights from HANES5, summary statistics on income from the Current Population Survey of March 2005.
13. HANES3 had detailed questions on diet, including items on pizza and beer. The statistics in the exercise were about right for the U.S. population age 18–24. See note 4 to chapter 4 for data sources.
14. For a review of the literature, see S. J. Pocock, M. Smith and P. Baghurst, "Environmental lead and children's intelligence: a systematic review of the epidemiological evidence," *British Medical Journal* vol. 309 (1994) pp. 1189–97.
15. "Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion," *British Medical Journal* vol. 297 (1988) pp. 319–28. The first analysis was a multiple regression, with $52 \times 200$ subjects pooled across centers; the second had separate regressions within each center. For commentary, see—

    D. A. Freedman and D. B. Petitti, "Salt and blood pressure: Conventional wisdom reconsidered," *Evaluation Review* vol. 25 (2001) pp. 267–87.
    N. Graudal and A. Galløe, "Should dietary salt restriction be a basic component of antihypertensive therapy? *Cardiovascular Drugs and Therapy* vol. 14 (2000) pp. 381–6.
    G. Taubes, "The (political) science of salt," *Science* vol. 281 (1998) pp. 898–907.

# Part IV. Probability

## Chapter 13. What Are the Chances?

1. For other views of chance, see—

    R. A. Fisher, *Statistical Methods and Scientific Inference*, 13th ed., reprinted by Oxford University Press, 1993, in J. H. Bennett, editor, *Statistical Methods, Experimental Design and Scientific Inference*.
    L. J. Savage, *Foundations of Statistics*, 2nd ed. (Dover, 1972).

    For discussion, see *Foundations of Science*, vol. 1 (1995) no. 1.
2. The 3rd edition was published in 1756, after de Moivre's death. It has been reprinted by Chelsea Publishing, New York, 1967.
3. *Statistical Abstract*, 2003, table 11.
4. W. Fairley and F. Mosteller, "A conversation about Collins," *University of Chicago Law Review* (1974).
5. The prosecutor calculated two "chances" for two "events," slipping back and forth between them. The first event was that the accused were guilty. The second event was that no other couple in Los Angeles matched the description. For a frequentist, the concept of chance does not apply so well. Even a Bayesian might find some difficulty here, because there is no reasonable chance model to connect the data with the hypothesis of guilt or innocence. (Also see note 6.)

    Were there other couples in Los Angeles matching the description? In principle, this might seem like a statistical issue, which could be settled by taking a sample. However, a calculation

will show that sampling the couples in the city does not settle the issue with any reasonable level of confidence: a complete census is needed.

6. The "characteristics" of DNA used in matching are the variable number of tandem repeats (VNTRs) between loci on non-coding segments of DNA. References—

   *Jurimetrics*, vol. 34, no. 1 (1993).

   National Academy of Sciences/National Research Council, *DNA Technology in Forensic Science* (Washington, D.C., 1992).

   National Academy of Sciences/National Research Council, *DNA Forensic Science: An Update* (Washington, D.C., 1996).

   Federal Judicial Center, *Reference Manual on Scientific Evidence*, 2nd ed. (Washington, D.C., 2000).

   The "prosecutor's fallacy" consists in confusing the rate at which defendant's DNA occurs in the population (however well or poorly that may be estimated) with the probability that defendant is innocent; more generally—at least from a Bayesian perspective—of confusing

   $$P\{\text{evidence} \mid \text{innocence}\} \quad \text{with} \quad P\{\text{innocence} \mid \text{evidence}\}.$$

   See W. C. Thompson and E. L. Schumann, "Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy," *Law and Human Behavior* vol. 11 (1987) pp. 167–87.

7. This exercise was suggested by D. Kahneman and A. Tversky, "Judgment under uncertainty: heuristics and bias," *Science* vol. 185 (1974) pp. 1124–31. Also see D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases* (Cambridge University Press, 1982).

## Chapter 14. More about Chance

1. From the dedication to the *Doctrine of Chances*.
2. Of course, $P(A \text{ or } B) = P(A) + P(B)$ if $P(A \text{ and } B) = 0$. More generally,

   $$P(A \text{ or } B) = P(A) + P(B \text{ but not } A),$$

   which is analogous to the multiplication rule for dependent events.
3. For a more historical account of the correspondence between Pascal and Fermat, see pp. 88–89 of F. N. David, *Games, Gods and Gambling* (Buckinghamshire, England: Charles Griffin & Co., 1962). Sandrine Dudoit (U.C. Berkeley) gave assistance exceptionelle on *franglais*.

## Chapter 15. The Binomial Coefficients

1. Apparently, Jia Xian discovered the binomial formula a little earlier—in the eleventh century. See Li Yan and Du Shiran, *Chinese Mathematics* (Oxford University Press, 1987, pp. 121 ff). Pascal's triangle is also called, perhaps more justly, Yang Hui's triangle.
2. The model is only approximate: there is a slightly better than even chance for a newborn to be male, and successive births in the family are slightly dependent.
3. This exercise was suggested by D. Kahneman and A. Tversky, "Judgment under uncertainty: heuristics and bias," *Science* vol. 185 (1974) pp. 1124–31.
4. On the Finnish twin study, see J. Kaprio and M. Koskenvuo, "Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs," *Social Science and Medicine* vol. 29 (1989) pp. 1083–89. For references on the health effects of smoking, see note 1 to chapter 2. Some researchers in artificial intelligence remain skeptical about the evidence: P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction, and Search*, 2nd ed. (MIT Press, 2000).

   Data in the table are for current smokers; males and females are pooled. Kaprio and Koskenvuo considered a number of potential confounders. On the following, they saw little difference between smokers and non-smokers: alcohol, blood pressure, cholesterol, diabetes, coffee consumption, education, occupation, marital status, Eysenck personality inventory. Smokers exercised less, which may increase the risk of coronary heart disease; however, they were also thinner, which may reduce their risk. For U.S. data, see D. Carmelli and W. F. Page, "Twenty-four year mortality in World War II U.S. male veteran twins discordant for cigarette smoking," *International Journal of Epidemiology* vol. 25 (1996) pp. 554–59.
5. *Statistical Abstract*, 1988, table 268; 1994, table 305; 2003, table 309.
6. *San Francisco Chronicle*, December 9, 1975. The research report itself is much more sober: D. J. Ullyot et al., "Improved survival after coronary artery surgery in patients with extensive coronary artery disease," *Journal of Thoracic and Cardiovascular Surgery* vol. 70 (1975) pp. 405–13.

7. *Bouman v. Block*, 940 F.2d 1211 (9th Cir. 1991).
8. For national data, see J. H. Pryor et al., *The American Freshman: National Norms for Fall 2005* (Higher Education Research Institute, UCLA, 2006).
9. This exercise was suggested by *Economic Report of the President 1974*, pp. 147 ff.
10. T. W. Teasdale et al., "Degree of myopia in relation to intelligence and educational level," *Lancet* (December 10, 1988) pp. 1351–54.

# Part V. Chance Variability

## Chapter 16. The Law of Averages

1. *An Experimental Introduction to the Theory of Probability* (University of Witwatersrand Press, 1964). Kerrich went to teach in South Africa after World War II.
2. We distinguish between the difference as a number (in "absolute terms") and the difference as a percent. Absolute values also come in at a more technical level. Let $X_n$ be the chance error after $n$ tosses, that is, the number of heads minus half the number of tosses. Then $X_n$ is a martingale, so $E\{X_{n+m}|X_n\} = X_n$; but $E\{|X_{n+m}||X_n\} > |X_n|$.
3. "Chance process" is used in a non-technical sense. A "number generated by a chance process" is the observed value of a random variable.
4. Computer programs are deterministic, and therefore cannot generate numbers in a truly random way. However, a program can generate a sequence of numbers which look quite random. One method involves a multiplier $M$, which is a very big number. A "seed" $x$ is chosen by the programmer: $x$ is between 0 and 1. The computer works out $M$ times $x$, which has an integer part and a decimal part:

$$\ldots.\ aaaaaaaaaaaa\,.\,bbbbbbbbbbbbb\ \ldots.$$

Digits to the left of the decimal point are printed out as the first random number, and the decimal part is used as the seed for the next random number. For more discussion, see

Jerry Banks, editor, *Handbook of Simulation* (Wiley, 1998).
P. L'Ecuyer, "Efficient and portable combined random number generators," *Communications of the ACM* vol. 31 (1988) pp. 742–74.
J. E. Gentle, W. Haerdle, and Y. Mori, editors, *Handbook of Computational Statistics* (Springer-Verlag, 2004).
D. Knuth, *The Art of Computer Programming* vol. II (Addison–Wesley, 1998).
P. A. W. Lewis and E. J. Orav, *Simulation Methodology for Statisticians, Operations Analysts, and Engineers* (Wadsworth & Brooks/Cole, 1988, chapter 5).
G. Marsaglia, "Random numbers fall mainly in the planes," *Proceedings of the National Academy of Sciences* vol. 60 (1968) pp. 25–28.
———, "A current view of random number generators," *Proceedings of the Sixteenth Symposium on the Interface between Computer Science and Statistics* (1985) pp. 3–10.

5. "Sum of draws from a box" is not a standard term but it is lighter than "sum of independent, identically distributed, random variables." "Box model" is not standard either, although it seems to be catching on.
6. Let $S_N$ be binomial with $N$ trials and success probability $p$. We claim that $P\{S_{2n+2} > n+1\} > P\{S_{2n} > n\}$ for all $p$ with $1/2 \le p < 1$. Indeed, let $f_{Nm}(p) = P\{S_N > m\}$. Then

$$f'(p) = NP\{S_{N-1} = m\}. \tag{1}$$

Let $g_n(p) = P\{S_{2n+2} > n+1\} - P\{S_{2n} > n\}$. By (1),

$$\frac{d}{dp}g_n(p) = \frac{(2n)!}{n!\,(n-1)!}\left[\frac{4n+2}{n}p(1-p) - 1\right]p^n(1-p)^{n-1}. \tag{2}$$

Now $(4n+2)/n > 4$. There is a $p_0$ with $1/2 < p < p_0$ such that $g_n'(p)$ is positive for $1/2 \le p < p_0$, negative for $p_0 < p < 1$, and 0 for $p = p_0$. Thus, $g_n$ increases between $1/2$ and $p_0$, then decreases between $p_0$ and 1. Suppose $n > 1$. Then $g_n(1) = 0$, and it suffices to check the claim for $p = 1/2$. However,

$$f_{2n,n}\left(\frac{1}{2}\right) = \frac{1}{2} - P\left\{S_{2n} = n \middle| p = \frac{1}{2}\right\}$$

is strictly increasing with $n$. If $n = 1$, the claim reduces to showing that $3p^2 - 4p + 1 < 0$ for $1/2 \le p < 1$, which is easily verified.

## Chapter 17. The Expected Value and Standard Error

1. Keno is the Las Vegas equivalent of bingo. There are 80 balls, numbered 1 through 80. On each play, 20 balls are chosen at random without replacement. If you bet on the single number 17, for example, you are betting that ball number 17 will be among the 20 that are chosen. Your chance of winning is $20/80 = 1/4$.
2. If $X_i$ are independent and identically distributed with mean $\mu$ and variance $\sigma^2$, then
$$E(X_1 + \cdots + X_n) = n\mu$$
and
$$\text{var}\,(X_1 + \cdots + X_n) = \text{var}\,X_1 + \cdots + \text{var}\,X_n = n\sigma^2.$$
The SE, which is the square root of the variance, is then $\sqrt{n}\,\sigma$. That is the square root law.
3. In this book, we use SD for data and SE for chance quantities (random variables). This distinction is not standard, and the term SD is often used in both situations.
4. In parts II and III, we used standard units for data, centering on the average and scaling by the SD. Here, we make the transition to random variables, centering on the expected value and scaling by the SE.
5. We are using "estimate" in its ordinary sense, of approximation. Statisticians also use "estimate" in a more technical sense, to be taken up in part VI.
6. Consider 11 cells labelled $-5, -4, \ldots, +4, +5$, and 10 balls. There are altogether
$$\binom{20}{10} = 184{,}756$$
ways to distribute the 10 balls into the 11 cells; each distribution defines a "box" with 10 tickets: for instance, if you put 4 balls into the cell "$-1$" and another 6 into the cell "3," you get the box $\boxed{4\ \boxed{-1}\text{'s}\ 6\ \boxed{3}\text{'s}}$. We wrote a computer program to check all 184,756 boxes; 5,448 of them have mean 0, but one of these consists of 10 $\boxed{0}$'s. The program checked the remaining possibilities to verify the intuitively obvious result. On the combinatorics, see section II.4 of W. Feller, *An Introduction to Probability Theory and its Applications*, vol. I, 3rd ed. (John Wiley & Sons, 1968). Also see note 6 to chapter 16. For an example with non-monotone behavior, consider the sum of 100 or 200 draws from the box
$$\boxed{5\ \boxed{-16}\text{'s}\ 5\ \boxed{16}\text{'s}}$$
Let $S_n$ be the sum of $n$ draws from the box, which is necessarily a multiple of 32. Thus, if $-15 \le S_{100} \le 15$ then $S_{100} = 0$; if $-30 \le S_{200} \le 30$ then $S_{200} = 0$. And $P\{S_{100} = 0\} > P\{S_{200} = 0\}$.
7. The idea of using "big" and "small" to label the values is due to the statistics group at Southern Methodist University.
8. E. O. Thorp, *Beat the Dealer* (Random House, 1966). Some side bets at baccarat also have positive expected values. So do some bets on the favorites (curiously) at race tracks. See R. T. Thaler and W. T. Ziemba, "Parimutuel betting markets: Racetracks and lotteries," *Journal of Economic Perspectives* vol. 2 (1988) pp. 161–74.

## Chapter 18. The Normal Approximation for Probability Histograms

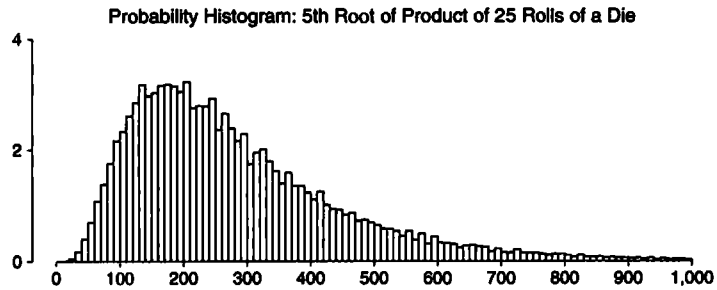
1. The Hewlett Packard HP 15C.
2. A mathematical discussion can be found in Chapter 7 of W. Feller, *An Introduction to Probability Theory and its Applications* vol. I, 3rd ed. (John Wiley & Sons, 1968).
3. If the number of draws is very large, it may be helpful to group ranges of values together, as in figure 10 (for products). Likewise if the tickets do not show whole numbers. Even if the number of draws is moderate, and there are whole numbers on the tickets, it may be helpful to use wider class intervals: this is so when the differences of the numbers on the tickets have a common divisor bigger than 1. For example, suppose you bet \$1 on the toss of a coin, 100 times. Your net gain is like the sum of 100 draws made at random with replacement from the box $\boxed{\boxed{-1}\boxed{+1}}$. The possible values for the net gain are even: $0, \pm 2, \pm 4, \ldots$ The histogram can be drawn with rectangles of width 1 centered at these values. However, with the method of section 4, rectangles of width 2 give better results. The problems in this book do not raise this sort of issue.
4. Computed on the HP 15C.
5. The continuity correction can be used when the tickets in the box are integers whose differences have no common divisor (except 1). This is called "aperiodicity." If the numbers have a common divisor bigger than 1, or the tickets have values other than integers, don't use the continuity correction without further thought. Also see notes 3 and 9.

6. A mathematical analysis of the skewness is provided by the Edgeworth expansion. See Chapter 16 in W. Feller, *An Introduction to Probability Theory and its Applications* vol. II, 2nd ed. (John Wiley & Sons, 1970).
7. The waves can be explained as follows. If the box were $\boxed{1}\boxed{1}\boxed{9}$, the possible values for the sum would be 25, 33, 41, ... separated by gaps of 8. If the box were $\boxed{2}\boxed{2}\boxed{9}$, the possible values for the sum would be 50, 57, 64, ... separated by gaps of 7. The box in figure 9 is intermediate between these two, and the peak-to-peak distance alternates between 7 and 8. Another way to look at it: the peaks reflect the distribution of the number of 9's among the 25 draws.
8. The shape of the histograms in figure 10 may be a little surprising. However, if $X_1, X_2, \ldots$ are the successive rolls of the die, then it is

$$(X_1 X_2 \cdots X_n)^{1/\sqrt{n}}$$

which is approximately log normal after centering. A probability histogram for the 5th root of the product of 25 rolls is shown below, and it has the right shape. The probabilities were computed using a combinatorial algorithm, and the wiggles are real. (The product of 25 rolls of a die has the form $2^a 3^b 5^c$ for non-negative integers $a, b, c$, lending itself to gaps and wiggles.)

The logarithm (base 10) of the product of 25 rolls is the sum of 25 logarithms. Each has mean 0.4762 and SD 0.2627, so the sum of 25 logs has expected value $25 \times 0.4762 \approx 11.91$ and standard error $\sqrt{25} \times 0.2627 \approx 1.31$. The sum of 25 logs is already quite close to normally distributed. Take the bottom panel in figure 10, for the product of 25 rolls. The axis cuts off at $10^{13}$, which is 13 on the log scale, or 0.83 in standard units. About 20% of the probability is to the right of this value. The width of each rectangle in the histogram is $10^{11}$. The first rectangle covers the interval from $-\infty$ to 11 on the log scale, which in standard units is $(-\infty, -0.69)$. This interval contains about 25% of the probability!




9. The tacit assumptions: nonzero SD, and a finite number of tickets in the box with integer values. Suppose for simplicity that the numbers in the box are aperiodic; let $\mu$ be their mean and $\sigma$ their SD. Let $h_n(x)$ be the histogram for the sum of $n$ draws, plotted by our convention: each rectangle has width 1, centered on a possible value. Let $\phi(z)$ be the standard normal density. Then

$$\sigma\sqrt{n}\, h_n(n\mu + \sigma\sqrt{n}z) \to \phi(z).$$

The "histogram in standard units" euphemizes this change of scale. See W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, 2nd ed. (John Wiley & Sons, 1971, pp. 517, 540).
10. Suppose the tickets in two boxes have the same average, and average absolute deviation from average. If they also have the same SD, the asymptotic behavior of the sums will be the same. If not, not. An example would be

A) $\boxed{-1}\boxed{1}$ B) $\boxed{-2}\boxed{0}\boxed{0}\boxed{2}$

In both boxes, the tickets average out to 0, and the average absolute deviation from average is 1. But the SD for box A is 1, while the SD for box B is about 1.4. Consequently, the sum of 100 draws from box B is about 1.4 times as spread out (by any reasonable measure of spread) than the sum of 100 draws from box A. It is the average and SD of the numbers in the box which control the asymptotic distribution of the sum: other measures of location and spread do not.
11. Let $n$ denote the number of draws, and $k$ the number of repetitions. The implicit condition is that $k/\sqrt{n}\log n \to \infty$. See D. A. Freedman, "A central limit theorem for empirical histograms," *Zeitschrift für Wahrscheinlichkeitstheorie* vol. 41 (1977) pp. 1–11.

# Part VI. Sampling

## Chapter 19. Sample Surveys

1. The chapter opening quote is from *The Adventure of the Copper Beeches*. We found it in Don McNeil, *Interactive Data Analysis* (John Wiley & Sons, 1977).
2. References on sampling

   LESS TECHNICAL

   N. M. Bradburn and S. Sudman, *Polls and Surveys* (Jossey-Bass Inc., 1988).
   A. Campbell, G. Gurin and W. Miller, *The Voter Decides* (Row-Peterson, Evanston, 1954).
   Jean M. Converse, *Survey Research in the United States: Roots and Emergence, 1890–1960* (University of California Press, 1987).
   Shari Seidman Diamond, Reference Guide on Survey Research, in *Reference Manual on Scientific Evidence*, 2nd ed. (Federal Judicial Center, Washington, D.C., 2000).
   George Gallup, *The Sophisticated Poll Watcher's Guide* (Princeton Opinion Press, 1972).
   Herbert Hyman et al., *Interviewing in Social Research* (University of Chicago Press, 1954).
   Frederick Mosteller et al., *The Pre-Election Polls of 1948* (Social Science Research Council, New York, 1949).
   Mildred Parten, *Surveys, Polls and Samples* (Harper & Row, 1950).
   F. F. Stephan and P. J. McCarthy, *Sampling Opinions* (John Wiley & Sons, 1958).
   Hans Zeisel and David Kaye, *Prove It with Figures* (Springer, 1997).

   MORE TECHNICAL

   W. G. Cochran, *Sampling Techniques*, 3rd ed. (John Wiley & Sons, 1977).
   Robert M. Groves et al., *Survey Methodology* (Wiley-Interscience, 2004).
   M. H. Hansen, W. N. Hurwitz and W. G. Madow, *Sample Survey Methods and Theory* (John Wiley & Sons, 1953).
   Leslie Kish, *Statistical Design for Research* (John Wiley & Sons, 1987).
   Seymour Sudman, *Applied Sampling* (Academic Press, 1976).
3. All quotes are from the *New York Times* (Oct. 1–15, 1936).
4. For help with the first and second editions, we thank Diane Colasanto, Laura Kalb, Jack Ludwig, Coleen McMurray, and Paul Perry of the Gallup Poll. Figure 3 was typeset from copy provided by the Gallup organization, and reproduced with their kind permission. For the third edition, we thank David Moore, Kim Neighbor, and Lydia Saad; the description of the 1992 survey is based on conversations with them. For the fourth edition, we again thank Lydia Saad.
5. See Parten, p. 393, and Stephan and McCarthy, pp. 241–70 (note 2).
6. For another discussion, see M. C. Bryson, "The *Literary Digest* poll: Making of a statistical myth," *American Statistician* vol. 30 (1976) pp. 184–85. Bryson agrees that the *Digest* poll was spoiled by non-response bias. However, he discounts selection bias as a problem, and questions whether the *Digest* really drew on phone books for its mailing list (that is, the list of people to be polled). Our primary source of information about the *Digest* poll was George Gallup—a shrewd, interested, and first-hand observer. He maintained that the *Digest* used phone books, lists of automobile owners, and its own subscription list as the source for the mailing list. This account is confirmed, at least in essentials, by others like Parten, or Stephan and McCarthy (see note 2).

   The *Digest* did not publish any very full account of its procedures that we could find. However, something can be learned by reviewing the issues of the *Digest* for the period August 22 through November 14, 1936. For instance,

   > The Poll represents thirty years' constant evolution and perfection. Based on "commercial sampling" methods used for more than a century by publishing houses to push book sales, the present mailing list is drawn from *every telephone book in the United States*, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail order and occupational data. [Aug. 22, p. 3, our italics.]

   The article goes on to explain that the list was put together for the 1924 election, but was subsequently revised by "trained experts." Most of the names on the list were held over from year to year, and the list was used for polls between elections (Aug. 29, p. 6). Drawing on lists of registered voters seems to have been an innovation for 1936, and such lists were used only for certain "big cities" (Oct. 17, p. 7). Clearly, the *Digest* expected to get the percentages right (Aug. 22, p. 3):

   > Once again, THE DIGEST was asking more than ten million voters—one out of four, representing every county in the United States—to settle November's election in October. Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be *triple-checked*, verified, *five times* cross-classified and totaled. When the last figure has

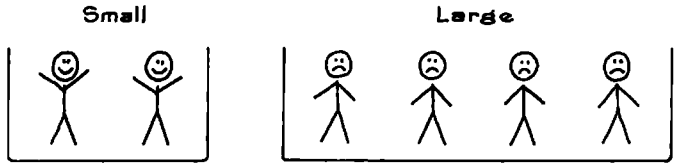
been totted and checked, if past experience is a criterion, the country will know *to within a fraction of 1 percent* the actual popular vote of forty millions. [Their italics.]

The *Digest* was off by 19 percentage points. Why? By modern standards, the *Digest*'s mailing list was put together in a somewhat arbitrary way, and it was biased: it excluded substantial, identifiable portions of the community. Bryson suggests that if the *Digest* had somehow managed to get 100% response from its list of 10 million names, it would have been able to predict the election results. This seems unlikely. As we say in the text, there were two main reasons for the *Digest*'s error: selection bias and non-response bias.

7. This 65% is typical of four-call probability samples in the late 1980s. The response rate declined from about 75% in 1975, and 85% in 1960. This decline is a major worry for polling organizations. In 2005, the best face-to-face research surveys in the U.S., interviewing a randomly-selected adult in a household, get response rates over 80%. Response rates for the Current Population Survey—around 95%—are discussed in chapter 22.
8. This section draws on the book by Mosteller et al. (note 2).
9. Stephan and McCarthy, p. 286 (note 2).
10. It is tempting to confuse quota sampling with stratified sampling, but the two are different. Suppose, for instance, that it is desired to draw a sample of size 200 from a certain town, controlling for sex; in fact, making the number of men equal to the number of women. A quota sampler could in principle hire two interviewers, one to interview 100 men, the other to interview 100 women. In other respects, the two interviewers would pick whomever they wanted. This is not such a good design. By contrast, a stratified sample would be drawn as follows:
    - Take a simple random sample of 100 men.
    - Independently, take a simple random sample of 100 women.

    This is a better design, because human bias is ruled out.
11. The list of units to be sampled is the "sampling frame," and the first step in taking a probability sample is drawing up the sampling frame. This can be quite difficult, and there is often some degree of mismatch between the frame and the population. With area samples, the frame is a list of geographic units.
12. Details of such designs are discussed in chapter 22. We suggest that stratification is needed to draw the sample in a way that keeps the costs reasonable, but in many polls the stratification does little to reduce sampling error. To take a hypothetical example, suppose a country consisted of two regions, East and West. In the East, 60% of the voters are Democrats; in the West, only 40% are. East and West are equal in size, so the overall percentage of Democrats is 50%. Now, two survey organizations take samples to estimate the overall percentage of Democrats. The first one uses a simple random sample of size $n$. The standard error is $50\%/\sqrt{n}$. The second one stratifies, taking a simple random sample of size $n/2$ in the East, and an independent simple random sample of size $n/2$ in the West. The standard error is $\sqrt{0.4 \times 0.6} \times 100\%/\sqrt{n}$. Since $\sqrt{0.4 \times 0.6} \approx 0.49$, the reduction in SE is minimal. Furthermore, in this artificial example, the difference between the regions is much larger than the difference observed in real elections. So the advantage of stratification in predicting real elections is even less. (By contrast, when sampling economic units like companies or establishments, stratification can really help to reduce variance; also see note 5 to chapter 20.)
13. The Gallup Poll uses variants of random-start list sampling. In the first 3 stages, probability is proportional to size; in effect, each unit appears on the list with multiplicity equal to its size. Within each of the four geographic regions, there is a stratum of rural areas, which is handled somewhat differently from the urban areas.
14. The Gallup organization explains "This method of selection within the household has been developed empirically to produce an age distribution by men and women separately which compares closely with the age distribution of the population."
15. Strictly speaking, for the Gallup Poll it is possible to compute sampling probabilities only for households, not for individuals—due to the rule used in selecting individuals within households. Non-response is another complication. We thank Ben King (Florida) for useful discussions on this point. Often, probability methods are designed so that each individual in the population will get into the sample with an equal chance, so the sample is "self-weighting." However, the Gallup poll interviews only one person in each household selected for the survey. This discriminates against people who live in large households; not enough of them are represented in the sample. (See sketch at top of next page.) An adjustment is made to correct for this bias, by giving more weight to the people from large households who do get into the sample. Household size is obtained from question 18, figure 3, p. 347.
16. Paul Perry, "A comparison of the voting preferences of likely voters and likely nonvoters," *Public Opinion Quarterly* vol. 37 (1973) pp. 99–109. Who has voted is a matter of public record; how they voted, of course, is not.

Household bias. Imagine selecting one of the two households below at random: then select a person at random from the selected household. This produces a sample of size one. A person in the small household has a better chance of getting into the sample than a person in the large household.




17. The Gallup Poll "secret ballot" is not secret; ballots are connected to questionnaires.
18. After 1992, the Gallup Poll changed the design. They stratified the sample by four census regions. Within each region, they chose a random sample of residential telephone banks, and dialed random numbers within sampled banks.
19. In 2005, for a good commercial telephone survey, about 1/3 of the telephone numbers dialed do not answer. If someone answers the phone, about 2/3 hang up rather quickly. However, if the interviewer gets through to a person, and engages them for a minute or two, the completion rate is around 95%.
20. L. Belmont and F. Marolla, "Birth-order, family-size, and intelligence," *Science* vol. 182 (1973) pp. 1096–1101. On the average, intelligence decreases with birth order and family size, even after controlling for family background. Also see R. B. Zajonc, "Family configuration and intelligence," *Science* vol. 192 (1976) pp. 227–36. However, the association may be to due to residual confounding by social class. See J. L. Rodgers, H. H. Cleveland, E. van den Oord, and D. C. Rowe, "Resolving the debate over birth order, family size, and intelligence," *American Psychologist* vol. 55 (2000) pp. 599—612. The Belmont-Marolla study is discussed again in exercise 40 on p. 575.
21. Kenneth Stampp, Professor Emeritus of History, University of California, Berkeley. This was a WPA project, and the subjects must have been in their seventies!
22. R. W. Fogel and S. L. Engerman, *Time on the Cross* (New York: W. W. Norton & Company, 1989, p. 39); *Evidence and Methods* (Little, Brown & Company, 1974, p. 37). A careful critique is by Richard Sutch, "The treatment received by American slaves," *Explorations in Economic History* vol. 12 (1975) pp. 335–438.
23. L. L. Bairds, *The Graduates* (ETS, Princeton, N.J., 1973).
24. Discussion by A. L. Cochrane in The Medical Research Council, *The Application of Scientific Methods to Industrial and Service Medicine* (HMSO, London, 1951, pp. 36–39).
25. A. C. Nielsen, *1987 Annual Report on Television*; *New York Times*, March 10, 1997, p. C1.
26. The story was published on September 11, 1988. The source was Raymond A. Eve and Dana Dunn, "Psychic powers, astrology and creationism in the classroom," *American Biology Teacher* vol. 52 (1990) pp. 10–21. The investigators got 190 responses out of their sample of 387 drawn from the list of 20,000 names, which in turn was a systematic sample from the National Register of High School Life Science and Biology Teachers. This is a good study which merits attention. Unfortunately, in the first few printings of the second edition, we relied on the newspaper description, which omitted crucial details about the sample; we drew the wrong conclusion about non-response bias.
27. Based on an example in Parten's book (note 2).
28. From *Time on the Cross* (note 22). Anne Arundel was the wife of the second Lord Proprietary of Maryland, Cecil Calvert. The two main slave auction houses of the time were at Annapolis (Arundel County) and Charleston (South Carolina). We thank Sharon Tucker for the Maryland history.
29. E. K. Strong, *Japanese in California* (Stanford University Press, 1933).
30. *San Francisco Chronicle*, December 10, 1987; letter by Stephen Peroutka to *New England Journal of Medicine* vol. 317 (1987) pp. 1542–43.
31. This example was suggested by D. Kahneman and A. Tversky, "Judgment under uncertainty: heuristics and bias," *Science* vol. 185 (1974) pp. 1124–31.

## Chapter 20. Chance Errors in Sampling

1. The example is loosely suggested by followup studies on Cycle I of the Health Examination Survey, a probability sample of persons age 18–79 in the U.S. The study was done in 1960–61. The sample size was 6,672, of whom 3,091 were male.
2. Sir Arthur Conan Doyle, *The Sign of Four* (J. B. Lippincott, 1899; Ballantine Books, 1974, p. 91). Holmes attributes the thought to Winwood Reade.

3. The histograms in figure 3, like the calculations in example 2, are based on sampling with replacement. In this example—with a sample of 400 from a population of 100,000—there is little difference between sampling with or without replacement. Details are in the next section. The vertical axis is drawn in percent per standard unit.
4. Data for the whole U.S. are available from *Statistical Abstract*, 2003: table 63 gives marital status; table 229, educational level for age 25+; table 693, personal income; and tables 490ff, income tax returns.
5. The issues may be different in other contexts. For instance, suppose you are sampling from two different strata, and want to allocate a fixed number of sampling units between the two. If the object is to equalize accuracy of the two estimated percentages, a reasonable first cut is to use equal sample sizes. If the object is to equalize accuracy of estimated numbers, or to estimate a percentage that is pooled across the strata, a larger sample should generally be drawn from the larger stratum. Gains in accuracy from stratification—as opposed to simple random sampling—should not be overestimated (note 12 to chapter 19).
6. Voting-age population by state comes from *Statistical Abstract*, 2006, table 408; election results by state from table 388. The population for NM was closer to 1.4 million; for TX, 16 million.

## Chapter 21. The Accuracy of Percentages

1. Sir Arthur Conan Doyle, *A Study in Scarlet* (J. B. Lippincott, 1893; Ballantine Books, 1975, p. 136).
2. The technique described in the text is a special case of what statisticians now call the "bootstrap" method for estimating standard errors. See Brad Efron and Rob Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, 1993). For a cautionary note, see L. D. Brown, T. T. Cai, and A. DasGupta, "Confidence intervals for a binomial proportion and asymptotic expansions,' *Annals of Statistics* vol. 30 (2002) pp. 160–201.
3. For college enrollments in the whole U.S., see *Statistical Abstract*, 2003, tables 278ff.
4. Family income data by educational level of head of household, for the whole U.S., is reported in *Statistical Abstract*, 2003, table 689.
5. *Statistical Abstract*, 2003, table 744. Exercises 4 and 5 more or less match these data.
6. Suppose we draw at random with replacement. As the sample size $n \to \infty$,
$$P\{\hat{p} - k\,\mathrm{SE} < p < \hat{p} + k\,\mathrm{SE}\}$$
tends to the area under the normal curve between $-k$ and $k$; this is less than 1.
7. This book takes a strict frequentist line. Other views are cited in note 1 to chapter 13. Many colleagues will feel that we shut our eyes and walked across an intellectual minefield in this section. We hope they will be charitable in their judgment of the results.
8. This picture was proposed by Juan Ludlow, CIMASS/UNAM, Mexico.
9. The standard errors applicable to simple random samples are often computed for samples of convenience. In some contexts, the results may be useful. For example, the objective may be to show that the sampling procedure at issue is quite different from simple random sampling (as in exercise 26 on p. 435).
10. *Statistical Abstract*, 2003, table 977. For national data on automobile ownership, see table 976.
11. D. Ravitch and C. E. Finn, Jr., *What Do Our 17-Year-Olds Know?* (Harper & Row, 1987). The sample was highly designed.
12. For national data, see *Statistical Abstract*, 2003, tables 1125–27.
13. The rules of Keno are explained in note 1 to chapter 17. The chance for a single number is 20/80, because there are 20 draws. The chance for a double number is $(20 \times 19)/(80 \times 79)$.

## Chapter 22. Measuring Employment and Unemployment

1. We are grateful to many people at Census Bureau for their help with previous editions, including Sherry Courtland, Charles Jones (deceased), Donna Kostanich, Marty Riche, and Jay Waite. For the 4th edition, we thank Louis Kincannon, Greg Weyland, and Cindy Taeuber.

   The Bureau of the Census is responsible for the sample design, collection, and production of data, as well as calculation of the estimates and their standard errors. The Bureau of Labor Statistics does the seasonal adjustments, and is responsible for the publication and economic interpretation of the results. Some useful references on the Current Population Survey—

   http://www.bls.gov/cps
   http://www.census.gov/cps
   Bureau of the Census, *The Current Population Survey: Design and Methodology*, Technical Paper No. 66 (2006).

*Employment and Earnings* vol. 52, no. 12 (December, 2005).
Technical Documentation, March 2005 Current Population Survey.
M. Thompson and G. Shapiro, "The Current Population Survey: An overview," *Annals of Economics and Social Measurement* vol. 2 (1973).

2. There are a few exceptions in the Northeast; Hawaii is also exceptional.
3. In November 2005, the sampling fraction was 1/2160 overall, 1/283 for Wyoming, 1/3286 for Texas. Rates will change over time, especially when the sample is redesigned after the 2010 Census. A number of interesting points are glossed over in the main text. (i) The discussion focused on the "unit frame," for the household population. There is a separate frame for group quarters, and an "area frame" to sample geographical areas which are sparesly population, or where addresses are poorly defined. (ii) Some large USUs are treated differently. (iii) Within a PSU, the Bureau takes a random-start list sample. The list is organized to reduce variance. In effect, this stratifies the USUs. (iv) Fairly detailed information is collected on persons age 15, and on military personnel living off-post; some information is collected on persons age 14 and below. These data are not published.
4. For the 1995 design, the precision of monthly estimates in the 11 largest states was equalized, as well as the precision of the annual averages in the remaining 40 states; the District of Columbia counts as a state, and "precision" means the coefficient of variation of the estimated number of unemployed persons. For 2005, there were separate controls on several large substate areas, including the county of Los Angeles and the city of New York.
5. Persons who worked in their own business or profession, or on their own farm, are counted as employed; so are persons who worked at least 15 hours (even without pay) in a family business. Persons on layoff, but expecting to be recalled, need not be looking for work in order to be counted as unemployed. There is another classification—"discouraged workers"—for persons who want a job, but are no longer actively looking for a job because they believe no jobs are available. The official definitions have remained essentially unchanged since the survey was first conducted in 1940. Some revisions were made in 1994, as part of the redesign of the survey; and the questionnaire changed appreciably. See note 13 below.
6. Inmates of penal and mental institutions, and the military, are excluded from "the civilian noninstitutional population." Data in tables 1–3 are not seasonally adjusted.
7. The total labor force equals the civilian labor force plus the military.
8. The procedure for getting the weights is sometimes called "ratio estimation." The technique actually used by the Bureau is a bit more complicated, since they also cross-classify by other demographic variables. Furthermore, they make an adjustment to correct for non-interviews, and for known demographic differences between the sample PSUs and the country, using Census data. They make another adjustment to the current estimates using information from the previous month's sample. Finally, they adjust the weights in an effort to compensate for differential coverage in the Census (note 12 below).
9. The procedure used in the 1970s involved linearizing the estimates first, and computing some building-block variances by the half-sample method. It is sketched by R. S. Woodruff, "A simple method for approximating the variance of a complicated estimate," *Journal of the American Statistical Association* vol. 66 (1971) pp. 411–14. In the 1980s, a partially balanced replication method was used. See Janice Lent, "Variance estimation for Current Population Survey small area estimates," *Proceedings of the Section on Survey Research Methods* (American Statistical Association, August, 1991). A complete description of procedures for the period 1995–2005 is available in Technical Paper No. 66.
10. The stratification reduces the standard errors, as does the use of ratio estimates. But the clustering pushes the standard errors up.
11. *Statistical Abstract*, 2003, tables 396 and 419.
12. Census undercount in 1980 is discussed in R. E. Fay, J. S. Passel, J. G. Robinson and C. D. Cowan, *The Coverage of the Population in the 1980 Census*, Bureau of the Census, 1988; also see *Survey Methodology* vol. 18, no. 1, June, 1992. For discussions of the undercount in 1990, and proposals for adjustment, see *Jurimetrics* vol. 34, no. 1 (fall 1993), *Statistical Science* (November 1994), and *Evaluation Review* (August 1996).

    On proposed adjustments for Census 2000, see *Society*, vol. 39 no. 1 (November, 2001); D. A. Freedman and K. W. Wachter, "On the likelihood of improving the accuracy of the census through statistical adjustment," in *Science and Statistics: A Festscrift for Terry Speed*, Institute of Mathematical Statistics Monograph 40 (2003).

    On coverage differences between the Current Population Survey (CPS) and the Census, see pp. G5–6 in the technical documentation to the March 2005 CPS.
13. The evidence suggests that the Bureau can find out reasonably well who has a full-time job, and who is outside the labor force. The problem is with a third group, the marginal workers who are

classified either as part-time workers, or with a job but not at work, or unemployed. For example, results from the reinterview program for the last quarter of 1987 can be tabulated as shown below.

Thus, 7,511 people were reinterviewed; 3,015 were classified as working full time in nonagricultural industries at the original interview, but 2,997 were classified that way—presumably correctly—at reinterview. The decrease is 0.6 of 1%. On the other hand, the number of part-time workers went up by 4.5%, and the number of unemployed went up by 3.7%. The overall number of unemployed—based on the original interviews—was estimated as about 7,000,000. Since 3.7% of 7,000,000 = 250,000, the bias in the estimate amounts to several hundred thousand people. The number of unemployed persons in these data is small, so the calculation is only to illustrate the idea. Also see K. W. Clarkson and R. F. Meiners, "Institutional changes, reported unemployment, and induced institutional changes," Supplement to *Journal of Monetary Economics* (1979).

| | *Labor force status at original interview* | | | | | | |
|---|---|---|---|---|---|---|---|
| Labor force status at reinterview | Agri-culture | Employed in nonagricultural industry: Full-time | Part-time | Not at work | Unem-ployed | Not in labor force | Total |
| Agriculture | 117 | 1 | 2 | 0 | 0 | 2 | 122 |
| Non-agriculture | | | | | | | |
| full-time | 0 | 2,967 | 22 | 2 | 3 | 3 | 2,997 |
| part-time | 1 | 45 | 1,187 | 5 | 4 | 28 | 1,270 |
| not at work | 0 | 2 | 0 | 137 | 2 | 4 | 145 |
| Unemployed | 0 | 0 | 2 | 2 | 226 | 21 | 251 |
| Not in labor force | 0 | 0 | 2 | 1 | 7 | 2,716 | 2,726 |
| Total | 118 | 3,015 | 1,215 | 147 | 242 | 2,774 | 7,511 |

Notes: After reconciliation, before weighting; 75% sample.
Source: Bureau of the Census, Statistical Methods Division

In 1994, there was a major revision to the CPS questionnaire; new "probe" questions were added on hours of work and duration of unemployment; the definitions of "discouraged workers" and involuntary part-time workers were changed. See the *Monthly Labor Review* for September 1993, and *Employment and Earnings* for February 1994. Changing the questions made a noticeable impact on the numbers, confirming that biases in the data (although small) are probably larger than sampling error. Also see T. J. Plewes, "Federal agencies introduce redesigned Current Population Survey," *Chance* vol. 7, no. 1 (1994) pp. 35–41.

In theory, ratio estimates can create small biases. In practice, however, with reasonably large samples the bias from this source is negligible. There is one problem the Bureau does not have: household bias (note 15 to chapter 19). The reason is that the sample includes all persons age 16 and over in the selected households, not just one person that the interviewer finds at home.

14. Based on an example in Hyman's book (note 2 to chapter 19).
15. http://ag.ca.gov/newsalerts/2005/05-018.htm

## Chapter 23. The Accuracy of Averages

1. The draws can be made with replacement, or without. In the second case, the number of draws—and the number of tickets left in the box—both have to be large; the correction factor may be needed for computing the SE (chapter 20, section 3). See T. Höglund, "Sampling from a finite population: a remainder term estimate," *Scandinavian Journal of Statistics* vol. 5 (1978) pp. 69–71. If the number of draws is small, the distribution of the sum depends strongly on the contents of the box, and may be quite far from normal: see chapter 18, or section 26.6 on the *t*-test.
2. *Statistical Abstract*, 2003 gives enrollments by age and sex for the whole U.S.: see tables 286–87.
3. *Statistical Abstract*, 2003, tables 953ff reports information about housing stock, based on the Census of 2000 and the American Housing Survey. Table 970 gives median rents for states. Table 971 does large metropolitan areas: San Francisco-Oakland-San Jose is the highest, at $968. "Specified" units are defined in the headnote to table 965.
4. *Statistical Abstract*, 2003, tables 977, 1126 gives data for the whole U.S. (Strange but true: more households have a TV than an oven.) For hours spent watching television, see tables 1125, 1127.
5. *Statistical Abstract*, 2003, tables 40, 1093.

6. *The Nation's Report Card: Mathematics 2000*, http://nces.ed.gov/nationsreportcard.
7. The study was done in 1976. A previous phase of the Carnegie survey is discussed in Martin Trow, editor, *Teachers and Students* (McGraw-Hill, 1975). In fact, a stratified sample was used. In 1992, there were about 3,600 institutions; the average enrollment was about 4,000: *Statistical Abstract*, 1994, table 275. Also see the *Digest of Educational Statistics*, although definitions are not exactly comparable across sources. In 2000, there were about 4,200 institutions, with an average entrollment of about 3,700. Over the period 1975–2000, the professoriate swelled by 60%, while student enrollments went up by about 40%.
8. See note 7, and pp. 6–7 of Trow's book.
9. E. S. Pearson and J. Wishart, editors, *Student's Collected Papers* (Cambridge University Press, 1942).
10. A. R. Jensen, "Environment, heredity and intelligence," *Harvard Educational Review* (1969, p. 20). The quote was edited slightly. For a recent discussion of the substantive issues, see J. P. Rushton and A. R. Jensen, "Thirty years of research on race differences in cognitive ability," *Psychology, Public Policy, and Law* vol. 11 (2005) 235-294.
11. 380 U.S. 202 (1965). In subsequent cases, courts have held that juries must be representative of the community; random sampling gives the permissible deviations from community percentage makeup. The leading case is *Castaneda* 430 U.S. 482 (1977), especially note 17 at 496; also see *Avery* 345 U.S. 559 (1953), where Justice Frankfurter held that "the mind of justice, not merely its eyes, would have to be blind to attribute such an occasion to mere fortuity." In *Batson* 476 U.S. 79 (1986) the Supreme Court narrowed the right to peremptory challenges.

    "Standard deviation analysis" is now used not only in jury selection cases, but in employment discrimination litigation, and even in antitrust matters. *McCleskey* 481 U.S. 279 (1987) suggests that, for better or for worse, in capital punishment cases the courts are reluctant to accept statistical evidence of discrimination. For some discussion of statistical evidence from various points of view, see—

    B. Black, "Evolving legal standards for the admissibility of scientific evidence," *Science* vol. 239 (1988) pp. 1508–12; vol. 241 (1988) pp. 1413–14.

    S. Fienberg, editor, *The Evolving Role of Statistical Assessments as Evidence in the Courts* (Springer-Verlag, 1989).

    M. Finkelstein, "The application of statistical decision theory to the jury discrimination cases," *Harvard Law Review* vol. 338 (1966) pp. 353–56.

    M. Finkelstein and B. Levin, *Statistics for Lawyers*, 2nd ed. (Springer-Verlag, 2001).

    D. Kaye and D. Freedman, *Reference Guide on Statistics*, in *Reference Manual on Scientific Evidence*, 2nd ed. (Federal Judicial Center, Washington, D.C., 2000).

    P. Meier, J. Sacks and S. L. Zabell, "What happened in Hazelwood: Statistics, employment discrimination, and the 80% rule," *American Bar Foundation Research Journal* (1984) pp. 139–86.

    D. W. Peterson, editor, "Statistical inference in litigation," *Law & Contemporary Problems* vol. 46, no. 4 (1983).

    D. L. Rubinfeld, "Econometrics in the courtroom," *Columbia Law Review* vol. 85 (1985) pp. 1048–97.
12. Lee R. Jones et al., *The 1990 Science Report Card* (U.S. Department of Education, Office of Educational Research and Improvement, Washington, D.C., 1992). See pp. 135 and 165.

# Part VII. Chance Models

## Chapter 24. A Model for Measurement Error

1. W. J. Youden, *Experimentation and Measurement* (National Science Teachers Association, Washington, D.C., 1962).
2. Such equipment is manufactured by Toledo scale, using four load cells. Railway cars can move up to 6 mph as they cross the weigh-bridge.
3. Data from http://www.wunderground.com
4. The error box was a bit complicated: 95% of the tickets followed the normal curve, with an average of 0 and an SD of 4 micrograms; the other 5% followed the normal curve with an average of 0 and an SD of 25 micrograms. Two normal curves were needed, one for the middle and one for the outliers.
5. The root-mean-square might be even better, since the average of the box is assumed to be 0.
6. The empirical distribution of the data on NB 10 is skewed and long-tailed (figure 2 in chapter 6). However, the 100-fold convolution of this distribution with itself is quite close to normal; the minor deviations from normality are described quite well by an Edgeworth expansion to order $1/n$.

7. W. J. Youden, "Enduring values," *Technometrics* vol. 14 (1972) pp. 1–11. Also see M. Henrion and B. Fischhoff, "Assessing uncertainty in physical constants," *American Journal of Physics* vol. 54 (1986) pp. 791–97.
8. Dependence between repeated measurements is often caused by observer bias: the person making the measurements subconsciously wants the second measurement to be close to the first one. The Bureau takes elaborate precautions to eliminate this kind of bias. For instance, the value of NB 10 is obtained by comparing total masses of different sets of weights. These sets are varied according to a design chosen by the Bureau. The person who actually makes the measurements does not know how these sets are related to one another, and so cannot form any opinion about what the scales "should" read.
9. By Michelson, Pease, and Pearson at the Irvine Ranch in 1929–33. The results were rounded off a bit in the exercise. Their average value for the speed of light, converted to miles per second, is about 186,270. The measurements were taken in several groups, and there is some evidence to show that the error SD changed from group to group.

   In essence, the speed of light is now a definition: "In 1983 the General Conference on Weights and Measures officially redefined the meter as the distance that light travels in vacuum in 1/299,792,458 of a second." See E. M. Purcell, *Electricity and Magnetism*, 2nd ed. (McGraw-Hill, 1985, Appendix E).
10. The quote is from R. D. Tuddenham and M. M. Snyder, *Physical Growth of California Boys and Girls from Birth to Eighteen Years* (University of California Press, 1954, p. 191). It was edited slightly. As the authors continue,

    > With the wisdom of hindsight, we recognized in the later years of the study that a more accurate estimate of the theoretical "true value" would have been not the first measurement recorded, nor even the "most representative," but simply the [average] of the set.

## Chapter 25. Chance Models in Genetics

1. We are grateful for expert advice (some of which we took) from Everett Dempster and Michael Freeling of the Genetics Department, University of California, Berkeley. G. A. Marx and D. K. Ourecky of the New York State Agricultural Experiment Station were also extremely helpful. Finally, we thank Ann Lane, University of Minnesota. Standard textbooks on molecular genetics include—

   B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 4th ed. (Garland Publishing, New York, 2002).

   A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart, *An Introduction to Genetic Analysis*, 8th ed. (W. H. Freeman & Co., 2004).

   Desmond S. T. Nicholl, *An Introduction to Genetic Engineering*, 2nd ed. (Cambridge, 2002).

   Cancer biology provides fascinating insights into genetics. There are two books by Robert A. Weinberg that provide an excellent introduction to the field: *One Renegade Cell* (Basic Books, 1999) and *Biology of Cancer* (Garland Science, 2006).
2. Strictly speaking, this refers only to one part of the seeds, the *cotyledons* or first leaves.
3. The term "gene" is slightly ambiguous, but refers to a region of DNA coding for a protein or major protein chain, or to a specific variant of DNA in that region. ("Allele" is perhaps the better term for a specific segment of DNA occupying a coding region.) Presumably, several proteins are needed to determine seed color. If so, the pure yellow and pure green strains would have many of the corresponding alleles in common, but would differ on one pair—the $y/y$ and $g/g$ in the text. Mendel himself referred to "entities" which controlled phenotypes.
4. Sperm are carried by the pollen, eggs are in the ovules. Technically, these are nuclei not cells.
5. Maternal and paternal genes are sometimes distinguishable due to "imprinting." See C. Sapienza, "Parental imprinting of genes," *Scientific American* vol. 263 (October, 1990) pp. 52ff. Also see K. Peterson and C. Sapienza, "Imprinting the genome—imprinting, genes, and a hypothesis for their interaction," *Annual Review of Genetics* vol. 27 (1993) pp. 7–31.
6. The location of the genes on the chromosomes was worked out by Lamprecht (*Agric. Hortique Genetica*, 1961). There is a discussion in English by S. Blixt (same journal, 1972). Also see S. Blixt, "The Pea," chapter 9, vol. 2, *Handbook of Genetics (Plants, Viruses and Protista)* edited by R. C. King, Plenum, 1974; S. Blixt, "Why didn't Gregor Mendel find linkage?" *Nature* vol. 256 (1975) p. 206; E. Novitski and S. Blixt, "Mendel, linkage, and synteny," *Biosciences* vol. 28 (1978) pp. 34–35.
7. *Experiments in Plant Hybridisation* (Oliver & Boyd, 1965, p. 53). That book reprints Mendel's original paper, and some commentaries by Fisher, based on an article in the *Annals of Science* vol. 1 (1936) pp. 115–37. Some geneticists are quite critical of Fisher's reasoning; see, for example, F. Weiling, "What about R. A. Fisher's statement of the 'too good' data of J. G.

Mendel's Pisum paper?" *Journal of Heredity* vol. 77 (1986) pp. 281–83. On balance, Fisher's argument seems persuasive.

8. This experiment used five characteristics, not just the one discussed here. One trial was repeated, since Mendel thought the fit was poor. He used 100 plants in each trial, making the total of 600 referred to in the text.
9. "On the correlation between relatives on the assumption of Mendelian inheritance," *Transactions of the Royal Society of Edinburgh* vol. 52 pp. 399–433.
10. *Biometrika* (1903). The factor 1.08 more or less adjusts for the sex difference in heights. The equation is rounded off from the one in the paper.
11. There were 1,078 families in the study, so chance variation on this scale is very unlikely.
12. To get equation (5) from equation (3), take the conditional expectation given father's height; with non-assortative mating, mother's height is replaced by its overall average value. In fact, however, the correlation between parental heights was about 0.25.
13. Chromosomes may not replicate exactly in ordinary cell division. The "telomeres" (chromosome ends) seem to get shorter when the cell does not manufacture the enzyme telomerase. References—
    - C. W. Greider and E. H. Blackburn, "Telomeres, telomerase, and cancer," *Scientific American* (February 1996) pp. 92–97,
    - M. Barinaga, "Cells count proteins to keep their telomeres in line," *Science* vol. 275 (1997) p. 928.
    - D. A. Banks and M. Fossel, "Telomeres, cancer, and aging," *Journal of the American Medical Asociation* vol. 278 (1997) pp. 1345–48.
    - A. G. Bodnar et al., "Extension of life-span by introduction of telomerase into normal human cells," *Science* vol. 279 (1998) pp. 349–52.
    - C. Bischoff et al., "No association between telomere length and survival among the elderly and oldest old," *Epidemiology* vol. 17 (2006) pp. 190–94.
14. This discussion ignores more-complicated phenomena like mutation and crossover.
15. This exercise is adapted from M. W. Strickberger, *Genetics*, 3rd ed. (Macmillan, 1985). The focus here is the color of the pods, which may be quite different from the color of the seeds.
16. Rasmusson, *Hereditas* vol. 20 (1935). This problem too is from Strickberger.

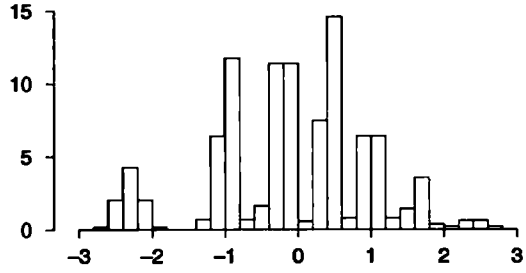
# Part VIII. Tests of Significance

## Chapter 26. Tests of Significance

1. From "Of experience," quoted in Jerome Frank, *Courts on Trial* (Princeton University Press, 1949).
2. Additional reading, in order of difficulty—
    - J. L. Hodges, Jr. and E. Lehmann, *Basic Concepts of Probability and Statistics*, 2nd ed. (SIAM, 2004).
    - L. Breiman, *Statistics with a View towards Applications* (Houghton Mifflin, 1973).
    - J. Rice, *Mathematical Statistics and Data Analysis*, 3d ed. (Duxbury Press, 2005).
    - P. Bickel and K. Doksum, *Mathematical Statistics*, 2nd ed. (Prentice Hall, 2001).
    - E. Lehmann, *Theory of Point Estimation*, 2nd ed. with G. Casella (Springer, 1998).
    - ———, *Testing Statistical Hypotheses*, 3rd ed. with J. Romano (Springer, 2005).
3. For a Bayesian, the frequentist $P$-value of a test can be substantially different from the posterior probability of the null hypothesis; indeed, the latter must depend on (i) the power of the test and (ii) the prior probability of the null. For more discussion, see J. Berger and T. Sellke, "Testing a point null hypothesis: The irreconcilability of $P$-values and evidence," *Journal of the American Statistical Association* vol. 82 (1987) pp.112–39.
4. In 2001, about 30% of people with full-time jobs were on flexible schedules: *Statistical Abstract*, 2003, table 606.
5. We are using ESP loosely to cover PK and clairvoyance as well. The experiment is described in C. Tart, *Learning to Use Extrasensory Perception* (University of Chicago Press, 1976). One subject did not, in fact, complete all the runs. And there are also results from a "Ten Choice Trainer," see exercise 6 on p. 487.
6. For the reasons given in the text, we do not consider it suitable to formulate the alternative hypothesis as drawing from a 0-1 box where the fraction of 1's is bigger than 1/4. It may be more reasonable to say that the number of correct guesses is stochastically larger than the sum of 7,500 draws from $\boxed{1\,0\,0\,0}$.
7. One such experiment was conducted by the former Professor W. Meredith, Psychology Department, University of California, Berkeley.

8. A. N. Doob et al., "Effect of initial selling price on subsequent sales," *Journal of Personality and Social Psychology*, vol. 11 (1969) pp. 345–50.
9. The anecdote about Student is reported in W. J. Youden, *Experimentation and Measurement* (Washington, D.C., 1963).
10. The $t$-test is one of the most popular statistical techniques, and we regret having to present it in a context which is both dry and partially hypothetical. (The story in the text is true, up to where they make a $t$-test; in practice, they don't.) We didn't run across any examples which were simultaneously real, interesting, and plausible. Our difficulty was the following. The $t$-test is used to compute significance levels. With small samples, some departures from normality can throw the computation off by a large factor. By way of illustration, the figure shows a probability histogram for the $t$-statistic, based on 10 draws made at random with replacement from the box $\boxed{-3}\boxed{-2}\boxed{5}$. The distribution is far from $t$-like.

Probability histogram for $t$-statistic based on 10 draws with replacement from $\boxed{-3}\boxed{-2}\boxed{5}$




The histogram was computed exactly, by considering all

$$\binom{12}{2} = 66$$

possible divisions of 10 objects into 3 groups; for each group, we computed the probability and the $t$-statistic. For 3 of the divisions, the $t$-statistic is undefined, but their mass is only $10^{-5}$; another 2% of the mass is outside the range $[-3, 3]$. The example can easily be modified so there is a smooth density with a shift parameter. For the combinatorics, see W. Feller, *An Introduction to Probability Theory and its Applications*, vol. I, 3rd ed. (John Wiley & Sons, 1968, section II.4).

To rely on the $t$-test, it seems to be necessary to know that the distribution of the errors is close to normal, without having a fair idea about the spread of the errors: if you knew the spread, you wouldn't be using the $t$-distribution. But how would you know the shape of the distribution without knowing its spread?

With large samples, departures from normality don't matter so much. Student's curves merge with the normal, and the $t$-statistic follows the normal curve (by the central limit theorem and the consistency of $\hat{\sigma}^2$ as an estimator of $\sigma^2$). This is one thing statisticians mean by the "robustness of the $t$-test." In our terms, this concept of robustness applies to the $z$-test not the $t$-test. Two references—

H. D. Posten, "The robustness of the one-sample $t$-test over the Pearson system," *Journal of Statistical Computation and Simulation* vol. 9 (1979) pp. 133–49.

E. Lehmann and W.-Y. Loh, "Pointwise vs. uniform robustness of some large sample tests and confidence intervals," *Scandinavian Journal of Statistics* vol. 17 (1990) pp. 177–87.

Small departures from independence can have large impacts on both the $z$-test and the $t$-test. Also see notes 12–13 below.

11. For present purposes, this is just a convention: the factor $\sqrt{n/(n-1)}$ could be absorbed into the multiplier derived from Student's curve. In some contexts, however, $\mathrm{SD}^+$ is preferred to the SD of the sample as estimator for the SD of the population: $(\mathrm{SD}^+)^2$ is unbiased, and this matters when pooling variances estimated from a large number of small samples.
12. The equation for the curve is

$$y = \text{constant}\left(1 + \frac{t^2}{d}\right)^{-(d+1)/2}$$

$$\text{constant} = 100\% \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)}$$

$$d = \text{degrees of freedom}$$

$$\Gamma = \text{Euler's gamma function}$$

The $t$-test was put on a rigorous mathematical footing by R. A. Fisher, who also showed that the procedure can give good approximations even when the errors did not follow the normal curve exactly: some departures from normality do not matter. This small-sample property is called "robustness" too. (But see note 10.)

13. If the tickets in the box follow the normal curve, then the probability histogram for the sum of the draws does too—even with only a few draws. Technically, the convolution of a normal curve with itself gives another normal curve. If the tickets in the box have a known distribution, which is not normal, statisticians can work out the probability histogram for the sum or average of the draws, using convolutions.
14. For national data, see J. H. Pryor et al., *The American Freshman: National Norms for Fall 2005* (Higher Education Research Institute, UCLA, 2006).
15. After Zeisel published the 1969 article, the next group of jurors chosen by Judge Ford was 24% female. References—

Hans Zeisel, "Dr. Spock and the case of the vanishing women jurors," *University of Chicago Law Review* vol. 37 (1969) pp. 1–18.

———, "Race bias in the administration of the death penalty: the Florida experience," *Harvard Law Review* vol. 95 (1981) pp. 456–68.

16. S. C. Truelove, "Therapeutic trials," in L. J. Witts, editor, *Medical Surveys and Clinical Trials* (Oxford University Press, 1959). Blinding the randomization is discussed in T. C. Chalmers, P. Celano, H. S. Sacks and H. Smith, Jr., "Bias in treatment assignment in controlled clinical trials," *New England Journal of Medicine* vol. 309 (1983) pp. 1358–61.
17. *Statistical Abstract*, 2003, tables 229, 1138, 1244 gives national data on education and reading. Also see *Reading At Risk: A Survey of Literary Reading in America* (National Endowment for the Arts, Washington, D.C., 2004). The latter publication takes a rather alarmist view of the prospects for the book, as the title indicates. By contrast, the data in *Statistical Abstract* suggest that books remain quite popular. For example, more people read books than surf the net.
18. These data originate with the Public Health Department of New York. We got them from Sandy Zabell, Professor of Statistics, Northwestern University. A reference is A. J. Izenman and S. L. Zabell, "Babies and the blackout: The genesis of a misconception," *Social Science Research* vol. 10 (1981) pp. 282–99. Apparently, the *New York Times* sent a reporter around to a few hospitals on Monday, August 8, and Tuesday, August 9, nine months after the blackout. The hospitals reported that their obstetrics wards were busier than usual—probably because of the general pattern that weekends are slow, Mondays and Tuesdays are busy. These "findings" were published in a front-page article on Wednesday, August 10, 1966, under the headline "Births Up 9 Months After the Blackout." That seems to be the origin of the baby-boom myth.

19. *Statistical Abstract*, 2003, table 681. The survey is hypothetical.
20. For a recent survey, see M. R. Rosenzweig and E. L. Bennett, "Psychobiology of plasticity: Effects of training and experience on brain and behavior," *Behavioural Brain Research* vol. 78 (1996) pp. 57–65. In fact, the experiment used not pairs but triplets, assigned at random to enriched, standard, and deprived environments. Data kindly provided by the investigators.

## Chapter 27. More Tests for Averages

1. Suppose $X$ and $Y$ are random variables, with variances $\sigma^2$ and $\tau^2$ respectively, and correlation $\rho$. Then $\text{var}(X - Y) = \sigma^2 + \tau^2 - 2\rho\sigma\tau$. If $\rho = 1$, the "chance errors" (i.e., departures from expected values) necessarily have the same sign, and offset each other—cancellation. Then the SE for the difference is $|\sigma - \tau|$. If $\rho = -1$, the errors reinforce each other, and the SE is $\sigma + \tau$. The case of independence corresponds to $\rho = 0$, and the SE is intermediate between the two extremes: $\text{SE} = \sqrt{\sigma^2 + \tau^2}$. If $X$ and $Y$ are independent, then $\text{SE}(X + Y) = \text{SE}(X - Y)$.
2. *NAEP 2004: Trends in Academic Progress*, http://nces.ed.gov/nationsreportcard. Average scores are as reported in the text; the sample was much larger, and highly designed; the SDs were close to 30. However, the SEs reported in the text are about right. Scores on these standardized performance tests seem to have bottomed out around 1970, with minor variations up and down since then. Using the pooled sample variance in the denominator of the test statistic would not be appropriate here, because the population variances may differ—even if the null hypothesis is true.
3. For the data source, see note 2. Three other tests are widely used for this sort of problem, besides the one presented in the text:

(i) Given the null hypothesis, the percentage of 1's in the two boxes is the same, and can be estimated by pooling the two samples:

$$\frac{107 + 132}{200 + 300} = \frac{239}{500} \approx 48\%.$$

On this basis, the common SD of the two boxes is estimated as

$$\sqrt{0.48 \times 0.52} \approx 0.50.$$

This pooled SD can be used to compute the SE in the denominator of the test statistic. (However, the pooled SD should not be used for other purposes, like putting confidence intervals around the difference.)

(ii) Fisher's "exact" test conditions on the total number of 1's in the two samples. The test statistic is the number of 1's in (say) the first sample; the sample sizes are fixed. The null distribution is the hypergeometric.

(iii) The $\chi^2$ statistic may be computed from the $2 \times 2$ table, and referred to a $\chi^2_1$-distribution.

Conditional on the total number of 1's in the two samples, test statistic (i) is a monotone function of the number of 1's in the first sample; so tests (i) and (ii) are equivalent. In effect, the hypergeometric has been approximated by the normal, which is fine for reasonably large samples. Likewise, (iii) is equivalent to (i) and (ii); indeed, $\chi^2 \equiv {Z_1}^2$, where $Z_1$ is the $z$-statistic computed with the pooled SD. If $Z$ is the $z$-statistic computed from the separate SDs, as in the text, then $Z^2 > {Z_1}^2$. However, as will be shown below, conditional on the total number of 1's in the two samples, $Z$ is a monotone function of the number of 1's in the first sample. So our test is equivalent to tests (i)-(ii)-(iii).

Some notation will be helpful. Let $\xi$ be the number of heads in $n$ tosses of a $p$-coin, and $\hat{p} = \xi/n$. Let $\zeta$ be the number of heads in $m$ tosses of a $q$-coin, and $\hat{q} = \zeta/m$. The two coins are independent. The null hypothesis is that $p = q$; the alternative, $p \neq q$. We condition on $s = \xi + \zeta$. Let $r = s/(m + n)$. This $r$ will be held constant.

For (i), the variance is computed as

$$v = r(1 - r)\left(\frac{1}{n} + \frac{1}{m}\right),$$

and the test statistic is

$$Z_1 = (\hat{p} - \hat{q})/\sqrt{v}.$$

For our test statistic, the variance is computed as

$$w = \frac{\hat{p}(1 - \hat{p})}{n} + \frac{\hat{q}(1 - \hat{q})}{m}$$

and the test statistic is

$$Z = (\hat{p} - \hat{q})/\sqrt{w}.$$

We view $m$, $n$, and $r = s/(m+n)$ as fixed; tacitly, we have assumed $0 < \hat{p}, \hat{q}, r < 1$. It will be convenient to introduce the new variable $x = \hat{p} - r$, so $\hat{q} = r - (nx/m)$. Clearly, $x$ takes only finitely many values. (Later, however, it will be convenient to think of $x$ as running through the whole interval from its minimum value to its maximum.) Now

$$\hat{p}(1-\hat{p}) = r(1-r) + (1-2r)x - x^2$$

$$\hat{q}(1-\hat{q}) = r(1-r) - \frac{n}{m}(1-2r)x - \frac{n^2}{m^2}x^2.$$

The two variances are related as follows:

$$w = v + bx - cx^2,$$

where

$$b = \left(\frac{1}{n} - \frac{n}{m^2}\right)(1-2r), \qquad c = \frac{1}{n} + \frac{n^2}{m^3}.$$

Finally, our test statistic is

$$Z = \frac{m+n}{m}x \Big/ \sqrt{v + bx - cx^2}$$

Of course, $w$, $v$, and $c$ are all positive; $b$ may be positive or negative. The monotonicity of $Z$ as a function of $x$ follows from the lemma below.

*Lemma.* Let $v, c > 0$ and let $b$ be real. Confine $x$ to the interval where $v + bx - cx^2 > 0$. Let

$$f(x) = x/\sqrt{v + bx - cx^2}$$

Then $f(x)$ is monotone increasing with $x$.

*Proof.* If $b \le 0$, this is trivial; so let $b > 0$. Then

$$\frac{df}{dx} = \frac{2v + bx}{2(v + bx - cx^2)^{3/2}} > 0.$$

4. See note 2.
5. *Statistical Abstract*, 2003, table 284. Most of the drop occurred between 1970 and 1980; the percentage bottomed out around 1995, and has been edging back up. Also see A. W. Astin et al., *The American Freshman: Thirty-Five Year Trends, 1966–2001* (Higher Education Research Institute, UCLA, 1991).
6. For national data, see J. H. Pryor et al., *The American Freshman: National Norms for Fall 2005* (Higher Education Research Institute, UCLA, 2006).
7. "Literacy among youths 12–17 years," *Vital and Health Statistics*, series 11, no. 131 (Washington, D.C., 1973). The sample design was like the Current Population Survey, and the investigators estimated the standard errors by the half-sample method. Simple random samples of the size indicated in the exercise will have standard errors about equal to the real ones.
8. "Intellectual development of children by demographic and socioeconomic factors," *Vital and Health Statistics*, series 11, no. 110 (Washington, D.C., 1971). For a discussion of the standard errors, see the previous note. The correlation between the children's test scores and parental education was 0.5, dropping to 0.3 when parental income was held constant. "Big city" means a population of 3 million or more. In fact, children in cities with a population in the range 1 to 3 million did best, averaging around 28 points.
9. For references to real trials on vitamin C, with interesting sidelights on blinding the randomization, see note 12, chapter 2.
10. In the text, we are using a strict null hypothesis: treatment has no effect on any subject. There is really only one number for each ticket, copied into both fields. The strict null has the charm of simplicity, and is appropriate in many cases. A weaker version of the null is sometimes used: the average response is the same for both treatments. This may be more realistic if the new treatment hurts some subjects but helps others. The alternative hypothesis usually does require two numbers for each subject, one for the response to treatment and one for the response to placebo. The discussion continues in note 11.
11. Consider a clinical trial to compare treatments A and B. We consider the weak form of the null, as in note 10; and the alternative, which does not constrain the responses at all. Suppose there are $N$ subjects, indexed by $i = 1, \ldots, N$. Let $x_i$ be the response of subject $i$ to treatment A; likewise,

$y_i$ is the response to B. For each $i$, either $x_i$ or $y_i$ can be observed, but not both. Let

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \qquad \tau^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2$$

$$\text{cov}\,(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

This model is sufficiently flexible to handle the weak form of the null hypothesis (note 10), as well as subject-to-subject heterogeneity under the alternative hypothesis. Thus, for instance, the average difference between treatments A and B—averaged over all the subjects in the study—is $\overline{x} - \overline{y}$. This "average causal effect" measures the difference between putting all the subjects into regime A, or putting all of them into regime B. The average causal effect is often the key parameter. And it is estimable, although the two responses are not simultaneously observable for any individual subject. Indeed, $\overline{x}$, $\overline{y}$, $\sigma^2$, and $\tau^2$ are all estimable; on the other hand, cov $(x, y)$ cannot be estimated by a sample covariance.

Responses in treatment and control are often modeled, for instance, as independent binomial with two different $p$'s, or independent normals with two different $\mu$'s. These parametric models seem less realistic. Independence of the two sample averages is generally wrong, and there is no reason to assume subjects are exchangeable within each treatment group. Such assumptions are *not* secured by randomization, which only makes the two groups comparable *as groups*. Thus, theoretical underpinnings are absent for, e.g., the $t$-test. It is surprising—and reassuring—that the permutation distributions of the conventional test statistics more or less coincide with the model-based distributions, at least in the contexts we are considering.

We now compute the variance of $\overline{X} - \overline{Y}$ under the alternative hypothesis, in our permutation setup. Let $S$ be a random subset of $\{1, \ldots, N\}$, with $n$ elements; this group gets treatment A, so $x_i$ is observed for $i \in S$. Let $T$ be a random subset of $\{1, \ldots, N\}$, with $m$ elements, disjoint from $S$. This group gets treatment B, so $y_j$ is observed for $j \in T$. We estimate the population means $\overline{x}$ and $\overline{y}$ by the sample means

$$\overline{X} = \frac{1}{n}\sum_{i \in S} x_i \qquad \overline{Y} = \frac{1}{m}\sum_{j \in T} y_j$$

By combinatorial calculations,

$$\text{var}\,\overline{X} = \frac{N-n}{N-1}\frac{\sigma^2}{n} \qquad \text{var}\,\overline{Y} = \frac{N-m}{N-1}\frac{\tau^2}{m}$$

$$\text{cov}\,(\overline{X}, \overline{Y}) = -\frac{1}{N-1}\,\text{cov}\,(x, y)$$

Thus

$$\begin{aligned}
\text{var}\,(\overline{X} - \overline{Y}) &= \frac{N-n}{N-1}\frac{\sigma^2}{n} + \frac{N-m}{N-1}\frac{\tau^2}{m} + \frac{2}{N-1}\,\text{cov}\,(x, y) \\
&= \frac{N}{N-1}\left(\frac{\sigma^2}{n} + \frac{\tau^2}{m}\right) + \frac{1}{N-1}\Big[2\,\text{cov}\,(x, y) - \sigma^2 - \tau^2\Big] \\
&\le \frac{N}{N-1}\left(\frac{\sigma^2}{n} + \frac{\tau^2}{m}\right)
\end{aligned}$$

because cov $(x, y) \le \sigma\tau$ and $2\sigma\tau - \sigma^2 - \tau^2 \le 0$. The "conservative estimate" in the text is $\sigma^2/n + \tau^2/m$. In practice, $\sigma^2$ and $\tau^2$ would be estimated by sample variances.

The signs may be a little perplexing. In general, we expect $x$ and $y$ to be positively correlated over all subjects. If too many subjects with high $x$-values are assigned to treatment A, then too few with high $y$-values are left for B. So the sample averages $\overline{X}$ and $\overline{Y}$ are negatively correlated. In principle, cov $(x, y)$ should be near its upper limit $\sigma\tau$, at least when $x$ and $y$ are highly correlated across subjects. Then the "conservative estimate" should be reasonably accurate for large samples. The strict null hypothesis in the text specifies that $x \equiv y$. Then $\sigma = \tau$, and the calculation is exact under the null hypothesis. Also see note 14 below. Of course, if $N$ is large relative to $m$ and $n$, then $\overline{X}$ and $\overline{Y}$ are nearly independent; again, the "conservative estimate" will be nearly right.

The impact of other variables may be handled as follows. Let $\eta$ denote treatment status. Let $\omega$ denote the state of other variables influencing the response. We assume there is a function $f$ such that the response of subject $i$ to treatment is $f(i, \eta, \omega)$. Let $\rho$ denote the assignment variable: if $\rho(i) = A$ then subject $i$ is assigned to treatment A, and likewise for B. We assume that $\rho$ and $\omega$ are independent: given $\omega$, the law of $\rho$ is uniform over all partitions of the subjects into a group $S$ of cardinality $n$ assigned to A and another group of cardinality $m$ assigned to B. The object of randomization, blinding, etc. is to secure this assumption. Then our argument can be done separately for each $\omega$, with

$$x_i = f(i, A, \omega) \quad \text{for } i \in S$$
$$y_j = f(j, B, \omega) \quad \text{for } j \in T$$

Few experiments are done on random samples of subjects. Instead, there is some initial screening process. Only subjects who pass the screen are randomized, and these subjects are best viewed as a sample of convenience. Therefore, some care is needed in setting up the inference problem. In our model, each subject has two potential responses, one to the treatment regime and one to the control regime. The "population" consists of pairs of responses. Both responses cannot be simultaneously observed for any subject. The experiment generates data not for the whole population, but for part of it. We observe responses to the treatment regime for subjects in the treatment group, and responses to the control regime for subjects in the control group. The statistical inference is from these observations to parameters characterizing the set of pairs of responses for the subjects that are randomized. The inference is not to some larger population of subjects—that kind of generalization would not be automatically justified by randomization. This is one aspect of Campbell's distinction between "internal validity" and "external validity:" see W. R. Shadish, T. D. Cook, W. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton Mifflin, 2002).

We are thinking primarily of experiments where subjects are divided into two random groups. However, similar comments apply if, for instance, subjects are paired by some ad hoc procedure; then a coin is tossed for each pair, choosing one subject for the treatment regime and one for the control regime. Again, the inference is to parameters characterizing the set of possible responses, and is made conditionally on the set of subjects and the pairing.

The model seems to go back to Neyman's early work on agricultural experiments. Some references:

J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczki* vol. 10 (1923) pp. 1–51, in Polish; English translation by D. Dabrowska and T. Speed, *Statistical Science*, vol. 5 (1990) pp. 463–80.

H. Scheffé, "Alternative models in the analysis of variance," *Annals of Mathematical Statistics* vol. 27 (1956) pp. 251–71.

J. L. Hodges, Jr. and E. Lehmann, *Basic Concepts of Probability and Statistics* (Holden-Day, 1964, section 9.4; 2nd ed. reprinted by SIAM, 2004).

D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology* vol. 66 (1974) pp.688–701.

J. Robins, "Confidence interval for causal parameters," *Statistics in Medicine* vol. 7 (1988) pp. 773–85.

P. Holland, "Causal inference, path analysis, and recursive structural equations models," *Sociological Methodology 1988*, C. Clogg, editor (American Sociological Association, Washington, D.C., Chapter 13.)

L. Dümbgen, "Combinatorial stochastic processes," *Stochastic Processes and their Applications* vol. 52 (1994) pp. 75–92.

D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, 2005).

Minor technical issues: (i) The relevant central limit theorem is for sampling without replacement (note 1, chapter 23). (ii) For small samples, the $t$-distribution may not provide a better approximation than the normal: the assumptions underlying the $t$-test do not hold.

12. A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," *Journal of Business* vol. 59, no. 4, part 2 (1986) pp. S251–78. Also see D. Kahneman and A. Tversky, "On the reality of cognitive illusions," *Psychological Review* vol. 103 (1996) pp. 582–96 (with discussion); D. Kahneman and A. Tversky, editors, *Choices, Values, and Frames* (Cambridge University Press, 2000); A. K. Sen, *Rationality and Freedom* (Harvard University Press, 2002).
13. B. J. McNeil, S. G. Pauker, H. C. Sox, Jr., and A. Tversky, "On the elicitation of preferences for alternative therapies," *New England Journal of Medicine* vol. 306 (1982) pp. 1259–62.
14. There were $80 + 87 = 167$ subjects in all (table 1). Of them, $40 + 73 = 113$ favored surgery; the remaining 54 favored radiation. The strict null hypothesis (note 10) specifies $x \equiv y$, so $\sigma = \tau$ and both are computable from the data. Indeed, on the null hypothesis, the percentage of doctors favoring surgery is $113/167 \times 100\% \approx 68\%$. Then

$$\sigma = \tau \approx \sqrt{0.68 \times 0.32} \approx 0.47$$

Likewise, the covariance between $\overline{X}$ and $\overline{Y}$ can be computed exactly. This term achieves the upper bound $\sigma\tau = \sigma^2$, because the correlation between $x$ and $y$ across subjects is 1. Now

$$\text{var}\,(\overline{X} - \overline{Y}) = \frac{N}{N-1}\left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2$$

The two forms of the test statistic (pooled or separate SDs, see note 3) are virtually identical. For example, if the null hypothesis defines the model, the r.m.s. difference between the values of the two statistics is only 0.013. Furthermore, the normal approximation is quite good: for either statistic, the chance of exceeding 2 in absolute value is about 4.8%, compared to the normal tail probability of 4.6%.

15. D. Kahneman and A. Tversky, "Choices, values, and frames," *American Psychologist* vol. 39 (1984) pp. 341–50.
16. In fact, the randomization was a bit more complicated. Inoculation required 3 separate injections over time, and hence the control group was given 3 injections (of the placebo) too. Vials containing the injection material were packed 6 to a box; 3 contalned the vaccine and had a common code number; the other 3 contained the placebo, with another common code number. Each vial had enough fluid for 10 injections.

    When the time came for the 1st round of injections, one vial was taken out of the box, and 10 children got their injections from that vial; the investigator recorded its code number against these 10 children; these 10 children got their 2nd and 3rd injections from the other 2 vials with the same code number in the box. The next 10 children got their 1st round injection from 1 of the 3 vials of the other group in that box (with a code number different from the 1st one used); the code number of the vial was recorded against them; and their subsequent injections were from the remaining 2 vials in the group.

    In effect, then, the children were blocked into pairs of groups of 10; a coin was tossed for each pair; one whole group went into treatment, and the other group into control, with a 50–50 chance. The calculation in the text is exact, on the plausible assumption that no 2 polio cases got injections from the same box. Otherwise, the calculation has to be modified. This particular trial is usually analyzed by the two-sample $z$-test, without taking account of the blocking (note 2 to chapter 1). We follow suit.
17. Barbara V. Howard et al., "Low-fat dietary pattern and risk of cardiovascular disease: The Women's Health Initiative randomized controlled dietary modification trial," *Journal of the American Medical Association* vol. 295 (2006) pp. 655–66.
18. D. Ravitch and C. E. Finn, Jr., *What Do Our 17-Year-Olds Know?* (Harper & Row, 1987, p. 52). The Soviet Union had the highest recognition factor.
19. http://www.gallup.com
20. References—

    K. Gray-Donald, M. S. Kramer, S. Munday et al., "Effect of formula supplementation in the hospital on duration of breast-feeding: A controlled clinical trial," *Pediatrics* vol. 75 (1985) pp. 514–18.

    K. Gray-Donald and M. S. Kramer, "Causality inference in observational vs. experimental studies: An empirical comparison," *American Journal of Epidemiology* vol. 127 (1988) pp. 885–92.

    Prior to running the controlled experiment, these investigators also ran an observational study, where both nurseries followed standard supplementation practice. There was a strong negative association between supplementation in the nurseries and breast-feeding later, as in the previous studies. Technically, assignment to the nurseries was not random. When a mother presented, she was assigned to the nursery with a bed available; this was done by clerical personnel not involved with the study. Eligibility was determined on objective criteria specified in the protocol. Unpublished data were kindly provided by the investigators.
21. Let $(X_i, Y_i)$ be independent and identically distributed pairs of random variables, with $E\{X_i\} = \alpha$, $\text{var}\,X_i = \sigma^2$, $E\{Y_i\} = \beta$, and $\text{var}\,Y_i = \tau^2$; let $\rho$ be the correlation between $X_i$ and $Y_i$, so $\text{cov}\,(X_1, Y_i) = \rho\sigma\tau$. Let $\overline{X} = (X_1 + \cdots + X_n)/n$ and $\overline{Y} = (Y_1 + \cdots + Y_n)/n$. The sample means are correlated, and $\text{var}\,(\overline{X} - \overline{Y}) = v/n$ with

$$v = \sigma^2 + \tau^2 - 2\rho\sigma\tau.$$

The variance $v$ would be estimated from sample data as

$$\hat{v} = \hat{\sigma}^2 + \hat{\tau}^2 - 2r\hat{\sigma}\hat{\tau},$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

and $r$ is the sample correlation coefficient,

$$r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/\hat{\sigma}\hat{\tau}.$$

The $z$-test would use the statistic $(\overline{X} - \overline{Y})/\sqrt{\hat{v}/n}$.

We now make the connection with the $z$-test based on the differences $X_i - Y_i$. Plainly, $\overline{X} - \overline{Y} = \overline{X - Y}$, the latter being $\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$. The differences $X_i - Y_i$ are independent and identically distributed, with $E\{X_i - Y_i\} = \alpha - \beta$ and $\text{var}\{X_i - Y_i\} = \sigma^2 + \tau^2 - 2\rho\sigma\tau = v$; of course, $\text{var}\{\overline{X - Y}\} = v/n = \text{var}\{\overline{X} - \overline{Y}\}$, where $v$ was defined above. The natural estimator for $v$ based on the differences is

$$\frac{1}{n}\sum_{i=1}^{n}\left[(X_i - Y_i) - (\overline{X - Y})\right]^2 = \hat{v},$$

coinciding with the variance estimator based on the paired data. (The equality takes a little algebra.) As a result, the $z$-statistic computed from the pairs must equal the $z$-statistic computed from the differences.

22. http://www.gallup.com
23. See note 22 for the source. The question was, "How would you rate the honesty and ethical standards of the people in these different fields—very high, high, average, low, or very low?" The percentage ratings of "very high or high" are shown in the table below, for some of the fields.

| | |
|---|---|
| Nurses | 82% |
| Druggists | 67% |
| Medical doctors | 65% |
| High school teachers | 64% |
| Clergy | 54% |
| Journalists | 28% |
| Building contractors | 20% |
| Lawyers | 18% |
| Congressmen | 14% |
| Car salesmen | 8% |
| Telemarketers | 7% |

24. A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science* vol. 211 (1981) pp. 453–458. Prices in the exercise were adjusted for inflation.
25. *The Third National Mathematics Assessment: Results, Trends and Issues* (Princeton: ETS/NAEP, 1983). The item is from the assessment, and the results are about as reported; the calculator group really did worse. However, it is not clear from the report whether the study was done observationally or experimentally.
26. P. H. Rossi, R. A. Berk and K. J. Lenihan, *Money, Work and Crime: Experimental Evidence* (San Diego: Academic Press, 1980, especially table 5.1). The study was done in 1976. We have simplified the experimental design, but not in any essential way; likewise, we changed the percents a little to make the testing problem sharper. Rossi et al. argue that income support did reduce recidivism, but the effect was masked by the impact on weeks worked. Their analysis has been criticized by H. Zeisel, "Disagreement over the evaluation of a controlled experiment," *American Journal of Sociology* vol. 88 (1982) pp. 378–96, with discussion.
27. S. J. Sherman, "On the self-erasing nature of errors of prediction," *Journal of Personality and Social Psychology* vol. 19 (1980) pp. 211–21.
28. William Epstein, as reported in the *New York Times*, September 27, 1988.

## Chapter 28. The $\chi^2$-Test

1. K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling," *Phil. Mag.*, series V, vol. 1 (1900) pp. 157–75.
2. If the chance model is right, each term is expected to be a bit less than one; the sum of all the terms is expected to be $n - 1$, where $n$ is the number of terms.
3. The equation for the curve is

$$y = \frac{100\%}{\Gamma(d/2)}\left(\frac{1}{2}\right)^{d/2} x^{(d/2)-1} e^{-x/2}$$

$d$ = degrees of freedom
$\Gamma$ = Euler's gamma function

4. The exact distribution was obtained using a program that stepped through all six-tuples of numbers adding up to 60, arranged in lexicographic order. It computed the $\chi^2$-statistic for each six-tuple, and the corresponding probability (using the multinomial formula). These probabilities were summed to give the answer—and the probability histogram in figure 2. The calculation seemed to be accurate to about 15 decimal places, since the sum of all the probabilities was $1 - 10^{-15}$. The wiggles in figure 2 are real.

   Many books recommend the Yates correction (subtracting 0.5 from the absolute difference before squaring, when this difference exceeds 0.5). With one degree of freedom, this is equivalent to the continuity correction (p. 317) and is a good thing to do. With more than one degree of freedom, numerical calculations show that it is often a bad thing to do. The histogram can be shifted much too far to the left. Numerical computations also show that with 5 observations expected per cell, and only a few degrees of freedom, the $\chi^2$-curve can be trusted out to the 5% point or so. With 10 observations expected per cell, the curve can be trusted well past the 1% point. Even if one or two cells in a moderate-size table have expecteds in the range 1–5, the approximation is often good.
5. When there are only two kinds of tickets in the box, the $\chi^2$-statistic is equal to the square of the $z$-statistic. Since the square of a normal variable is $\chi^2$ with 1 degree of freedom, the $\chi^2$-test will in this case give exactly the same results as a (two-tailed) $z$-test. Also see note 3 to chapter 27.
6. The data for this example, and for exercise 9 on p. 532, were kindly supplied by the California State Lottery through their statistical consultant Don Ylvisaker (UCLA).
7. In some cases (e.g., with only a few observations per cell), it is advisable to group the data.
8. *UCLA Law Review*, vol. 20 (1973) p. 615.
9. See note 7 to chapter 25.
10. A. R. Luria, *The Working Brain* (Basic Books, New York, 1973).
11. The HANES design involved a cluster sample, so there is some dependence in the data, which the $\chi^2$-test would not take into account. The half-sample method could be used to generate the null distribution. Women are consistently more right-handed than men, in all age groups. See *Anthropometric Reference Data and Prevalence of Overweight: United States, 1976–80.* Data from the National Health Survey, series 11, no. 238. (U.S. Department of Health and Human Services, Washington, D.C.). The numbers in table 5 are close to the real data, and make the arithmetic easier to follow.
12. Of course, if the test is done conditional on the marginals, the expecteds may be viewed as given. Also see note 3 to chapter 27.
13. Unweighted counts from a CD-ROM supplied by the Census Bureau, for the March 2005 Current Population Survey. The $\chi^2$-test does not take the design of the sample into account, but the difference is real.
14. *UCLA Law Review*, vol. 20 (1973) p. 616.
15. Unweighted counts from a CD-ROM supplied by the Census Bureau. The table is restricted to civilians. The $\chi^2$-test does not take the design of the sample into account. In many such surveys, across all age groups, the never-married men are less successful at work. For women, however, the unemployment rate for never-marrieds is about the same as for the married group. Also see R. M. Kaplan and R. G. Kronick, "Marital status and longevity in the United States population," *Journal of Epidemiology and Community Health* vol. 60 (2006) pp. 760–5.
16. This exercise is adapted from data supplied by IRRI.
17. Paraphrased from evidence presented at an extradition hearing for James Smyth, Federal District Court (N.D. Cal., 1993). See Defense brief of December 10, 1993 (pp. 7–8), Plaintiffs' exhibit 72.15, and Declaration of Robert Koyak. The District Court's decision not to extradite on grounds of probable discrimination was reversed on appeal.

## Chapter 29. A Closer Look at Tests of Significance

1. 225 U.S. 391, quoted from Jerome Frank, *Courts on Trial* (Princeton University Press, 1949).
2. When he was editor of the *Journal of Experimental Psychology*, Arthur Melton defended the practice in these words:

   > The next step in the assessment of an article involved a judgment with respect to the confidence to be placed in the findings—confidence that the results of the experiment would be repeatable under the conditions described. In editing the *Journal* there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were [only] significant at the .05 level, whether by one- or two-tailed test! This has not implied a slavish worship of the .01 level or any other level, as some critics may have implied. Rather, it reflects a belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying they were the product of the way the ball bounced.

There is a better way to make sure results are repeatable: namely, to insist that important experiments be replicated. The quote comes from an editorial in the *Journal* vol. 64 (1962) pp. 553–57. We found it in an article by David Bakan, reprinted in J. Steger, editor, *Readings in Statistics* (Holt, Rinehart and Winston, 1971). Also see note 4 below.

3. The history is on the authority of G. A. Barnard, formerly the professor of statistics, Imperial College of Science and Technology.
4. Unfortunately, even a relatively modest amount of data-snooping can produce off-scale $P$-values. Of course, the problems created for $P$-values should not stop investigators from looking at their data. One good research strategy is to *cross-validate:* develop the model on half the data, then see how well the fit holds up when the equations are applied to the other half. Real replication is even better. Replication is a crucial idea, and we do not do it justice in the text. References on data snooping and replication include—

   R. Abelson, *Statistics as Principled Argument* (Lawrence Erlbaum Associates, Hillsdale, N.J., 1995).

   T. K. Dijkstra, editor, *On Model Uncertainty and its Statistical Implications.* Springer Lecture Notes No. 307 in *Economics and Mathematical Systems* (1988).

   A. S. C. Ehrenberg and J. A. Bound, "Predictability and prediction," *Journal of the Royal Statistical Society*, series A, vol. 156, part 2 (1993) pp. 167–206.

   D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge, 2005).

   M. Oakes, *Statistical Inference* (ERI, Chestnut Hill, 1986).
5. The example is stylized, but the problem is real. We are assuming an incidence rate of 1 per 100,000 per year, and using a Poisson model. Despite concerns about environmental pollution, liver cancer rates have been falling steadily in the U.S. since the 1930s. For discussion and other references, see D. Freedman and H. Zeisel, "From mouse to man: The quantitative assessment of cancer risks," *Statistical Science* vol. 3 (1988) pp. 3–56, with discussion. Also see B. N. Ames, L. S. Gold and W. C. Willett, "The causes and prevention of cancer," *Proceedings of the National Academy of Sciences, U.S.A.* vol. 92 (1995) pp. 5258–65. For a controversial example of a cluster, see S. W. Lagakos, B. S. Wessen and M. Zelen, "An analysis of contaminated well water and health effects in Woburn, Massachusetts," *Journal of the American Statistical Association* vol. 81 (1986) pp. 583–614, with discussion. There is a fascinating account of the Woburn litigation by Jonathan Harr, *A Civil Action* (Random House, 1995). Also see R. B. Schinazi, "The probability of a cancer cluster due to chance alone," *Statistics in Medicine* vol. 19 (2000) pp. 2195–98.
6. In other cases, it is harder to correct the $P$-value for data snooping. See the book by Dijkstra, cited in note 4. For some discussion of the impact on journal publications, see—

   L. J. Chase and R. B. Chase, "A statistical power analysis of applied psychological research," *Journal of Applied Psychology* vol. 61 (1976) pp. 234–37.

   K. Dickersin, S. Chan, T. C. Chalmers, H. S. Sacks and H. R. Smith, Jr., "Publication bias and clinical trials," *Journal of Controlled Clinical Trials* vol. 8 (1987) pp. 343–53.

   A. Tversky and D. Kahneman, "Belief in the law of small numbers," *Psychological Bulletin* vol. 2 (1971) pp. 105–10.

   C. B. Begg and J. A. Berlin, "Publication bias and dissemination of clinical research," *Journal of the National Cancer Institute* vol. 81 (1989) pp. 107–15.
7. "The Lipid Research Clinics Primary Prevention Trial Results," *Journal of the American Medical Association* vol. 251 (1984) pp. 351–64. The investigators quote $z \approx -1.92$, based on lifetable analysis and blocking. The protocol did not state whether one- or two-tailed tests would be used; it noted "significant morbidity and mortality associated with cholesterol-lowering agents"; and declared that a significance level of 1% "was chosen as the standard for showing a convincing difference between treatment groups." There was a strong suggestion that fatal and non-fatal heart attacks would be analyzed separately—in which case the differences are not significant. See *Journal of Chronic Diseases* vol. 32 (1979) pp. 609–31. The investigators do not appear to have followed protocol. Also see *Journal of Clinical Epidemiology* vol. 43 no. 10 (1990) pp. 1021ff. There are less-formal accounts by T. J. Moore, *Heart Failure* (Random House, 1989) and *Lifespan* (Simon & Schuster, 1993).

   Another experiment is reported by H. Buchwald et al., "Effect of partial ileal bypass surgery on mortality and morbidity from coronary heart disease in patients with hypercholesterolemia," *New England Journal of Medicine* vol. 323 (1990) pp. 946–55. But see G. D. Smith and J. Pekkanen, "Should there be a moratorium on the use of cholesterol lowering drugs?" *British Medical Journal* vol. 304 (1992) pp. 431–34: the evidence from several trials suggests that cholesterol-lowering drugs actually increase the death rate. On the other hand, a large Scandinavian study on Simvastatin obtained a 30% reduction in mortality, among subjects with a history of heart disease. See "Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study," *Lancet* vol. 344 (November 19, 1994) pp. 1383–89.

There is also the Scottish study on pravastatin, see the *New England Journal of Medicine* (November 16, 1995). For a review, see A. M. Garber, W. S. Browner and S. B. Hulley, "Cholesterol screening in asymptomatic adults, revisited," *Annals of Internal Medicine* vol. 124 (1996) pp. 518–31.

8. K. R. Rao, editor, "The Ganzfeld debate," *Journal of Parapsychology* vol. 49, no. 1 (1985) and vol. 50, no. 4 (1986). The discreteness of the distributions matters, and significance probabilities must be computed by convolution.
9. The evaluation of bioassay results is a complicated issue, but the multiple-endpoint problem is a real one. Many chemicals do seem to cause liver cancer but prevent leukemia in mice. See the paper by Freedman and Zeisel referenced in note 5. Also see T. S. Davies and A. Monro, "The rodent carcinogenicity bioassay produces a similar frequency of tumor increases and decreases: Implications for risk assessment," *Regulatory Toxicology and Pharmacology* vol. 20 (1994) pp. 281–301; T. H. Lin et al., "Carcinogenicity tests and inter-species concordance," *Statistical Science* vol. 10 (1995) pp. 337–53.
10. T. C. Chalmers, R. S. Koff and G. F. Grady, "A note on fatality in serum hepatitis," *Journal of Gastroenterology and Hepatology* vol. 69 (1965) pp. 22–26.
11. The confusion between "statistical significance" and importance gets worse with correlation coefficients. Instead of looking at the value of $r$, some investigators will test whether $r = 0$, and then use $P$ as the measure of association. Regression coefficients often get the same treatment. However, it is the analysis of variance which presents the problem in its most acute form: some investigators will report $P$-values, $F$-statistics, everything except the magnitude of their effect. For some discussion, see P. E. Meehl, "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology," *Journal of Consulting and Clinical Psychology* vol. 46 (1978) pp. 806–34.
12. On the other hand, there may be noticeable differences in reading abilities between big-city children and rural children, in later ages. See I. S. Kirsch and A. Jungeblut, *Literacy: Profiles of America's Young Adults* (ETS/NAEP, Princeton, N.J., 1986).
13. The 6 points comes from a rough-and-ready regression analysis of auxiliary data, and includes selection effects. Other indicators of school quality are discussed in review exercise on p. 94 and exercise 2 on p. 506.
14. This is a close paraphrase of a comment (taken out of context) by D. T. Campbell, "Reforms as experiments," *American Psychologist* vol. 24 (1969) pp. 409–29. The reference was supplied by the late Merrill Carlsmith, formerly professor of psychology, Stanford University.
15. M. J. Mahoney, "Publication prejudices: An experimental study of confirmatory bias in the peer review system," *Journal of Cognitive Therapy and Research* vol. 1 (1977) pp. 161–75. The experimental design, and the quotes, have been simplified a little.
16. Daniel McFadden, "The revealed preferences of a government bureaucracy: Empirical evidence," *Bell Journal of Economics* vol. 7 (1971) pp. 55–72. The study period was 1958–66. The "effect" of a variable is a coefficient in a model; of course, the model may be open to question. This reference was supplied by Chris Achen, professor of political science, University of Michigan.
17. Paraphrase of testimony by W. Hogan and J. Kalt (Harvard) in a 1987 administrative hearing on violations of oil price controls. Elasticity is a price coefficient in a regression model.
18. To paraphrase Keynes, the significance tester who thinks he doesn't need a box model may just have a naive one. J. M. Keynes, *The General Theory of Employment, Interest, and Money* (Harcourt Brace Jovanovich, 1935, pp. 383–84).

    > Practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist.

19. *Statistical Abstract*, 2003, table 11.
20. This study was discussed in section 4 of chapter 2; also see note 7 to that chapter, for references. In this example, $z \approx 5$ so $P$ is rather small. We can interpret $P$ as a descriptive statistic. Altogether there were 933 candidates, of whom 825 were men and 108 were women. If you think that sex and admissions were unrelated, comparing admission rates for men and women is like comparing the admission rate for any group of 825 people with the admission rate for the remaining group of 108 people. (After all, there are many irrelevant splits, based on fingerprints and so forth.) There are

$$\binom{933}{825} \approx 7 \times 10^{143}$$

possible ways to split the 933 candidates into two groups, one of size 825 and the other of size 108. For each split, compute $z$. This population of $z$-values is close to normally distributed, so the observed $z$-value of 5 is quite unusual. See D. Freedman and D. Lane, "A nonstochastic interpretation of reported significance levels," *Journal of Business and Economic Statistics* vol. 1

(1983) pp. 292–98. The idea goes back to R. A. Fisher. See E. J. G. Pitman, "Significance tests which may be applied to samples from any population," *Journal of the Royal Statistical Society* Series B vol. 4 (1936) pp. 119–30.

21. *Project Follow Through Classroom Evaluation*, published by SRI at Menlo Park, California. The senior investigator was Jane Stallings. The quotes were edited slightly. The study was done in 1972–73.
22. This assumes the control average of 60 to be known without error. In fact, SRI made a two-sample $t$-test. However, the SRI scoring procedure was bound to introduce dependence between treatment and control scores—it was based on pooled ranks.
23. These are real numbers, from 1976. About half the TAs had participated in grading the final, and many had graded similar finals in previous years. Over time, the graduate students did learn how to handle Statistics 2 problems.
24. F. Mosteller and R. Rourke, *Sturdy Statistics* (Addison-Wesley, 1973, p. 54).
25. T. A. Ryan, B. L. Joiner and B. F. Ryan, *Minitab Student Handbook* (Duxbury Press, Boston, 1976, p. 228).
26. "Intellectual development of children by demographic and socioeconomic factors," *Vital and Health Statistics* series 11, no. 110 (Washington, D.C., 1971).
27. R. S. Erikson, J. P. McIver and G. C. Wright, Jr., "State political culture and public opinion," *American Political Science Review* vol. 81 (1987) pp. 797–813. The analytic technique was multiple regression on dummy variables for demographic categories (e.g., low income, etc.); then dummies were added for regions and states. Adding in the state dummies increased the adjusted $R^2$ from 0.0898 to 0.0953, but the $F$ to enter was 8.35, with 40 degrees of freedom in the numerator—and 55,072 in the denominator. The authors say that the state effects are significant in practical terms as well; the $R^2$'s suggest otherwise. The authors acknowledge that state dummies may be proxies for omitted variables, but argue against this interpretation. The papers cited in this note and the next are discussed by D. A. Freedman, "Statistical models and shoe leather," in P. Marsden, editor, *Sociological Methodology 1991* (American Sociological Association, Washington, D.C., chapter 10). Also see D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, 2005).
28. J. L. Gibson, "Political intolerance and political repression during the McCarthy era," *American Political Science Review* vol. 82 (1988) pp. 511–39. "Effects" are coefficients in a path model. Presumably, the author would view the randomness in the estimates as generated by the model. On the other hand, the adequacy of the model may be open to question.
29. The experiment is discussed by C. E. M. Hansel, *ESP: A Scientific Evaluation* (Charles Scribner's Sons, 1966, chapter 11). The numbers have been changed to simplify the arithmetic. The point of the experiment was to illustrate the fallacy discussed in the text. The reference was supplied by Charles Yarbrough, Santa Rosa, Calif.
30. The random number generator on the Aquarius itself does not seem to have been tested, but the generator is similar to ones that were tested. In ESP research, nothing is simple, and Tart would not agree with much of what we write: C. Tart et al., "Effects of immediate feedback on ESP performance: A second study," *Journal of the American Society for Psychical Research* vol. 73 (1979) pp. 151–65. For a lively discussion of the issues, see Martin Gardner, *Science: Good, Bad, and Bogus* (Avon Books, 1981, chapters 18 and 31).
31. Reproduced by permission of the publisher, Harcourt Brace Jovanovich, Inc.
32. Based on a question used by A. Tversky and D. Kahneman. Also see p. 298 in Steger's book, referenced in note 2 above.
33. See p. 68 of Mosteller and Rourke, note 24.
34. F. Arcelus and A. H. Meltzer, "The effect of aggregate economic variables on congressional elections," *American Political Science Review* vol. 69 (1965) pp. 1232–69, with discussion. This reference was supplied by Chris Achen. The argument uses a regression model, and is therefore more subtle than indicated in the exercise. (Of course, the validity of the model is open to question.) However, the investigators' position on hypothesis testing is brutal; see the rejoinder by Arcelus and Meltzer to the comments by Goodman and Kramer.
35. *Statistical Abstract*, 1988, table 21; *Statistical Abstract*, 1994, table 26; *Statistical Abstract*, 2003, table 17.
36. *Statistical Abstract*, 1994, tables 616 and 621. *Employment and Earnings* vol. 52, no. 12 (December, 2005), table A-2.
37. *Statistical Abstract*, 2003, table 284. Also see A. W. Astin et al., *The American Freshman: Thirty-Five Year Trends, 1966–2001* (Higher Education Research Institute, UCLA, 1991). Most of the change occurred between 1970 and 1980.
38. R. E. Just and W. S. Chern, "Tomatoes, technology and oligopsony," *Bell Journal of Economics* vol. 11 (1980) pp. 584–602. For discussion, see R. Daggett and D. Freedman, "Econometrics and the law: A case study in the proof of antitrust damages," in L. M. LeCam and R. A. Olshen, editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* vol. 1,

pp. 123–72 (Wadsworth, Belmont, California, 1985). Just and Chern estimated both linear and log-linear demand functions; the $t$-test reported in the exercise was applied to the coefficient of price in a linear demand function.

39. For national data, see *Statistical Abstract*, 2003, table 1244. By this measure, dining out was the most popular activity, followed by reading, and entertaining at home.
40. The quote is from D. L. Hartl, Letter, *Nature* vol. 372 (1994) p. 398; we thank David Kaye (Arizona State University) for calling it to our attention. Also see note 6 to chapter 13.
41. June 27, 1993.
42. Paraphrased from evidence presented at an extradition hearing for James Smyth, Federal District Court (N.D. Cal., 1993). Defense Exhibit 31, *Secondary Analysis of the School Leavers Survey* (1989), Standing Advisory Commission on Human Rights, by Cormack et al.
43. Data are from Thomas H. Cohen and Steven K. Smith (2004), *Civil Trial Cases and Verdicts in Large Counties 2001*, Bureau of Justice Statistics, U.S. Department of Justice. Results were simplified a little. Jury awards have declined over the period 1991–2001. Interestingly enough, judges tend to be more generous to plaintiffs than are juries.
44. See R. C. Lewontin, "Sex, lies, and social science," in *New York Review of Books*, April 20, May 25, and August 10, 1995. Lewontin is reviewing R. T. Michael et al., *Sex In America: A Definitive Survey* (Little Brown, 1994), which is a popularized version of E. O. Laumann et al., *The Social Organization of Sexuality: Sexual Practices in the United States* (University of Chicago Press, 1994). Also see Devon D. Brewer et al., "Prostitution and the sex discrepancy in reported number of sexual partners," *Proceedings of the National Academy of Sciences of the U.S.A.*, vol. 97 (2000) pp. 12385–388. Brewer et al. find that female prostitutes—who have very large numbers of male partners—are substantially under-represented in the survey; and "men are reluctant to acknowledge that their reported partners include prostitutes."
45. John A. Dossey et al., *Can Students Do Mathematical Problem Solving?* (U.S. Department of Education, Office of Educational Research and Improvement, Washington, D.C., 1992, pp. 141, 172).
46. *Brock v. Merrell Dow Pharmaceuticals, Inc.*, 874 F.2d 307, 311–12 (5th Cir.), modified, 884 F.2d 166 (5th Cir. 1989), cert. denied, 494 U.S. 1046 (1990); D. H. Kaye and D. A. Freedman, *Reference Guide on Statistics*, 2nd ed. (Federal Judicial Center, Washington, D.C., 2000, p. 121).
47. See W. T. Keeton, J. L. Gould, and C. G. Gould, *Biological Science*, 5th ed. (W. W. Norton & Company, 1993, p. 445).
48. *Statistical Abstract*, 2003: table 66 gives 108 million households, table 305 gives 2.11 million burglaries reported to the police, table 321 gives 3.14 million burglaries reported to the survey. The survey uses a highly designed sample, but a simple random sample of 50,000 gives (roughly) the right standard errors. Also see J. P. Lynch and L. A. Addington, *Understanding Crime Statistics* (Cambridge, 2007).
49. The randomization included blocking, not accounted for here. The averages were published; the SDs were kindly provided by J. D. Neaton (professor of biostatistics, University of Minnesota). An interesting sidelight: logistic regressions fitted to the Framingham data predicted a very substantial reduction in mortality due to the modest-looking decrements in risk factors (3 mm in blood pressure, 5 mg/dl in serum cholesterol, 13% in smoking). There was some concern that smoking was under-reported by the treatment group, and an adjustment was made for this by blood chemistry. References—

    "Multiple Risk Factor Intervention Trial," *Journal of the American Medical Association* vol. 248 (1982) pp. 1465–77.

    "Statistical design considerations in the NHLI Multiple Risk Factor Intervention Trial (MRFIT)," *Journal of Chronic Diseases* vol. 30 (1972) pp. 261–75.

    "Mortality rates after 10.5 years for participants in the Multiple Risk Factor Intervention Trial," *Journal of the American Medical Association* vol. 263 (1990) pp. 1795–1801.
50. http://www.gallup.com
51. *Waisome v. Port Authority*, 948 F.2d 1370, 1376 (2nd Cir. 1991); D. H. Kaye and D. A. Freedman, *Reference Guide on Statistics*, 2nd ed. (Federal Judicial Center, 2000, Washington, D.C., p. 124). The quote is edited slightly.
52. M. S. Kanarek et al., "Asbestos in drinking water and cancer incidence in the San Francisco Bay Area," *American Journal of Epidemiology* vol. 112 (1980) pp. 54–72. There was no relationship between asbestos in the water and lung cancer for blacks or women. Data in the paper strongly suggest that smoking was a confounder. For more discussion, see D. A. Freedman, "From association to causation: Some remarks on the history of statistics," *Statistical Science*, vol. 14 (1999) pp. 243–58; reprinted in *Journal de la Société Française de Statistique*, vol. 140 (1999) pp. 5–32 and in *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century* (Lawrence Erlbaum Associates, 2003, pp. 45–71), edited by J. Panaretos.
53. See note 20 to chapter 19. Children with no siblings are an exception, scoring slightly below first-borns in two-child families.

# Answers to Exercises

## Part I. Design of Experiments

### Chapter 2. Observational Studies

Set A, page 20

1. False. The population got bigger too. You need to look at the number of deaths relative to total population size. The population in 2000 was about 281 million, and in 1970 it was about 203 million: 2.4 out of 281 is smaller than 1.9 out of 203, so the death rate was lower in 2000. There was a very considerable increase in life expectancy between 1970 and 2000.

   *Comment.* Between 1970 and 2000, the population got older, on average, so the reduction in death rates is even more impressive.

2. The basic facts: richer families are more likely to volunteer for the experiment, and their children more vulnerable to polio (section 1 of chapter 1).

   (a) From line 1 of the table, the polio rates in the two vaccine groups were about the same. If (for example) the consent group in the NFIP study had been richer, their rate would have been higher.
   (b) From line 3 of the table, the polio rates in the two no-consent groups were about the same.
   (c) From line 2 of the table, the polio rate in the NFIP control group was quite a bit lower than the rate in the other control group.
   (d) The no-consent group is predominantly lower-income, and the children are more resistant to polio. The NFIP control group has a range of incomes, including the more vulnerable children from the higher-income families.
   (e) The ones who consent are different from the ones who don't consent (p. 4).

   *Comment on (c).* The NFIP controls had a whole range of family backgrounds. The controls in the randomized experiment were from families who consented to participate. These families were richer, and their children more vulnerable to polio. The NFIP design was biased against the vaccine.

3. Children who were vaccinated might engage in more risky behavior—a bias against the vaccine. On the other hand, the placebo effect goes in favor of the vaccine. (The similarity of rates in line 1 of table 1, p. 6, suggests biases are small.)

4. No, because the experimental areas were selected in those parts of the country most at risk from polio. See section 1 of chapter 1.

5. The people who broke the blind found out whether or not they were getting vitamin C. The ones who knew they were getting vitamin C for prevention tended to get fewer colds. Those on vitamin C for therapy tended to get shorter colds. This is the placebo effect. Blinding is important.

6. $558/1{,}045 \approx 53\%$, and $1{,}813/2{,}695 \approx 67\%$. Adherence is lower in the nicotinic acid group. Something went wrong with the randomization or the blind. (For example, nicotinic acid might have unpleasant side effects, which causes subjects to stop taking it.)

7. In trial (i), something must have gone wrong with the randomization. The difference between 49.3% and 69.0% shows that the treatment group smoked less to begin with, which would bias any further comparisons. The difference cannot be due to the treatment, because baseline data say what the subjects were like before assignment to treatment or control. (More about this in chapter 27.)

8. Option (ii) explains the association, option (i) does not. Choose (ii). See p. 20.

9. (a) Yes: 39 deaths from breast cancer in the treatment group, versus 63 in the control group.
   (b) The death rate in the treatment group (screened and refused together) is about the same as the death rate in the control group because screening has little impact on deaths from causes other than breast cancer.
   (c) Compare A) the control group with B) those who refused screening in the treatment group. Group A includes women who would accept screening as well as those who would refuse. On average, then, group A is richer than group B. Neither group is affected by screening, and group A has a higher death rate from breast cancer.
   (d) Most deaths are from causes other than breast cancer; those rates are not affected by screening. However, the women who refuse screening are poorer and more vulnerable to most diseases. That is why their death rates are higher.

   *Comments.* (i) In part (a), you should compare the whole treatment group with the whole control group. This is the "intention to treat" principle. It is conservative, that is, it understates the benefit of screening. (If all the women had come in for screening, the benefit would have been higher.) You should not compare the "examined" with the "refused" or with the controls: that is biased against treatment, see exercise 10(a).

   (ii) The Salk vaccine field trial could have been organized like HIP: (1) define a study population of, say, 1,000,000 children; (2) randomize half of them to treatment and half to control, where treatment is the invitation to come in and be vaccinated; (3) compare polio rates for the whole treatment group versus the whole control group. In this setup, it would not be legitimate to compare just the vaccinated children with the controls; you would have to compare the whole treatment group with the whole control group. The design actually used in the Salk field trial was better, because of the blinding (section 1 of chapter 1); however, this seems to have been a relatively minor issue for HIP, and the design they used is substantially easier to manage.

10. (a) This is not a good comparison. There is a bias against screening. The comparison between the "examined" and "refused" groups is observational, even though the context is an experiment: it is the women who decide whether to be examined or not. This is just like adherence to protocol in the clofibrate trial (section 2). There are confounding variables, like income and education, to worry about. These matter. The comparison is biased against screening because the women who come in for examination are richer, and more vulnerable to breast cancer.
    (b) This is not a good theory: the overall death rate in the treatment group from diseases other than breast cancer is about the same as that in the control group, and the reduction in breast cancer death rate is due to screening.
    (c) False. Screening detects breast cancers which are there and would otherwise be detected later. That is the point of screening.

*Comments.* (i) In the HIP trial, the number of deaths from other causes is large, and subject to moderately large chance effects, so the difference $837-879 = -42$ is not such a reliable statistic. More about this in chapter 27. The comparison of 1.1 and 1.5 in 10(a) is very unreliable, because the number of breast cancers is so small—23 and 16. However, the difference between 39 and 63 in 9(a) is hard to explain as a chance variation.

(ii) In part 10(c), within the treatment group, the screened women had a higher incidence rate of diagnosed breast cancer, compared to the women who refused. The two main reasons: (1) screening detects cancers; (2) breast cancer—like polio and unlike most other diseases—hits the rich harder than it hits the poor, and the rich are more likely to accept screening.

(iii) The benefits of mammography for women age 50–70 are now generally recognized; there remains some question whether the benefits extend to women below the age of 50. For references, see note 14 to chapter 2.

11. The women who have been exposed to herpes are the ones who are more active sexually; this evidence is not convincing. (See example 2 on p. 16.)

    *Comment.* In the 1970s, herpes (HSV-2) was thought to be causal. In the 1980s, new evidence from molecular biology suggested that HSV was not a primary causal agent, and implicated strains of human papilloma virus (HPV-16,18). For references, see note 4 to chapter 2.

12. If a woman has already aborted in a previous pregnancy—and is therefore more at risk in her current pregnancy—a physician is likely to tell her to cut down on exercise. In this instance, exercise is a marker of good health, not a cause.

13. False. Altogether, 900 out of 2,000 men are admitted, or 45%; while 360 out of 1,100 women are admitted, or 33%. This is because women tend to apply to department B, which is harder to get into. See section 4.

14. (a) 39 out of 398 is like 40 out of 400, or 10 out of 100, or 10%.
    (b) 25% (c) 25% (d) 50%

15. (a) 10%. That's spread over a $10,000 range, so for the next three parts, guess about 1% in each $1,000 range.
    (b) 1% (c) 1% (d) 2%

# Part II. Descriptive Statistics

## Chapter 3. The Histogram

Set A, page 33

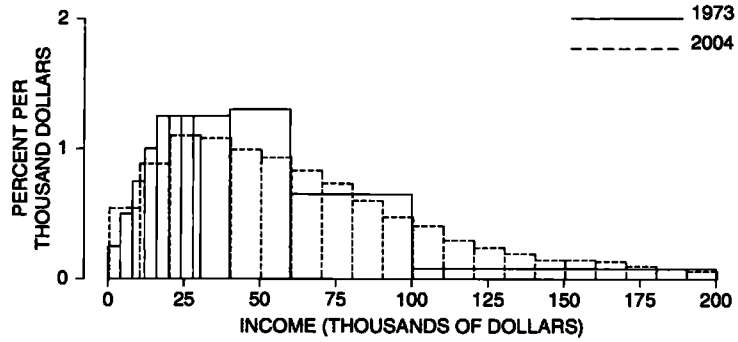
1. (a) 2% (b) 3% (c) 4% (d) 5% (e) 15% (f) 15%
2. More between $10,000 and $11,000.
3. (a) B (b) 20% (c) 70%
4. (a) Well over 50%. (b) Well under 50%. (c) About 50%.
5. Class (b).

6. There were more in the range 90 to 100.

7. A (ii), B (i), C (iii)

8. The figure does not adjust for inflation, so the comparison is not a good one.

*Comment.* In 1973, a dollar bought roughly 4 times as much as in 2004. The figure below compares the 2004 histogram with the 1973 histogram—corrected for this change in purchasing power. Family income went up by a factor of about 4 in "nominal" dollars, but in "real" dollars—corrected for inflation—there was not that much improvement. (We shifted the 2004 histogram to the right a little; data on the consumer price index are from *Statistical Abstract*, 1993, table 756; 2003, table 713; table 690 in the latter publication suggests about a 15% increase in real family income over the period 1980–2000; prices indices are not the most reliable of statistics, because they may not reflect quality improvements.)




Set B, page 38

1. The 1991 histogram is shown in figure 5 on p. 39, and the reason for the spikes is discussed on that page.

2. Smooths out the graph between 0 and 8.

3. The educational level went up. For example, more people finished high school and went on to college in 1991 than in 1970.

   *Comment.* In this century, there has been a remarkable and steady increase in the educational level of the population. In 1940, only 25% of the population age 25+ had finished high school. By 1993, this percentage was up to 80%, and still climbing. In that year, about 7% of the population age 25+ had completed a master's degree or better. In 2005, about 85% of the population age 25+ had a high school degree, and 9% had a master's degree or better.

4. Went up.

Set C, page 41

1. 15% per $100.

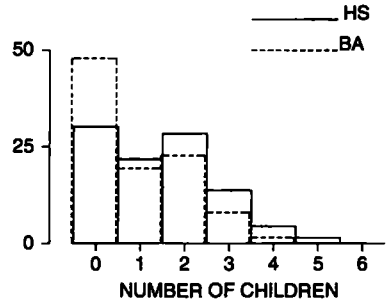
2. Option (ii) is the answer, because (i) doesn't have units, and (iii) has the wrong units for density.

3. 1,750, 2,000, 1, 0.5. The idea on density: If you spread 10 percent evenly over 1 cm = 10 mm, there is 1 percent in each mm, that is, 1 percent per mm.

4. (a) 1.5% per cigarette × 10 cigarettes = 15%.
   (b) 30% (c) 30% + 20% = 50% (d) 10% (e) 3.5%

Set D, page 44

1. (a) qualitative
   (b) qualitative
   (c) quantitative, continuous
   (d) quantitative, continuous
   (e) quantitative, discrete
2. (a) Number of children is a discrete variable.
   (b)




   (c) Better-educated women have fewer children.

Set E, page 46

1. On the whole, the mothers with four children have higher blood pressures. Causality is not proved, there is the confounding factor of age. The mothers with four children are older. (After controlling the age, the Drug Study found there was no association left between number of children and blood pressure.)
2. Left: adds 10 mm  Right: adds 10%

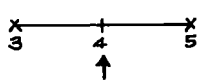
Set F, page 48

1. (a) 7%  (b) 5%
   (c) The users tend to have higher blood pressures.
2. Use of the pill is associated with an increase in blood pressure of several mm.
3. The younger women have slightly higher blood pressures.
   *Comment.* This is a definite anomaly. Most U.S. studies show that systolic blood pressure goes up with age. By comparison, the younger women in the Contraceptive Drug Study have blood pressures which are too high, while the older women have blood pressures which are too low. This probably results from bias in the procedure used to measure blood pressures at the multiphasic, which tended to minimize the prevalence of blood pressures above 140 mm.

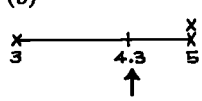
## Chapter 4. The Average and the Standard Deviation

Set A, page 60

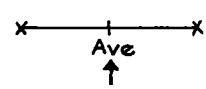
1. (a)




(b)




(c)



*Comment.* With two numbers, the average is half way between. If you add bigger numbers to the list, the average moves up. (Smaller numbers move it down.) The average is always somewhere between the smallest and biggest number on the list.

2. If the average is 1, the list consists of ten 1's. If the average is 3, the list consists of ten 3's. The average cannot be 4: it has to between 1 and 3.
3. The average of (ii) is bigger, it has the large entry 11.
4. $(10\times 66 \text{ inches} + 77 \text{ inches})/11 - 67 \text{ inches} = 5$ feet 7 inches. Or reason this way: the new person is 11 inches taller than the old average. So he adds 11 inches/11 = 1 inch to the average.
5. 5 feet $6\frac{1}{2}$ inches. As the number of people in the room goes up, each additional person has less of an effect on the average.
6. 5 feet 6 inches + 22 inches = 7 feet 4 inches: it's a giraffe.

7. The Rocky Mountains are at the right end, Kansas is around 0 (sea level), and the Marianas trench is at the left end.
8. The conclusion does not follow, the data are cross-sectional not longitudinal. The men with higher diastolic blood pressures are likely to die earlier; they will not be represented in the graph. Furthermore, men with higher blood pressure are more likely to be put on medications that reduce blood pressure.
9. During the recessions, firms tend to lay off the workers with lowest seniority, who are also the lowest paid. This raises the average wage of those left on the payroll. When the recession ends, these low-paid workers are rehired.

   *Comment.* It matters who is included in an average—and who is excluded.

Set B, page 65

1. (a) 50 (b) 25 (c) 40
2. (a) median = average (b) median = average
   (c) median is to the left of the average—long right-hand tail at work.

3. 20
4. The average has to be bigger than the median, so guess 25. (The exact answer is 27.)
5. The average: long right-hand tail.
6. (a) 1 (b) 10 (c) 5 (d) 5
   ("Size" means, neglecting signs.)

Set C, page 67

1. (a) average = 0, r.m.s. size = 4
   (b) average = 0, r.m.s. size = 10.
   On the whole, the numbers in list (b) are bigger in size.
2. (a) 10 (to one decimal place, the exact answer is 9.0).
   (b) 20 (to one decimal place, the exact answer is 19.8).
   (c) 1 (to one decimal place, the exact answer is 1.3).
   The average of the lists is 0; the r.m.s. operation wipes out the signs.
3. For both lists, it's 7; all the entries have the same size, 7.
4. The r.m.s. size is 3.2.
5. The r.m.s. size is 3.1.
   *Comment.* The r.m.s. in exercise 5 is smaller than in exercise 4. There is a reason. Suppose we are going to compare each number on a list to some common value. The r.m.s. size of the amounts off depends on this value. For some values the r.m.s. is larger, for others the r.m.s. is smaller. When is the r.m.s. smallest? It can be proved mathematically that the r.m.s. size of the amounts off is smallest for the average.
6. The errors are way bigger than 3.6, which is supposed to be the r.m.s. size. Something is wrong with the computer.

Set D, page 70

1. (a) 170 cm is 24 cm above average, the SD is 8 cm, so 24 cm represents 3 SDs.
   (b) 2 cm is 0.25 SDs.
   (c) $1.5 \times 8 = 12$ cm, the boy is $146 - 12 = 134$ cm tall.
   (d) shortest, $146 - 18 = 128$ cm; tallest, $146 + 18 = 164$ cm.
2. (a) 150 cm—about average; 4 cm is only 0.5 SDs.
   130 cm—unusually short; 16 cm is 2 SDs.
   165 cm—unusually tall.
   140 cm—about average.
   (b) About 68% were in the range 138 to 154 cm (ave $\pm$ 1 SD), and 95% were in the range 130 to 162 cm (ave $\pm$ 2 SD).
3. biggest, (iii); smallest, (ii).
   *Comment.* All three lists have the same average of 50 and the same range, 0 to 100. But in list (iii), more of the numbers are further away from 50. In list (ii), more of the numbers are closer to 50. There is more to "spread" than the range.
4. (a) 1, since all deviations from the average of 50 are $\pm 1$.
   (b) 2 (c) 2 (d) 2 (e) 10

*Comment.* The SD says how far off average the entries are, on the whole. Just ask yourself whether the amounts off are on the whole more like 1, 2, or 10 in size.

5. 25 years. The average is maybe 30 years, so if 5 years were the answer, many people would be 4 SDs away from the average; with 50 years, everybody would be within 1 SD of the average.
6. (a) (i) (b) (ii) (c) (v)
7. In trial (i), something went wrong: the treatment group is much heavier than the control group. (See exercise 7 on p. 22.)
8. The averages and SDs should be about the same, but the investigator with the bigger sample is likely to get the tallest man, as well as the shortest. The bigger the sample, the bigger the range. The SD and the range measure different things.
9. Guess the average, 69 inches. You have about 1/3 of a chance to be off by more one SD, which is 3 inches.
10. 3 inches. The SD is the r.m.s. deviation from average.

Set E, page 72

1. The SD of (ii) is larger; in fact, the SD of (i) is 1, the SD of (ii) is 2.
2. No, the SD is different from the average absolute deviation, so the method is wrong.
3. No, the 0 does count, so the method is wrong.
4. (a) All three classes have the same average, 50.
   (b) Class B has the biggest SD; there are more students far away from average.
   (c) All three classes have the same range. There is more to spread than the range; see exercise 3 on p. 70.
5. (a) (i) average = 4; deviations = −3, −1, 0, 1, 3; SD = 2.
   (ii) average = 9; deviations = −3, −1, 0, 1, 3; SD = 2.
   (b) List (ii) is obtained from list (i) by adding 5 to each entry. This adds 5 to the average, but does not affect the deviations from the average. So, it does not affect the SD. Adding the same number to each entry on a list does not affect the SD.
6. (a) (i) average = 4; deviations = −3, −1, 0, 1, 3; SD = 2.
   (ii) average = 12; deviations = −9, −3, 0, 3, 9; SD = 6.
   (b) List (ii) is obtained from list (i) by multiplying each entry by 3. This multiplies the average by 3. It also multiplies the deviations from the average by a factor of 3, so it multiplies the SD by a factor of 3. Multiplying each entry on a list by the same positive number just multiplies the SD by that number.
7. (a) (i) average = 2; deviations = 3, −6, 1, −3, 5; SD = 4.
   (ii) average = −2; deviations = −3, 6, −1, 3, −5; SD = 4.
   (b) List (ii) is obtained from list (i) by changing the sign of each entry. This changes the sign of the average and all the deviations from the average, but does not affect the SD.
8. (a) This would increase the average by $250 but leave the SD alone.
   (b) This would increase the average and SD by 5%.
9. The r.m.s. size is 17, and the SD is 0.

10. The SD is much smaller than the r.m.s. size. See p. 72.

11. No.

12. Yes; for instance, the list 1, 1, 16 has an average of 6 and an SD of about 7.

## Chapter 5. The Normal Approximation for Data

Set A, page 82

1. (a) 60 is 10 above average; that's 1 SD. So 60 is +1 in standard units. Similarly, 45 is −0.5 and 75 is +2.5.
   (b) 0 corresponds to the average, 50. The score which is 1.5 in standard units is 1.5 SDs above average; that's $1.5 \times 10 = 15$ points above average, or 65 points. The score 22 is −2.8 in standard units.

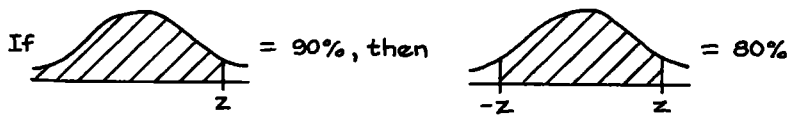
2. The average is 10; the SD is 2.
   (a) In standard units, the list is +1.5, −0.5, +0.5, −1.5, 0.
   (b) The converted list has an average of 0 and an SD of 1. (This is always so: when converted to standard units, any list will average out to 0 and the SD will be 1.)

Set B, page 84

1. (a) 11% (b) 34% (c) 79%
   (d) 25% (e) 43% (f) 13%

2. (a) 1 (b) 1.15

3. (a) 1.65
   (b) 1.30. It's NOT the same $z$ as in (a).

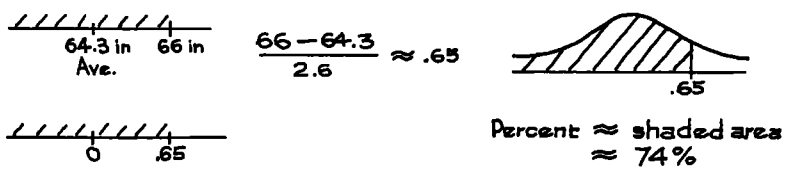



4. (a) 100% − 39% = 61%.
   (b) impossible without further information

5. (a) 58% ÷ 2 = 29% (b) 50% − 29% = 21%.
   (c) impossible without further information.

Set C, page 88

1. (a)




   (b) 69% (c) 0.2 of 1%.

2. (a) 77% (b) 69%
3. In figure 2, the percentage of women with heights between 61 inches and 66 inches is exactly equal to the area under the histogram and approximately equal to the area under the normal curve.

Set D, page 89

1. (a) 75% (b) \$29,000
   (c) 75%. Reason: 90% − 10% = 80% are in the range \$15,000 to \$135,000; and \$15,000 to \$125,000 is about the same range but a little smaller.
2. 5, 95.
3. \$7,000.
4. The area to the left of the 25th percentile has to be 25% of the total area, so the 25th percentile must be quite a bit smaller than 25 mm.
5. (a) It has fatter tails.
   (b) The interquartile range is about 15.

Set E, page 92

1. She was 2.15 SDs above average, at the 98th percentile.
2. The score is 0.85 SDs above average, which is $0.85 \times 100 \approx 85$ points above average. That's $535 + 85 = 620$.
3. 2.75 points—0.50 SDs below average.

Set F, page 93

1. (a) The average is

$$\frac{5}{9} \times (98.6 - 32) = 37.0$$

   The SD is

$$\frac{5}{9} \times 0.3 = 0.17$$

   (b) In standard units, the change of scale washes out, so the answer is 1.5.

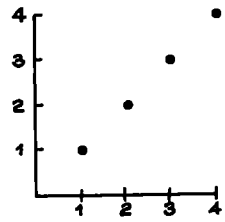
## Chapter 7. Plotting Points and Lines

Set A, page 111

1. A = (1, 2) B = (4, 4) C = (5, 3) D = (5, 1) E = (3, 0).
2. $x$ up by 3, $y$ up by 2.
3. Point D.

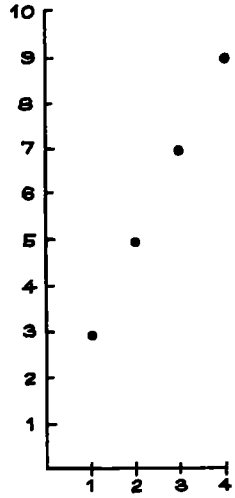
Set B, page 112

1. The four points all lie on a line.




2. The maverick is (1, 2) and it is above the line.
3. The points all lie on a line.

| $x$ | $y$ |
|---|---|
| 1 | 3 |
| 2 | 5 |
| 3 | 7 |
| 4 | 9 |




4. (1, 2) is out; (2, 1) is in.
5. (1, 2) is in; (2, 1) is out.
6. (1, 2) is in; (2, 1) is out.

Set C, page 114

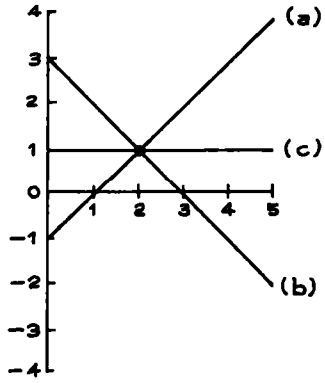
1.

| | *Fig. 16* | *Fig. 17* | *Fig. 18* |
|---|---|---|---|
| Slope | −1/4 in per lb | 5 | 1 |
| Intercept | 1 in | −10 | 0 |

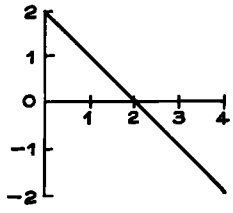
*Note:* In Figure 18, the axes cross at (2, 2).

Set D, page 115

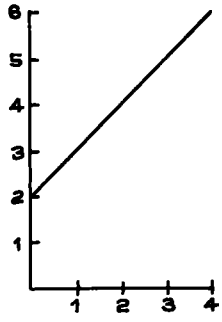
1.




2. On the line.
3. On the line.
4. Above the line.
5.



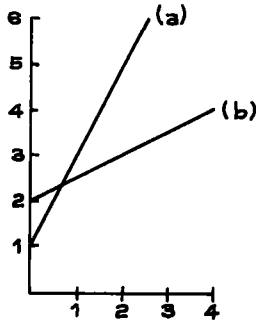
6.



Set E, page 116

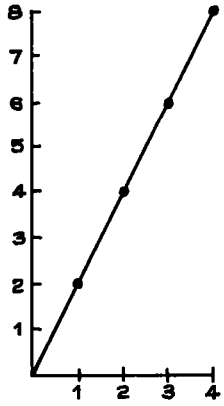
1.

| | *Slope* | *Intercept* | *Height at* $x = 2$ |
|---|---|---|---|
| (a) | 2 | 1 | 5 |
| (b) | 1/2 | 2 | 3 |

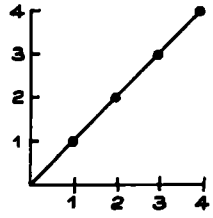


2. (a) $y = \frac{3}{4}x + 1$ (b) $y = -\frac{1}{4}x + 4$ (c) $y = -\frac{1}{2}x + 2$

3. They are all on the line $y = 2x$.




4. They are all on the line $y = x$.




5. (a) on the line. (b) above the line. (c) below the line.

6. All three statements are true. If you understand exercises 4, 5, and 6, you are in good shape for part III.

# Part III. Correlation and Regression

## Chapter 8. Correlation

Set A, page 122

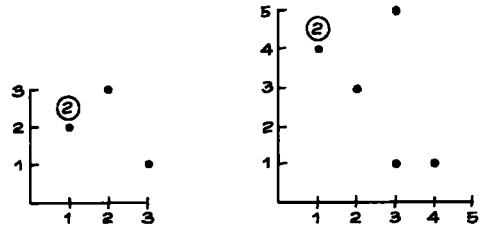
1. (a) shortest father, 59 inches; his son, 65 inches.
   (b) tallest father, 75 inches; his son, 70 inches.
   (c) 76 inches, 64 inches.
   (d) two: 69 inches, 70 inches.
   (e) ave = 68 inches. (f) SD = 3 inches.

2.

| $x$ | $y$ |
|---|---|
| 1 | 4 |
| 2 | 3 |
| 3 | 1 |
| 4 | 1 |
| 4 | 2 |

3. (a) ave $x = 1.5$ (b) SD of $x = 0.5$
   (c) ave $y = 2$ (d) SD of $y = 1.5$

4.




5. (a) A, B, F (b) C, G, H (c) ave $\approx 50$
   (d) SD $\approx 25$ (e) ave $\approx 30$
   (f) False. (g) False, the association is negative.

6. (a) 75 (b) 10 (c) 20
   (d) The final. (e) The final. (f) True.

Set B, page 128

1. (a) Negative. The older the car, the lower the price.
   (b) Negative. The heavier the car, the less efficient.

2. Left: ave $x = 3.0$, SD $x = 1.0$, ave $y = 1.5$, SD $y = 0.5$, positive correlation.
   Right: ave $x = 3.0$, SD $x = 1.0$, ave $y = 1.5$, SD $y = 0.5$, negative correlation.

3. The left hand diagram has correlation closer to 0, it's less like a line.

4. The correlation is about 0.5.

5. The correlation is nearly 0.

   *Comment.* Psychologists call this "attenuation." If you restrict the range of one variable, that usually cuts the correlation down.

6. (a) All the points on the scatter diagram would lie on a line sloping up, so the correlation would be 1.
   (b) Close to 1; this is like part (a), with some noise thrown into the data.

   *Comment.* In the March 2005 Current Population Survey, the correlation between the ages of the husbands and wives was about 0.93; the husbands were, on average, 2.3 years older than their wives.

7. (a) Nearly $-1$: the older you are, the earlier you were born; but there is some fuzz, depending on whether your birthday is before or after the day of the questionnaire.
   (b) Somewhat positive.

8. (a) Somewhat positive. Although wife's income must be less than family income, the two are positively associated.
   (b) Nearly $-1$. If family income is practically constant, the more the wife makes, the less the husband can make.

   *Comment.* In the March 2005 Current Population Survey, the correlation between wife's income and total income was about 0.70. Among families with total income in the range \$80,000–\$90,000, the correlation between husband's income and wife's income was about $-0.98$.

9. False: see p. 126.

Set C, page 131

1. (a) True. (b) False.
2. Dashed.
3. He is one SD above average in height and must weigh $140 + 20 = 160$ pounds.
4. (a) Yes. (b) No. (c) Yes.

Set D, page 134

1. (a) ave of $x = 4$, SD of $x = 2$
   ave of $y = 4$, SD of $y = 2$

| *Standard units* $x$ | $y$ | *Product* |
|---|---|---|
| −1.5 | 1.0 | −1.50 |
| −1.0 | 1.5 | −1.50 |
| −0.5 | 0.5 | −0.25 |
| 0.0 | 0.0 | 0.00 |
| 0.5 | −0.5 | −0.25 |
| 1.0 | −1.5 | −1.50 |
| 1.5 | −1.0 | −1.50 |

$r$ = average of products $\approx -0.93$

   (b) $r = 0.82$, by calculation.
   (c) No calculation is necessary: $r = -1$. The points all lie on a line sloping down, $y = 8 - x$.
2. About 50%.
3. About 25%.
4. About 5%.

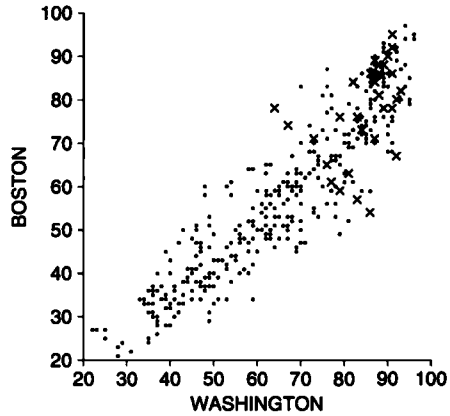
## Chapter 9. More about Correlation

Set A, page 143

1. (a) About the same.
   (b) The maximum has to be bigger than the minimum.
2. No: the correlation between $x$ and $y$ is the same as the correlation between $y$ and $x$.
3. $r$ stays the same.
4. $r$ stays the same.
5. $r$ changes.
6. (a) Up. (b) Down. (c) Reverses the sign.
7. (a) 1 (b) Goes down.
   (c) $r$ will be less than 1—measurement error.
8. The correlation would go down (to about 0.25, in fact).

9. The correlation for the whole year is bigger; for example, it will be very cold in the winter, very hot in the summer—in both cities.

   *Comment.* This is another example of "attenuation" (exercise 5 on p. 130). In the scatter diagram below, the crosses show the data for June 2005 ($r = 0.42$); the dots show the data for days in other months; the correlation for all 365 days is 0.92. Focusing on June restricts the range of the temperatures, and attenuates (weakens) the correlation.




10. Data set (iii) is the same as (ii), with $x$ and $y$ switched; so $r$ is 0.7857. Data set (iv) comes from (i), by adding 1 to each $x$-value, so $r$ is 0.8571. Data set (v) comes from (i) by doubling each y-value, so $r$ is 0.8571 too. Data set (vi) comes from (ii) by subtracting 1 from each $x$-value, and multiplying each $y$-value by 3, so $r$ is 0.7857.

## Set B, page 145

1. Each diagram separately has correlation near 0.6. But all together, things look much more like a line, and the correlation is closer to 0.9—this is attenuation in reverse.
2. Somewhat more than 0.67. This is like the previous exercise: when you put all the children together, the data are much more linear. Also see exercise 9 on p. 144.
3. Yes; the only difference is a change of scale.
4. Yes; it's like any of the diagrams in the previous exercise, so $r \approx 0.7$.

## Set C, page 148

1. (i) should be summarized using $r$, (ii) and (iii) should not.
2. False: like diagram (iii) in exercise 1.
3. Nearly 1. There is a strong association, but the relationship is quadratic not linear, so the correlation cannot be 1.
4. Both are false. You need to look at the scatter diagram to check for outliers or non-linearity.

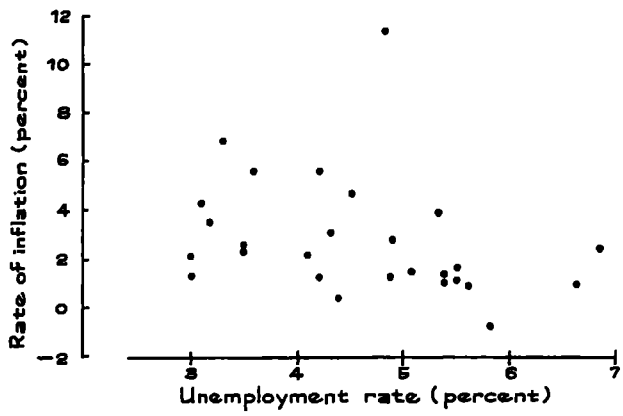
Set D, page 149

1. (a) Diagram is not given. (b) True.
   (c) This cannot be determined from the data (but is true by other studies).
2. No. This correlation might well exaggerate the strength of the relationship—it's based on rates.

Set E, page 152

1. Duration is only measured to the nearest 2 million years; this variable is not easy to determine very accurately.
2. Yes, and this would exaggerate the strength of the association.
3. (a) True. (b) True. (c) True. (d) False.
   Moral: association is not the same as causation.
4. Probably, but this doesn't follow from the data. It could be, for example, that people who have trouble reading watch more television—so causality runs in the other direction. After all, the correlation between $x$ and $y$ equals the correlation between $y$ and $x$.
5. The best explanation is the association between coffee drinking and cigarette smoking. Coffee drinkers are likelier to smoke, smoking causes heart trouble.
6. This is an observational study, not a controlled experiment, and plotting points from the fifties or seventies on the graph just makes a mud pie.

The Phillips "Curve" for the period 1949–74.




*Source: Economic Report of the President*, 1975.

## Chapter 10. Regression

Set A, page 161

1. (a) 67.5 (b) 45 (c) 60
   Work for (a). A score of 75 is 1 SD above average. However, $r$ is only 0.5. If you take the students who are 1 SD above average on the midterm, their average score on the

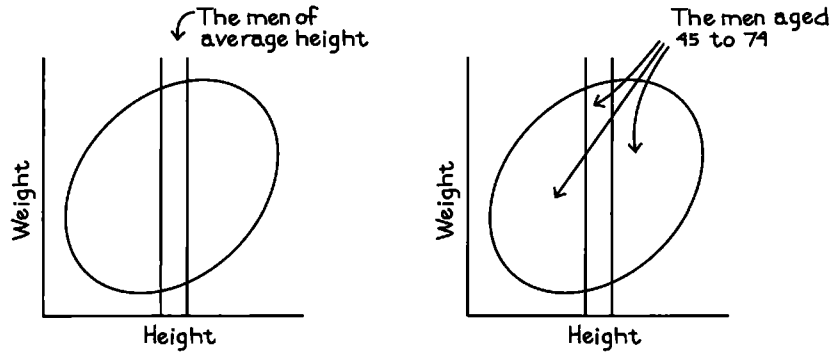
final will only be about 0.5 SDs above average on the final, that is, $0.5 \times 15 = 7.5$ points. So, the estimated average score on the final for this group is $60 + 7.5 = 67.5$.

*Comment.* The regression estimates always lie on a line—the regression line. More about this in chapter 12.

2. (a) 190 pounds (b) 173 pounds
   (c) −68 pounds (d) −206 pounds.

   *Comment on (c).* This is getting ridiculous, but the Public Health Service didn't run into any little men 2 feet tall, so the regression line doesn't pay much attention to this possibility. The regression line should be trusted less and less the further away it gets from the center of the scatter diagram.

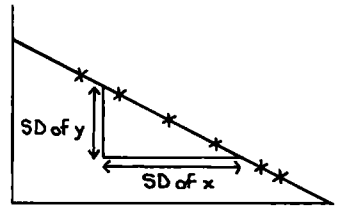
3. False. Think of the scatter diagram for the heights and weights of all the men. Take a vertical strip over 69 inches, representing all the men whose height was just about average. Their average weight should be just about the overall average. But the men aged 45–74 are represented by a different collection of points, some of which are in the strip, and many of which aren't. The regression line says how average weight depends on height, not age. (The older men actually weigh a little more than average—middle-age spread has set in.)




4. These women have completed 12 years of schooling, which is 2 years below average. They are $2/2.4 \approx 0.83$ SDs below average in schooling. The estimate is that they are below average in income, but not by 0.83 SDs—only by $r \times 0.83 \approx 0.28$ SDs of income. In dollars, that's $0.28 \times \$26{,}000 \approx \$7{,}300$. Their average income is estimated as

$$\text{overall average} - \$7{,}300 = \$32{,}000 - \$7{,}300 = \$24{,}700.$$

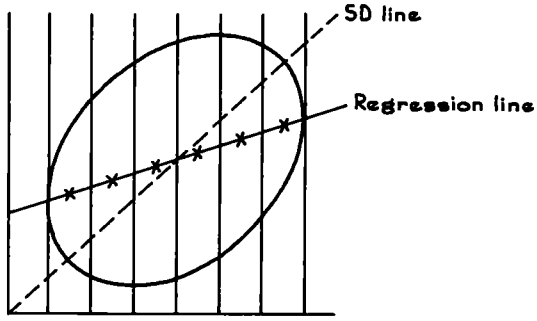
5. The points must all lie on the SD line, which slopes down; the rate is one SD of $y$ per SD of $x$.



Set B, page 163

1. (a) True: the graph of averages slopes upward. Generally, men with higher incomes have wives with higher incomes. People often choose mates with similar educational levels and family backgrounds, which tends to bring incomes into line as well.
   (b) Chance error. The data are from a sample, and there are only 4 couples behind the dot.
   (c) The regression estimates would be a little too low: the line runs below the dots.

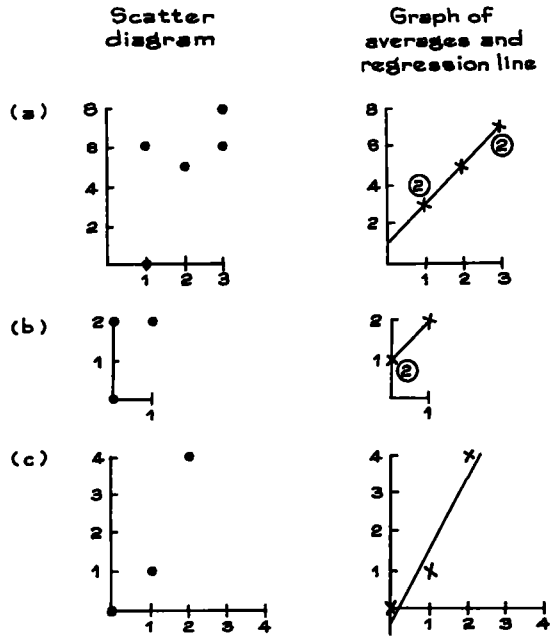
2.



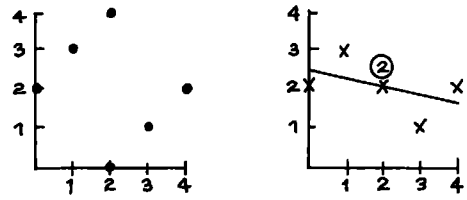

The crosses fall on the solid regression line, the dashed line is the SD line.

3. For the two diagrams on the left, the SD line is dashed and the regression line is solid. For the two on the right, the SD line is solid and the regression line is dashed. Moral: the regression line isn't as steep as the SD line.

4.



4. (d)

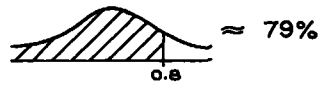



Set C, page 167

1. (a) 67.5 (b) 45 (c) 60 (d) 60
   This exercise is about individuals; exercise 1 on p. 161 was about groups. The arithmetic for parts (a–c) is the same; pp. 165–66.

2. (a) 79% (b) 38% (c) 50% (d) 50%
   Work for (a):

   In standard units, his SAT score was 1.3. The regression prediction for his first-year score is $0.6 \times 1.3 \approx 0.8$ in standard units.




   This corresponds to a percentile rank of 79%. In example 2, the predicted percentile rank was only 69%, which is closer to 50%. That is because the correlation was lower in example 2. There is more regression to the mean in example 2.

3. (a) The SD line—dashed. (b) The regression line—solid.

4. (a) There is a minimum age for marriage.
   (b) Age is reported as a whole year; there are a lot of husbands age 30, but none aged 30.33; likewise for the wives.

5. False. The regression line says how average weight depends on height, not on age. See exercise 3 on p. 161.

Set D, page 174

1. No, this looks like the regression effect. Imagine a controlled experiment. At one airport, the instructors discuss the ratings with the pilots. At another, the instructors keep the ratings to themselves. Even at the second airport, the ratings on the two landings will not be identical—differences come in. So the regression effect appears: on the average, the bottom group improves a bit, and the top group falls back. That is probably all the air force saw in their data.

2. No. It looks like the tutoring had an effect—regression would only take them closer to the average, but they got to the other side.

3. The sons of the 61-inch fathers are taller, on the average, than the sons of the 62-

inch fathers. This is just chance variation. By the luck of the draw, Pearson got too many families where the father was 61 inches tall and the son was extra tall.

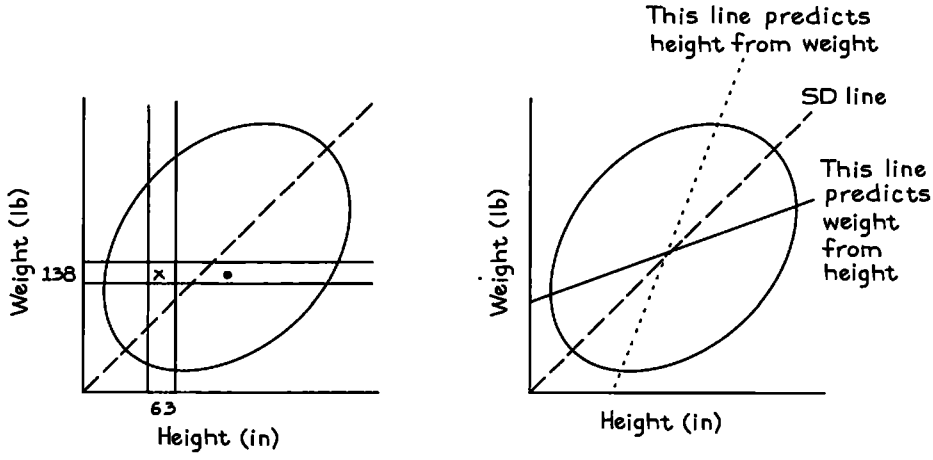
*Comment.* There were only 8 families where the father was about 61 inches tall, and 15 where the father was 62 inches—lots of room for chance error.

Set E, page 175

1. False. There are two completely different groups of men here. (See the diagram below.) The ones who are 63 inches tall are in the vertical strip. They average 138 pounds in weight, as shown by the cross. The ones who weighed 138 pounds are in the horizontal strip. Their average height is shown by a heavy dot, and it's a lot more than 63 inches.

   Remember, there are two regression lines—

   - one for weight on height,
   - one for height on weight.




2. False. The fathers only average 69 inches; you have to use the other line.
3. False. This is just like exercises 1 and 2. (A typical student at the 69th percentile of the first-year tests should be at the 58th percentile on the SAT; use the other line.)

## Chapter 11. The R.M.S. Error for Regression

Set A, page 184

1. B is <u>tall</u> and <u>chubby</u>, while D is <u>short</u> and <u>skinny</u>.
2. (a) False. (b) True.
3. Prediction errors = −7, 1, 3, −1, 4; r.m.s. error = 3.9.
4. (a) 0.2 (b) 1 (c) 5.
5. A few thousand dollars.
6. The one with the smaller r.m.s. error should be used, as it will be more accurate overall.

7. (a) 8 points—one r.m.s. error. (b) 16 points—two r.m.s. errors.
8. (a) \$20,000. (b) The horizontal line. See p. 183.

Set B, page 187

1. $\sqrt{1-0.6^2} \times 10 = 8$ points.
2. (a) Guess the average, 65.
   (b) 10. If you use the regression line, the r.m.s. error is given by the formula (exercise 1). If you use the average, the r.m.s. error is the SD. (See exercises 9–10 on p. 71.)
   (c) Use the regression line, and the r.m.s. error is given by the formula as 8 points (exercise 1).
3. Generally, it helps to have more information. The r.m.s. error will be smaller for person B, by the factor $\sqrt{1-0.6^2} = 0.8$. See p. 186.

Set C, page 189

1. (a) (iii) (b) (ii) (c) (i)
2. (a) (i) (ii) (b) not used (c) (iii)
3. (a) SD of $y \approx 1$
   (b) SD of residuals $\approx 0.6$
   (c) SD of $y$ in strip $\approx 0.6$, about the same as the SD of the residuals.

   *Comment.* The vertical scatter in the strip is about the same as the r.m.s. error of the regression line—but the vertical scatter in the whole diagram is a lot more than the vertical scatter in the strip.

Set D, page 193

1. (a) True.
   (b) True; the scatter diagram is homoscedastic, so the subjects are off the regression line by similar amounts in each vertical strip.
   (c) False, because the scatter diagram is heteroscedastic; 9 points is a sort of average amount off, but the prediction errors are going to be bigger with high scores.
2. (a) $\sqrt{1-0.5^2} \times 2.7 \approx 2.3$ inches.
   (b) 71 inches—regression method.
   (c) 2.3 inches. The scatter diagram is homoscedastic, so the sons' heights are off the regression line by similar amounts, for any father's height. The amount off is the r.m.s. error of the line.
   (d) The prediction is 68 inches, and it is likely to be off by 2.3 inches or so.
3. (a) $\sqrt{1-0.37^2} \times \$20{,}000 \approx \$18{,}600$.
   (b) \$24,500—regression method.
   (c) This cannot be determined from the information given. The \$18,600 is sort of the average amount off the line. But the scatter diagram is heteroscedastic, so the amount off the line changes from strip to strip. The spread in incomes is larger for more highly educated people, so the amount off will be larger than \$18,600.
   (d) The prediction is \$7,100. The amount off cannot be determined, but will be less than \$18,600.

4. The husband is between 20 and 30 years of age.

5. (a) 50, 15 (b) 50, 15 (c) 0.95 (d) 25, 5
   (e) 0.5—attenuation. See exercise 9 on p. 144 and exercises 1–2 on pp. 145–146.

6. (a) The SD for all the wives is much bigger. That is the main point of exercises 4–6. See the comments below.
   (b) The two SDs are about the same.

   *Comments.* If you just take the families where the husband is 20 to 30 years of age, the wives are going to be much more similar in age, their SD drops from about 15 years to about 5 years. If you take the husbands born in March, that does not cut down the variability in the ages of their wives. Smaller samples do not generally have smaller SDs (exercise 8 on p. 71). But if you restrict the range of $x$, that will generally reduce the SD of $y$.

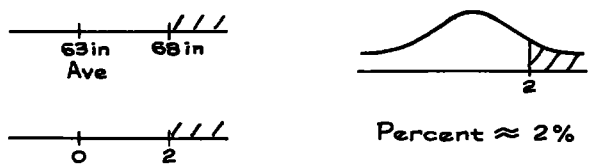
7. (a) 68 inches, the average.
   (b) 3 inches, the SD.
   (c) Regression. If one twin is 6 ft 6 in, guess 6 ft $5\frac{1}{2}$ in for the other one.
   (d) $\sqrt{1-0.95^2} \times 3 \approx 0.9$ inches.

   *Comments.* (i) If $r = 1$, you should guess that the height of the second twin equals the height of the first one. But $r$ is a little less than 1. So you regress the second twin back toward the mean—a little bit.

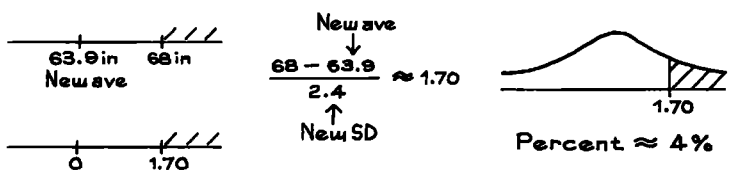
   (ii) The answer to (d) is quite a bit smaller than the answer to (b). When $r = 0.95$, there is quite a large reduction in r.m.s. error when you use the regression line.

Set E, page 197

1. (a)




   (b) new average ≈ 63.9 inches, new SD ≈ 2.4 inches




2. (a) 14% (b) 33%

3. (a) 38% (b) 60%

## Chapter 12. The Regression Line

Set A, page 207

1. (a) \$2,000 × 8 + \$5,000 = \$21,000
   (b) \$2,000 × 12 + \$5,000 = \$29,000
   (c) \$2,000 × 16 + \$5,000 = \$37,000

2. (a) 240 ounces = 15 pounds (b) 20 ounces.
   (c) 3 ounces of nitrogen yields 18 lb 12 oz of rice, 4 ounces of nitrogen yields 20 pounds of rice.
   (d) Controlled.
   (e) Yes. The line fits quite well ($r = 0.95$), and 3 ounces is close to a value that was used.
   (f) No. That's too far away from the amounts used.
3. (a) Predicted son's height = 0.5 × father's height + 35 inches.
   (b) Predicted father's height = 0.5 × son's height + 33.5 inches.

   *Comment.* There are two regression lines, one predicts son's height from father's height, the other predicts father's height from son's height (section 5 of chapter 10).
4. This testimony is overstatement. Associations in the data may be due to confounding. Without doing the experiment, or working very hard at the observational data, you can't be sure what the impact of interventions will be.

Set B, page 210

1. With 12 years of education, height is predicted as 69.75 inches; with 16 years, height is predicted as 70.75 inches. Going to college clearly has no effect on height. This observational study picked up a correlation between height and education due to some third factor in family background.
2. 439.16 cm, 439.26 cm. Hanging a bigger weight on the wire makes it stretch more. You can trust the regression line in exercise 2 because it is based on an experiment. In exercise 1, the line was fitted to data from an observational study.
3. (a) $540 + 110 = 650$ (b) 540 (c) Greater than (p. 208).
4. (a) 540 (b) 540 (c) Greater than (p. 208).

   *Comment.* if you use the average value of $y$ to predict $y$, the r.m.s. error is the SD of $y$; see p. 183.
5. The regression line makes the smallest r.m.s. error (p. 208).

# Part IV. Probability

## Chapter 13. What Are the Chances?

Set A, page 225

1. (a) (vi) (b) (iii) (c) (iv) (d) (i)
   (e) (ii) (f) (v) (g) (vi)
2. About 500.
3. About 1,000.
4. About 14.
5. Box (ii), because $\boxed{3}$ pays more than $\boxed{2}$, and the other ticket is the same.

Set B, page 227

1. (a) The question is about the second ticket, not the first: see part (a) of example 2. The answer is 1/4.

(b) 1/3; there are 3 tickets left after $\boxed{2}$ is drawn.

2. (a) 1/4 (b) 1/4
   With replacement, the box stays the same.

3. (a) 1/2 (b) 1/2
   The chances for the 5th toss of the penny do not depend on the results of the first 4 tosses.

4. (a) 1/52 (b) 1/48
   This is like example 2 on p. 226.

Set C, page 229

1. (a) 12/51 (b) $13/52 \times 12/51 = 1/17 \approx 6\%$.

2. (a) 1/6 (b) $1/6 \times 1/6 \times 1/6 = 1/216 \approx 1/2$ of 1%.

3. (a) 4/52 (b) $4/52 \times 4/51 \times 4/50 \approx 5/10{,}000$.
   *Comment.* In this exercise, the cards are dependent; in exercise 2, the rolls were independent.

4. "At least one ace" is the better option: you would choose an exam in which you had to get at least one question right out of six, over an exam in which you had to get all six right.

5. This is fine, it's the multiplication rule.

6. The coin has to land "tails, heads"; the chance is 1/4.

7. (a) 1/8
   (b) $1 - 1/8 = 7/8$
   (c) 7/8; you get at least one tail when you don't get three heads: so (b) and (c) are the same.
   (d) 7/8; just switch heads and tails in (c).

Set D, page 232

1. (a) independent: if you get a white ticket, there is 1 chance in 3 to get "1" and 2 chances in 3 to get "2"; if you get the black ticket, the chances for the numbers stay the same.
   (b) independent
   (c) dependent: with the white tickets, there is only 1 chance in 3 to get "2"; with the black tickets, there are 2 chances in 3.

2. (a,b) independent (c) dependent
   *Comment.* This kind of box will come up again in chapter 27. Here is the argument for (a). Suppose you draw a ticket, and see the first number is 4 but don't see the second number: the chance that the second number will be 3 is 1/2. Likewise if the first number is 1. That is independence.

3. Ten years is 520 weeks, so the chance is $(999{,}999/1{,}000{,}000)^{520} \approx 0.9995$.
   *Comment.* In the New York State Lotto, your chance of winning something is about 1/12,000,000.

4. This is false. It's like saying someone doesn't have a temperature because you can't find the thermometer. To figure out whether two things are independent or not, you pretend to know how the first one turned out, and then see if the chances for the second change. The emphasis is on the word "pretend."

5. (a) 5% (b) 20%

   To figure (a) out, suppose you have 80 men and 20 women in the class. You also have 15 cards marked "freshman" and 85 cards marked "sophomore." You want to give out a card to each student, so that as few women as possible get "sophomore." The strategy is to give a sophomore card to each man; you are left with 5, which have to go to 5 women. The 15 freshman cards go to the other 15 women.

   *Comment.* If year and sex are independent, the percentage of sophomore women would be 85% of 20% — 17%, between the two extremes.

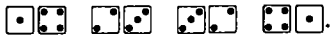
6. Same as previous exercise: the chance of getting a sophomore woman equals the percentage of sophomore women in the class.

7. False. The calculation assumes that the percentage of women is the same across all age groups, and it isn't: women live longer than men. (Actually, women age 85 and over accounted for nearly 1.1% of the U.S. population in 2002.)

8. If the subject draws the ace of spades from the small pile, he has 13 chances in 52 to draw a spade from the big deck, and win the prize. Likewise if he draws the deuce of clubs. Or any other card. So the answer is $13/52 = 1/4$.

## Chapter 14. More about Chance

Set A, page 240

1.  . The chance is 4/36.

2. There are 25 possible results; for 5 of them, the sum is 6. So the chance is 5/25. (The figure is not shown.)

3. Most often, 7; least often, 2, 12. (Use figure 1 to get the chance of each total, as in exercise 1.)

4. (a) 2/4 (b) 2/6 (c) 3/6

Set B, page 242

1. False. The question is about the number of children who had either cookies or ice cream, including the gluttons who had both. The number depends on the choices made by the children, and two possibilities are shown in the table.

| *Cookies only* | *Ice cream only* | *Both* | *Neither* |
|---|---|---|---|
| 12 | 17 | 0 | 21 |
| 3 | 8 | 9 | 30 |

   In the first case, 12 children had cookies only, 17 children had ice cream only, 0 had both, and 21 had neither. So $12 + 17 = 29$ had cookies or ice cream. The second line shows another possibility, where 9 children had both cookies and ice cream. In this situation, the number with cookies or ice cream is $3 + 8 + 9 = 20$. Just as a check: the number with cookies is $3 + 9 = 12$, and the number with ice cream is $8 + 9 = 17$, as given in the problem. But the number with cookies or ice cream is not $12 + 17$, because the addition double counts the 9 gluttons. The number who had cookies or ice cream depends on the number of gluttons who had both.

2. (a) 4/20 (b) 8/20 (c) 12/20 (d) 14/20

   *Comment.* (4 + 8 + 12)/20 gives the wrong answer to (d)—by double-counting some dots and triple-counting others.

3. They are the same.

4. False. Simply adding the two chances double counts the chance of [•][⁚]. See example 5 on p. 242.

5. False. There is 1 chance in 10 of getting [7] on any particular draw, but these events are not mutually exclusive.

6. True. 100% − (10% + 20%) = 70%. Use the addition rule, and p. 223 for the subtraction.

## Set C, page 246

1. (a) 1/52 of the contestants step forward.
   (b) 1/52 of the contestants step forward; example 2 in chapter 13.
   (c) The ones who got both the ace of hearts on the first card and the king of hearts on the second card step forward twice. (In terms of getting the weekend, that's overkill.) The fraction who step forward twice is 1/52 × 1/51.
   (d) False; as (c) shows, the events aren't mutually exclusive, so addition double counts the chance that both occur.

   *Comment.* The chance in (d) is

$$1/52 + 1/52 - 1/52 \times 1/51.$$

2. (a) 1/52 of the contestants step forward.
   (b) 1/52 of the contestants step forward.
   (c) If you get the ace of hearts on the first card, you can't get it on the second card; nobody steps forward twice.
   (d) True; as (c) shows, the events are mutually exclusive, so addition is legitimate.

   *Comment.* In exercise 2, the two ways to win are mutually exclusive; not so in exercise 1. Addition is legitimate in exercise 2, not in 1.

3. (a,b) True; see example 2 in chapter 13.
   (c) False. "Top card is the jack of clubs" and "bottom card is the jack of diamonds" aren't mutually exclusive, so you can't add the chances.
   (d) True. "Top card is the jack of clubs" and "bottom card is the jack of clubs" are mutually exclusive.
   (e,f) False; these events aren't independent, you need the conditional chances.

4. (a) False; $1/2 \times 1/3 = 1/6$, but A and B may be dependent: you need the conditional chance of B given A.
   (b) True; see section 4 of chapter 13.
   (c) False. ("Mutually exclusive" implies dependence, and the chance is actually 0.)
   (d) False; $1/2 + 1/3 = 5/6$, but you can't add the chances because A and B may not be mutually exclusive.
   (e) False; if they're independent, they have some chance of happening together, so they can't be mutually exclusive: don't add the chances.
   (f) True.

   *Comment.* If you have trouble with exercises 3 and 4, look at example 6, p. 244.

5. See example 2 in chapter 13.
   (a) 4/52 (b) 4/51 (c) $4/52 \times 4/51$

Set D, page 250

1. (a) (i) (b) (i) (ii)
   (c) (iii) (d) (ii) (iii)
   (e) (i) (ii) (f) (i)

2. Bets (a) and (f) say the same thing in different language. So do (b) and (e). Bet (d) is better than (c).

3. (a) 3/4 (b) 3/4 (c) 9/16 (d) 9/16 (e) $1 - 9/16 = 7/16$

4. (a) Chance of no aces $= (5/6)^3 \approx 58\%$, so chance of at least one ace $\approx 42\%$. Like de Méré, with 3 rolls instead of 4.
   (b) 67% (c) 89%

5. $1 - (35/36)^{36} \approx 64\%$

6. The chance that the point 17 will not come up in 22 throws is $(31/32)^{22} \approx 49.7\%$. The chance that it will come up in 22 throws is therefore $100\% - 49.7\% = 50.3\%$. So this wager (laid at even money) was also favorable to the Master of the Ball. Poor Adventurers.

7. The chance of surviving 50 missions is $(0.98)^{50} \approx 36\%$. Deighton is adding chances for events that are not mutually exclusive.

## Chapter 15. The Binomial Coefficients

Set A, page 258

1. The number is 4.

2. The number is 6.

3. (a) $(5/6)^4 = 625/1{,}296 \approx 48\%$
   (b) $4(1/6)(5/6)^3 = 500/1{,}296 \approx 39\%$
   (c) $6(1/6)^2(5/6)^2 = 150/1{,}296 \approx 12\%$
   (d) $4(1/6)^3(5/6) = 20/1{,}296 \approx 1.5\%$
   (e) $(1/6)^4 = 1/1{,}296 \approx 0.08$ of 1%
   (f) Addition rule: $(150 + 20 + 1)/1{,}296 \approx 13\%$.

4. This is the same as exercise 3(a–c). Rolling an ace is like drawing a red marble, while 2 through 6 correspond to green. To see why, imagine two people, A and B, performing different chance experiments:

- A rolls a die four times and counts the number of aces.
- B draws four times at random with replacement from the box $\boxed{R\,G\,G\,G\,G\,G}$ and counts the number of R's.

The equipment is different, but as far as the chance of getting any particular number of reds is concerned, the two experiments are equivalent.

- There are four rolls, just as there are four draws.
- The rolls are independent; so are the draws.
- Each roll has 1 chance in 6 to contribute one to the count (ace); similarly for each draw (red).

5. The chance of getting exactly 5 heads is $\frac{10!}{5!\,5!}\left(\frac{1}{2}\right)^{10} = \frac{252}{1{,}024} \approx 25\%$. The chance of getting exactly 4 heads is $\frac{10!}{4!\,6!}\left(\frac{1}{2}\right)^{10} = \frac{210}{1{,}024} \approx 21\%$. The chance of getting exactly 6 heads is the same. By the addition rule, the chance of getting 4 through 6 heads is $672/1{,}024 \approx 66\%$.

6. You need the chance of getting 7, 8, 9, or 10 heads when a coin is tossed 10 times. Use the binomial formula, and the addition rule:

$$\frac{10!}{7!\,3!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{8!\,2!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{9!\,1!}\left(\frac{1}{2}\right)^{10} + \frac{10!}{10!\,0!}\left(\frac{1}{2}\right)^{10} = \frac{176}{1{,}024} \approx 17\%.$$

*Comment.* Looks like chance, not vitamins.

# Part V. Chance Variability

## Chapter 16. The Law of Averages

Set A, page 277

1. The error is 50 in absolute terms, 5% in percentage terms.
2. The error is 1,000 in absolute terms, 1/10 of 1% in percentage terms. Compare this with the previous exercise: the chance error has gone up in absolute terms (from 50 to 1,000) but down in percentage terms (from 5% to 1/10 of 1%).
3. False. The chance stays at 50%. See p. 274.
4. (a) Ten tosses. As the number of tosses goes up, you are more and more likely to be close to 50% heads, less and less likely to be above 60% heads. Here, chance variability in the percentages helps you, a small number of tosses is better than a large number.
   (b) One hundred tosses: now chance variability in the percentages hurts you—because you want to be close to 50%. With more tosses, there is less chance variability in the percentages. More tosses are better.
   (c) One hundred tosses; like (b).
   (d) Ten tosses. As the number of tosses goes up, there is less and less chance for the number of heads to exactly equal the expected number. Let's take a more extreme case: suppose you toss the coin 1,000,000 times. The chance of getting exactly 500,000 heads—rather than 500,001 or 500,043 or 499,997 or some other number close to 500,000—is quite slim.

5. Option (i) is better. This is just like exercise 4(a).
6. Option (ii), the reason is chance error.
7. It's about the same with or without replacement.
8. Same. Both have 50% $\boxed{-1}$'s and 50% $\boxed{+1}$'s.
9. Eventually, the chance error would be large and negative. Then, it would get positive again. In absolute terms, the swings get wilder and wilder.

Set B, page 280

1. $47 \times 1 + 53 \times 2 = 153$.
2. (a) 100, 200 (b) 50, 50 (c) $50 \times 1 + 50 \times 2 = 150$.
3. (a) 100, 900.
   (b) $33 \times 1 + 33 \times 2 + 33 \times 9 \approx 400$.
   *Comment.* 400 isn't halfway between 100 and 900.
4. Guess 500 in all three cases; (iii) is best, (i) worst.
5. The chance for "1" is 1 in 10; the chance for "3 or less" is 3 in 10; the chance for "4 or more" is 7 in 10—there are 7 numbers from 4 through 10 inclusive. Drawing at random from boxes is discussed in chapters 13–14.
6. Box (i) is better, it has fewer $-1$'s, and the same 2.
7. Options (i) and (ii) do it. Your net gain is the sum of your wins and losses, taking signs into account.

Set C, page 284

1. (i) and (ii) are the same. (iii) means that all ten draws must be "1," which is worse than (i).
2. Option (i) is no good; the sum of the draws is unrelated to the net gain. Option (ii) is no good; it says you win \$17 with 2 chances in 36 on a single play, but your chances are 2 in 38. Option (iii) is right. If in doubt, review example 1 on p. 283.
3. Your net gain is like the sum of 10 draws made at random with replacement from the box

   | 1 ticket $\boxed{\$36}$ 215 tickets $\boxed{-\$1}$ |

   This is a terrible game.

## Chapter 17. The Expected Value and Standard Error

Set A, page 290

1. (a) $100 \times 2 = 200$ (b) $-25$ (c) 0 (d) $66\frac{2}{3}$
   *Comment on (d).* The "expected value" need not be one of the possible values. It's like saying that the average family has 2.1 children. This is sensible, even though the "average family" is a statistical fiction.
2. This is the same as the expected value for the sum of two draws from the box $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}\boxed{6}$. So the answer is $2 \times 3.5 = 7$ squares.

3. The model is given on pp. 283–284. The average of the the numbers in the box is
$$(\$35 - \$37)/38 = -\$2/38 \approx -\$0.05$$
(To compute the average, you have to add up the tickets in the box; $\boxed{+\$35}$ adds \$35 to the total, but the 37 $\boxed{-\$1}$'s take \$37 away; then you have to divide by the number of tickets in the box, which is 38.) The expected net gain is equal to $100 \times (-\$.05) = -\$5$. You can expect to lose around \$5.

4. The box is on p. 283. The average of the box is
$$(\$18 - \$20)/38 = -\$2/38 \approx -\$0.05$$
(The average is the total of the numbers in the box, divided by 38; the 18 tickets marked "+\$1" contribute \$18 to the total, while the 20 tickets marked "−\$1" take \$20 away.) The expected net gain is $100 \times (-\$0.05) = -\$5$.

   *Comment.* Exercises 3 and 4 show that with either bet (number or red-or-black), you can expect to lose 1/19 of your stake on each play.

5. −\$50. Moral: the more you play, the more you lose.

6. The average of the box is $(18x - \$20)/38$. To be fair, this has to equal 0. The equation is $18x - \$20 = 0$. So $x \approx \$1.11$. They should pay you \$1.11.

7. The Master of the Ball should have paid 31 pounds, just as the Adventurers thought. Moral: the Adventurers may have the fun, but it is the Master of the Ball who has the profit.

Set B, page 293

1. (a) The average of the box is 4; the SD is 2. So the expected value for the sum is $100 \times 4 = 400$; the SE for the sum is $\sqrt{100} \times 2 = 20$.
   (b) Around 400, give or take 20 or so.
   (c) Guess 400, off by 20 or so. Parts (b) and (c) interpret the numbers in (a).

2. The net gain is like the sum of 100 draws from the box $\boxed{-\$1}\boxed{\$1}$. The average of the box is \$0; the SD is \$1. The sum of 100 draws has expected value \$0; the SE for the sum is $\sqrt{100} \times \$1 = \$10$. So your net gain will be around \$0, give or take \$10 or so.

3. With option (ii), the numbers are too close to 50; no number is more than 5 away. With option (iii), the numbers alternate much too regularly. Option (i) is it.

4. The expected value is 150, the observed value is 157, the chance error is 7, the standard error is 10.

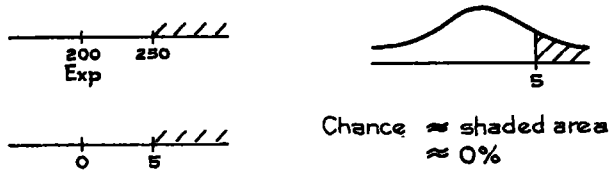
5. Multiplying the number of draws by 4 multiplies the expected value by 4 and the SE by $\sqrt{4} = 2$. The expected value for the sum of 100 draws is $4 \times 50 = 200$, and the SE is $2 \times 10 = 20$.

6. (a) is true, (b) is false: the expected value for the sum of the draws can be computed exactly, as
$$\text{number of draws} \times \text{average of box}$$
(c) is false, (d) is true: the sum will be off its expected value, and the SE tells you by about how much.

7. Yes. The chance is small, but positive. If you wait long enough, events of small probability do happen.

## Set C, page 296

1. (a) Smallest, 100; largest, 400.
   (b) The average of the box is 2; the SD is 1. The sum has an expected value of $100 \times 2 = 200$; the SE for the sum is $\sqrt{100} \times 1 = 10$. The sum will be around 200, give or take 10 or so.
   (c)




2. (a) Largest, 900; smallest, 100. (b) Chance ≈ 68%

3. (a) The expected value is 0, so the sum is around 0, and your best hope is chance variability in the sum—you want the sum to be far from its expected value. Chance variability goes up with the number of draws, choose 100.
   (b) Same as (a).
   (c) Now chance variability in the sum works against you, because you want the sum to be close to its expected value; choose 10.

4. (i) Expected value for sum = 500, SE for sum = 30.
   (ii) Expected value for sum = 500, SE for sum = 20.
   Both sums will be around 500, but sum (i) will be further away. In (a) and (b), chance variability helps—choose (i). In (c), chance variability hurts—choose (ii).

5. 98%.

6. Either they win \$25,000 (with chance $20/38 \approx 53\%$) or they lose \$25,000 (with chance $18/38 \approx 47\%$). The answer is 50%.
   *Comment.* The casino is much happier with a lot of small bets, where the profit is almost guaranteed, than with one big bet, where there is a lot of risk.

7. One number will pay off \$35,000, but the other 37 will lose, so the gambler loses \$2,000 for sure.
   *Comment.* The casino likes the gamblers to spread their bets.

8. Option (ii) is right; the SE doesn't go up by a full factor of 2, but only $\sqrt{2} \approx 1.4$.

## Set D, page 299

1. (a) No, replace the 5 by $7 - (-2) = 9$. (b) Yes. (c) Yes.
   (d) No—the list shows 3 different numbers, so the short-cut doesn't apply.

2. The net gain is like the sum of 100 draws from the box

$$\boxed{\$2}\ \boxed{-\$1}\ \boxed{-\$1}\ \boxed{-\$1}$$

   The average of the box is $(\$2 - \$1 - \$1 - \$1)/4 = -\$0.25$. The SD is

$$[\$2 - (-\$1)] \times \sqrt{1/4 \times 3/4} \approx \$1.30.$$

   The net gain in 100 plays will be around $100 \times (-\$0.25) = -\$25$, give or take $\sqrt{100} \times \$1.30 = \$13$ or so.

3. (a) From the point of view of the house, a dollar bet on the house special is like one draw from the box

| 5 tickets $\boxed{-\$6}$ 33 tickets $\boxed{+\$1}$ |

The average of the box is $[5 \times (-\$6) + 33 \times \$1)]/38 \approx \$.08$. So the house expects to make about 8 cents per dollar bet. As far as the house is concerned, this is a great bet.

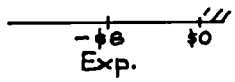
(b) The player's net gain is like the sum of 100 draws at random with replacement from the same box with the signs reversed:

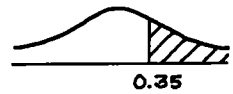
| 5 tickets $\boxed{+\$6}$ 33 tickets $\boxed{-\$1}$ |

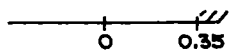
The average of the box is $-\$.08$; the SD is

$$[\$6 - (-\$1)] \times \sqrt{5/38 \times 33/38} \approx \$2.37.$$

The player's expected net gain in 100 plays is $-\$8$, give or take \$24 or so.




Chance ≈ shaded area
≈ 36%

4. The expected net gain in 100 one-dollar bets on a section is $-\$5$; the SE is \$14. The expected net gain in 100 bets on red is $-\$5$; the SE is \$10. Options (i) and (ii) have the same expected net gain. But (i) has the bigger SE, that is, more variability: (a) is false, (b) and (c) are true.

Set E, page 303

1. (a) You can't add up words, so box (i) is out. With box (iii), you get 2 chances in 3 to go up each time, and it should only be 1 in 2. Box (ii) is the one.
   (b) Average of box = 0.5 and SD of box = 0.5 too. The sum of 16 draws has an expected value of $16 \times 0.5 = 8$; the SE is $\sqrt{16} \times 0.5 = 2$. The number of heads will be around 8, give or take 2 or so.
2. New box: $\boxed{0}\boxed{0}\boxed{0}\boxed{0}\boxed{1}$. It's $\pm 3$ SE, chance is about 99.7%.
3. New box: $\boxed{0}\boxed{1}$. It's 1 SE or more, chance is about 16%.
4.

| *Group of 100 tosses* | *Observed value* | *Expected value* | *Chance error* | *Standard error* |
|---|---|---|---|---|
| 1–100 | 44 | 50 | −6 | 5 |
| 101–200 | 54 | 50 | +4 | 5 |
| 201–300 | 48 | 50 | −2 | 5 |
| 301–400 | 53 | 50 | +3 | 5 |

5. Expect about 68—example 5 on p. 301; actually, you see 69.
6. (a,b) About 99.7%—it's 3 SEs.

   *Comment.* When the number of tosses goes up from 10,000 to 1,000,000, the percentage of heads gets closer to 50%: the 99.7%-interval shrinks from

   $$50\% \pm 1.5\% \quad \text{to} \quad 50\% \pm 0.15\%.$$
7. Expected is 30, observed is 33, chance error is 3, SE is about 3.5.
8. Put in five 0's and five 1's. Tell it to draw 1,000 times.
9. It's fine. The number of aces isn't supposed to be 16.67 exactly, it's only supposed to be around 16.67.

## Chapter 18. The Normal Approximation for Probability Histograms

### Set A, page 312

1. Between 70 and 80 inclusive.

2. (a) Between 6.5 and 10.5.
   (b) Between 6.5 and 7.5—the left and right edges of the rectangle over 7.

3. (a) 7
   (b) 7: tallest bar in 2nd panel.
   (c) No, this is just chance variation. In fact 4 is less likely than 5, as the probability histogram in the bottom panel shows.
   (d) (iii). The top panel is an empirical histogram—it shows observed percentages, not chances.

4. (a) 3, 6
   (b) Bottom panel—the probability histogram shows chances. The values 2 and 3 are equally likely for the product.
   (c) Look at the second panel: 3 appeared more often. Chance variation again.
   (d) The value 14 is impossible for the product. Reason: there are only two ways to factor 14, as $1 \times 14$ or $2 \times 7$; no die can show 7 or 14.
   (e) The bottom panel is a probability histogram, so areas under it represent chances: 11.1% is the chance of getting a product of 6 when you roll a pair of dice.

5. A goes with (i) and B with (ii). B is lower, more spread out, and farther to the right. Box (ii) has a bigger average and a bigger SD.

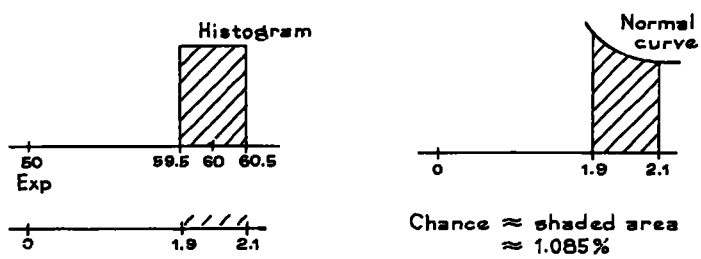
6. False. The probability histogram for the sum tells you the chances for the sum. It doesn't tell you how the draws turned out. The shaded area represents the chance that the sum will be in the range from 5 to 10 inclusive. (The box had 85 tickets marked 0, 2 tickets marked 1, and 13 tickets marked 2.)

### Set B, page 318

1. (i) Exactly 6 heads. (ii) 3 to 7 heads exclusive.
   (iii) 3 to 7 heads inclusive.

2. The area between 51.5 and 52.5 under the histogram gives the exact chance. The normal curve is only an approximation (but a very good one).

3. The expected number of heads is 50; the SE is 5. You want the area of the rectangle over 60 in figure 3, p. 315.

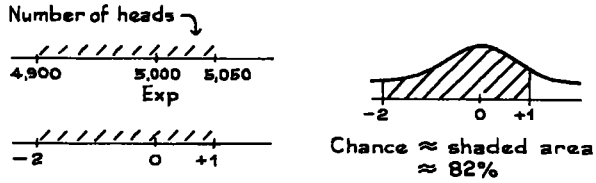



*Comment.* The exact chance is 1.084%.

4. From exercise 3, about one group in a hundred should have 60 heads. In fact, exactly one group in the hundred does (#6,901–7,000).

5. The expected number of heads is 5,000; the SE is 50.

(a)



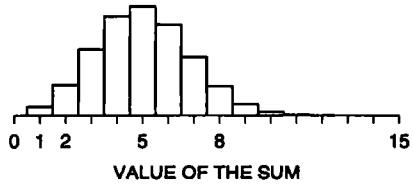

(b) chance ≈ 2% (c) chance ≈ 16%.

6. (a) Yes. The blocks are big. (b) No. Small blocks.

*Comment on (a).* Keeping track of the edges changes the estimate from 50% to 54%.

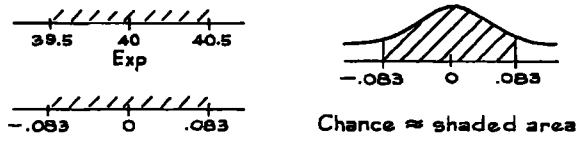
Set C, page 324

1. (a)




(b) 3 is more likely than 8: the block over 3 is bigger.

2. The number of heads in 400 tosses of this biased coin is like the sum of 400 draws from the box $\boxed{9\ \boxed{0}\text{'s}\ \boxed{1}}$. The expected number of heads is 40, and the SE is 6. You want the area of the rectangle over 40, at the bottom of figure 6 on p. 320.




From the table, this area is between 4% and 8%. (Actually, the area is 6.6%, and so is the chance.)

3. The normal curve is lower than the histogram around 1, so the estimate would be too low.

4. Yes. Big blocks.

5. A (ii), B (i), C (iii). The more lopsided the box, the more skewed the histogram.

*Comment.* With 25 draws from the box $\boxed{24\ \boxed{0}\text{'s}\ \boxed{1}}$, you cannot expect to get many $\boxed{1}$'s. The leftmost rectangle in the probability histogram gives the chance that the sum will be zero—the draws are all $\boxed{0}$. This chance is 36%. The next rectangle gives the chance that the sum will be one—one $\boxed{1}$ among the draws, and 24 $\boxed{0}$'s. This chance is 38%. And so forth. (The chances can be worked out using the binomial formula, chapter 15.)

6. (i) 100 (ii) 400 (iii) 900

   The histograms get closer to the normal curve as the number of draws goes up.

7. Choose (i).

   *Comment.* Chances are given by areas under probability histograms. Often, the corresponding area under the normal curve is a good approximation, but not here—the curve is much higher than the histogram, so the area under the curve is much bigger than the area under the histogram.

8. Most likely, 105; least likely, 101; expected value, 100.

   *Comment.* There is a trough in this histogram near the expected value. (With 100 draws the trough has disappeared.)

9. (a) Much smaller than 50%. The value 276,000 is 0.276 million, about half-way between the 0.2 and the 0.4 on the horizontal axis. The area to the right of this point is much smaller than 50%. (This histogram has a very long right-hand tail, and the expected value is a lot bigger than the median.)

   (b) $1{,}000{,}000/100 = 10{,}000$

   (c) 400,000 to 410,000 is a lot more likely, relatively speaking. The box just to the right of 400,000 is relatively much higher than the box just to the left. Products have quite irregular probability histograms.

# Part VI. Sampling

## Chapter 19. Sample Surveys

Set A, page 349

1. The population consists of all undergraduates registered in the current term. The parameter is the percentage of these undergraduates living at home.

2. (a) This is a probability method: it is perfectly definite, chance enters in a planned way—when you choose that random starting point between 1 and 100—and nobody has any discretion as to who gets in the sample.

   (b) The method is different from simple random sampling. For instance, two people whose names are adjacent on the list have no chance to get into the sample together. (Simple random samples are defined in section 4.)

   (c) The sample is unbiased: each person has an equal chance of getting into the sample.

3. Choose (ii). See pp. 334, 339, and 342.

4. The population and the sample are the same, namely, all men age 18 in the Netherlands in 1968; there is no room for sampling error.

5. Doing a survey by telephone could introduce bias, because telephone subscribers are probably different from non-subscribers. However, the percentage of non-subscribers is so small that this bias can usually be ignored. (If you are estimating small percentages, or are interested in the sort of people who might not have telephones, this bias can matter.) Using telephone books would introduce serious bias, since there are many unlisted numbers. See section 7.

   *Comment.* About 95% of households in the U.S. have telephones, according to *Statistical Abstract*, 2006, table 1117. The corresponding figure in 1980 was 93%.

6. No. You might expect the respondents interviewed by blacks to be much more critical. (And they were.)

7. No, this parish might have been quite different from the rest of the South. (It was: Plaquemines is sugar country, and sugar required more highly skilled labor than cotton.)

8. No. First, the ETS judgment about "representative" schools may have been biased. Next, the schools may not have used good methods to draw a sample of their own students.

   *Comment.* There are about 3,600 institutions of higher learning in the U.S., including junior colleges, community colleges, teachers' colleges. About 1,000 of them are very small, altogether enrolling only 10% of the student population. At the other end, there are about 100 schools with enrollments over 20,000—and these account for about one third of the student population.

9. Quite a bit different from. Non-respondents generally differ from respondents—early respondents probably differ from late ones. (In the study, the percentage with TB was quite a bit higher among the last 200 respondents: perhaps those people did not want to have their illness confirmed.)

10. A description of the sample design would be more reassuring than a sales pitch followed by a disclaimer.

11. With 200 replies out of 20,000 questionnaires, nonresponse bias is an overwhelming problem. With 200 responses out of 400 questionnaires, the response rate is adequate to show something important: a substantial fraction of high-school biology teachers hold creationist views.

12. False. The serious problem is non-response bias. Additional people brought into the sample to build it back up to planned size are likely to differ from non-respondents, and do not fix the problem of non-response bias.

## Chapter 20. Chance Errors in Sampling

Set A, page 361

1.

| | |
|---|---|
| population | box |
| population percentage | 40% |
| sample | draws |
| sample size | 1,000 |
| sample number | number of 1's among the draws |
| sample percentage | percentage of 1's among the draws |
| denominator for sample percentage | 1,000 |

2. The box model: make 400 draws from a box with 10,000 $\boxed{1}$'s and 15,000 $\boxed{0}$'s. The average of the box is 0.40, and the SD is about 0.5, so the expected value for the sum is $400 \times 0.4 = 160$ and the SE for the sum is $\sqrt{400} \times 0.5 \approx 10$.

   (a) EV for number = 160 and SE for number = 10.

   (b) EV for percent = $(160/400) \times 100\% = 40\%$, and

   $$\text{SE for percent} = (10/400) \times 100\% = 2.5\%.$$

   (c) 40%, 2.5%.

   *Comments.* (i) Parts (b) and (c) call for the same numbers, in part (c) you have to interpret the results. (ii) The expected value for the sample percentage is the population percentage (p. 359).

3. The SE for the number of heads is $\sqrt{10{,}000} \times 0.5 = 50$. The SE for the percent is $(50/10{,}000) \times 100\% = 0.5$ of 1%.

4. (a) and (b) are both true.

   *Comment.* When drawing at random from a 0–1 box, the EV for the percentage of 1's among the draws equals the percentage of 1's in the box. This is so whether the draws are made with or without replacement. The equality is exact.

5. False. They forgot to change the box. The number of 1's is like the sum of 400 draws from the box

$$\boxed{0\,0\,0\,1\,0}.$$

6. 10%+1%. The number of red marbles in the sample is $90 \pm 9$. If the number is 1 SE too high, it's $90 + 9$: now convert to percent out of 900. Our SE for a percentage is added to or subtracted from the expected value, not multiplied.

7. The total distance advanced equals the total number of spots thrown. This is like the sum of 200 draws (at random with replacement) from the box

$$\boxed{1\,2\,3\,4\,5\,6}.$$

   The average of this box is 3.5, and the SD is 1.7. So he can expect to advance around $200 \times 3.5 = 700$ squares, give or take $\sqrt{200} \times 1.7 \approx 24$ squares or so.

8. Sherlock Holmes is forgetting about chance error.

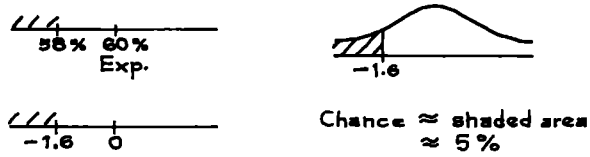
Set B, page 366

1. (a) The expected value for the percentage of reds in the sample equals the percentage of reds in the population. (Population = box, sample = draws.) See p. 359.
   (b) As the number of draws goes up, the SE for the number of reds in the sample goes up but the SE for the percentage of reds goes down. See p. 360.

2. The first thing to do is to set up a box model. There should be 30,000 tickets in the box, one for each registered voter; 12,000 are marked 1 (Democrat) and 18,000 are marked 0. The number of Democrats in the sample is like the sum of 1,000 draws from the box. The fraction of 1's in the box is 0.4. The expected value for the sum is $1{,}000 \times 0.4 = 400$. The SD of the box is $\sqrt{0.4 \times 0.6} \approx 0.49$. The SE for the sum is $\sqrt{1{,}000} \times 0.49 \approx 15$.

   (a) The expected value for the percent is 400 out of 1,000, or 40%. The SE for the percent is 15 out of 1,000, or 1.5%. (No surprise about the expected value: 40% of the registered voters are Democrats.)
   (b) The percentage of Democrats in the sample will be around 40.0%, give or take 1.5% or so. Parts (a) and (b) require the same calculations; in (b), you have to to interpret the results.
   (c) This is $\pm 0.67$ SE, the chance is about 48%.

3. (a) There should be 100,000 tickets in the box, one for each person in the population, of which 60,000 are marked 1 (married) and 40,000 are marked 0. The number of married people in the sample is like the sum of 1,600 draws from the box. The expected value for the sum is $1{,}600 \times 0.6 = 960$. The SD of the box is $\sqrt{0.6 \times 0.4} \approx 0.5$. The SE for the sum is $\sqrt{1{,}600} \times 0.5 = 20$. The number of married people in the sample will be 960, give or take 20 or so. Now 960 out of 1,600 is 60%, and 20 out of 1,600 is 1.25%. So 60% of the people in the sample will be married, give or take 1.25% or so.




(b) There should be 100,000 tickets in the box, of which 10,000 are marked 1 (income over $75,000) and the other 90,000 are marked 0. There are 1,600 draws. The chance is about 9%.

(c) The box has 100,000 tickets, of which 20,000 are marked 1 (college degree) and the other 80,000 are marked 0. There are 1,600 draws. The chance is about 68%.

4. The shaded area represents the chance of drawing a sample in which 22% or more of the sample persons earn more than $50,000 a year.

5. (a) the chance that the sample will have 88 high earners
   (b) the chance that the sample will have 22% high earners
   (c) 88 is 22% of 400, so the same chance is described in two different ways. No coincidence at all.

Set C, page 370

1. Option (iii) is right. That is the point of the section.

2.

| *Number of draws* | *SE for percentage of 1's among draws* |
|---|---|
| 2,500 | 1% |
| 25,000 | 0.27 of 1% |
| 100,000 | 0% |

*Comment.* After 100,000 draws, there are no more tickets in the box, and no uncertainty about the percentage of 1's among the draws.

3. The sample size should be 2,500.

4. The SE is the same for all three boxes, because all three have the same fraction of 1's, so the same SD.

5. SE with = 20%; SE without $= \sqrt{\dfrac{10-4}{10-1}} \times 20\% \approx 16\%$.

*Comment.* This is an artificial example where the number of draws is a large fraction of the number of tickets in the box, so the correction factor really kicks in.

## Chapter 21. The Accuracy of Percentages

Set A, page 379

1. (a) observed (b,c) estimated from the data as

*Comment.* There is a big difference between chapter 20 and chapter 21. In chapter 20, you knew the composition of the box, and could compute the expected value and SE exactly. Here, the composition of the box has to be estimated from the data. In chapter 20, you reason forward, from the box to the draws. Here, you reason backward, from the draws to the box.

2. The first step is to set up the model. (We need the box model to compute the SE for the sum of draws.) There are 100,000 tickets in the box, some marked 1 (currently enrolled in college) and the others 0 (not enrolled). Then 500 draws are made from the box to get the sample. The number of college students in the sample is like the sum of the draws. The fraction of 1's in the box is unknown, but can be estimated by the fraction of 1's observed in the sample, which is $194/500 \approx 0.388$. So the SD of the box is estimated as $\sqrt{0.388 \times 0.612} \approx 0.49$. The SE for the sum is $\sqrt{500} \times 0.49 \approx 11$. The 11 is the likely size of the chance error in the 194. The SE for the percentage of 1's is $(11/500) \times 100\% = 2.2\%$. The percentage of persons 18–24 in the town who are college students is estimated as 38.8%. The estimate is likely to be off by 2.2% or so. The estimate is 38.8%, and the give-or-take number is 2.2%.

3. The estimate is 48%, give or take 5% or so.

4. The estimate is 2.8%, give or take 0.8 of 1% or so.

5. The estimate is 46.8%, give or take 2.5% or so.

6. No. Most people work for the few large establishments.

7. SE = 2%.

8. (a) $18.0\% \pm 1.9\%$ (b) $21.0\% \pm 2.0\%$ (c) $24.5\% \pm 2.2\%$

   *Comment.* The third person is off by a couple of SEs in estimating the percentage of 1's in the box; even so, the estimated standard error is only off by 0.2 of 1%. The bootstrap method is good at estimating SEs.

9.

| | *Known to be* | *Estimated from the data as* |
|---|---|---|
| Observed value | 30.8% | N/A |
| Expected value | N/A | 30.8% |
| SE | N/A | 1.5% |
| SD of box | N/A | 0.46 |
| Number of draws | 1,000 | N/A |

Set B, page 383

1. (a) observed (b,c) estimated from the data as
   See exercise 1 on p. 379.

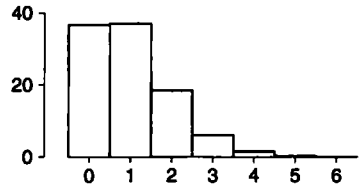
2. (a) $38.8\% \pm 4.4\%$ (b) $38.8\% \pm 6.6\%$ (c) $38.8\% \pm 3.3\%$

   *Comments.* As the confidence level goes up, the confidence interval gets longer. However, as the sample size goes up, the confidence interval gets shorter.

3. (a) Expect 1 red marble among the draws, give or take 1 or so.
   (b) It is impossible to draw fewer than 0 red marbles, so the chance is 0.
   (c) About 16%.
   (d) No. If the probability histogram looks like the normal curve, then the chance of drawing fewer than 0 red marbles can be read off the curve. Since $16\% \neq 0\%$—see (b) and (c)—the histogram does not look like the curve.

   *Comment.* The histogram is shown at the top of the next page.

4. False. The normal approximation cannot be used here. As best we can estimate from the sample, 1% of the marbles in the box are red, and 99% are blue. This is




the box in exercise 3. The probability histogram for the percentage of reds among 100 marbles drawn from this box does not look like the normal curve. (With 100 draws out of 10,000, there is little difference between sampling with or without replacement.) If the sample were bigger, or the box were less lopsided, the normal curve would be fine.

Set C, page 386

1. Probabilities are used when reasoning from the box to the draws; confidence levels are used when reasoning from the draws to the box.

2. (a) The chance error is in the observed value.
   (b) The confidence interval is for the population percentage.

3. (a) 18.0% ± 3.8%, covers.
   (b) 21.0% ± 4.0%, covers.
   (c) 24.5% ± 4.4%, just misses.

4. (a) True.
   (b) False. The EV is computed exactly; the chance error is in the sample percentage of reds, not in the expected value.
   (c) True.
   (d) False. Confidence intervals are for parameters, not sample data. See pp. 385–386.
   (e) True.

   *Comment on (b).* The SE tells you the likely size of the chance error in the percentage of reds among the draws. The 50%, however, is a property of the box and does not depend on how the draws turn out: there is no chance error in the 50%. For instance, if you draw 100 times and get 53 reds, the sample percentage of reds is 53%, and the chance error—in the 53%—is +3%. If you get 42 reds, the percentage of reds among the draws is 42%, and the chance error in the 42% is −8%. But the expected value stays the same, no matter how the draws turn out. Also see exercise 6 on p. 294.

5. (a) True. (b) True. (c) True. (d) True.
   (e) False; the sample percentage is 53%, you don't need a confidence interval for that.

6. (a) True.
   (b) True.
   (c) False. The sample percentage is known, and in the interval.
   (d) False. If you view the interval as fixed, the chance is either 0 or 1. Moral: the chances are in the sampling procedure, not the population. That is why statisticians use the term "confidence interval."

7. False. The SE for the percentage measures the likely size of the difference between

one sample percentage and the population percentage; not the difference between two sample percentages.

*Comment.* The SE for the difference between two sample percentages has to be bigger, because both are subject to chance variability; by contrast, the population percentage isn't varying. See chapter 27 for more about the difference between two sample percentages.

8. True. Probabilities are used when you reason forward, from the box to the draws; confidence levels are used when reasoning backward, from the draws to the box: see pp. 385–386.

Set D, page 388

1. Theory says, watch out for this man. What population is he talking about? Why are his students like a simple random sample from the population? Until he can answer these questions, don't pay much attention to the SEs he calculates.
2. This is not a simple random sample: you are guaranteed to get 25 students from each class, a simple random sample won't do that. The procedure does not apply.

Set E, page 390

1. This isn't a simple random sample, the formulas don't apply.
2. This is fine.
3. (a) "altered voter enthusiasm"
   (b) Chance variation—the Gallup Poll is based on a random sample.
   (c) As table 2 shows, chance errors of several percentage points are quite possible. Maybe late September is not such a good guide to early November after all. (On the other hand, Bush did win.)

## Chapter 22. Measuring Employment and Unemployment

Set A, page 403

1. (a) True.
   (b) False. The Bureau would divide up the sample into groups, by race, age, and so on, then weight up each group separately; section 4.
2. 151.4 million $\pm$ 0.1 million; section 5.
3. This is a simple random sample of households, and the inference is about households. The SD of the box is estimated as $\sqrt{0.80 \times 0.20} = 0.40$. The SE for the sum is $\sqrt{100} \times 0.40 = 4$. The SE for the percentage is 4%.
4. This is a simple random sample of households, but a cluster sample of people. (The household is the cluster.) The inference is about people. So, you need more information to estimate the SE—the formulas for simple random samples do not apply (section 5).

   *Comment on exercises 3 and 4.* In exercise 3, you have a simple random sample of households, and make an inference about households—the percent where all occupants are vaccinated. In exercise 4, you are making an inference about people from a cluster sample of people.
5. The SE for the percentage is only 0.2 of 1%, so a discrepancy of $55\% - 52\% = 3\%$

is almost impossible to explain as a chance error. People like to say they voted, even if they didn't.

6. The one for white males; it is based on a lot more people.

## Chapter 23. The Accuracy of Averages

Set A, page 413

1. (a) $7{,}611/100 = 76.11$ (b) $73.94 \times 100 = 7{,}394$
2. The SE for the average is 1. The answer to (a) is almost 100%. The answer to (b) is 68%. Don't confuse the SE for the average of the draws with the SD of the box.
3. (a) False. (b) True.
   To repeat, do not mix up the SE for the average of the draws with the SD of the box.
4. (a) The expected value for the average of the <u>draws</u> equals the average of the <u>box</u>.
   (b) As the number of draws goes up, the SE for the <u>sum</u> of the draws goes up but the SE for the <u>average</u> of the draws goes down.
5. The SE for the sum of the draws is $\sqrt{100} \times 20 = 200$. The SE for the average is $200/100 = 2$. The average of the draws will be around 50, give or take 2 or so. This is still true if the draws are made without replacement, because only a small fraction of the tickets in the box are drawn out. On the other hand, if you draw 100 tickets at random without replacement from a box of 100 tickets, the SE is 0.
6. The chance that the average of the draws is between 2.25 and 2.75.
7. The percentage of times $\boxed{4}$ came up in the 50 draws.
8. (a) The chance that the sum will be 90.
   (b) The chance that the average will be 3.6.
   (c) $3.6 = 90/25$, so the same chance is described in two different ways. No coincidence at all. See exercise 5 on p. 366.
9. (a), (c), (e) are true; (b), (d), (f) are false. You know the contents of the box; you can compute the expected value for the average without error; however, there is chance error in the average of the draws. See exercise 6 on p. 294, exercises 4–6 on pp. 386–387.
10. The average of the draws is just their sum, divided by 25 (the number of draws). So 25 changes to 1, 50 to 2, and 55 to $55/25 = 2.2$.

Set B, page 420

1.

| | |
|---|---|
| population | box |
| population average | average of the box |
| sample | draws |
| sample average | average of the draws |
| sample size | number of draws |

2. (a) "SD of box" makes sense; "SE for box" does not.
   (b) "SE for average of draws" makes sense; "SE for average of box" does not.
   The term "SD" applies to a list of numbers; "SE" applies to a chance process. The tickets in the box (and their average) are fixed, but the draws are random.

3. (a,b) Estimated from the sample as. The SD of the sample is \$19,000; this is used to estimate the SD of the box. The SE is based on the estimated SD; so it too is an estimate. If you do not know what is in the box, you have to estimate the SD and the SE from the data.
   (c) observed.

4. 95% of 50 $\approx$ 48.

5. (a) Each organization takes its sample average as the center of its confidence interval. The sample averages are different, because of chance variation.
   (b) The sample SDs are different (chance variation), so the estimated SEs are different. That is why the lengths of the intervals are different.
   (c) 49.

6. The box has 30,000 tickets, one for each registered student, showing his or her age. The data are like 900 draws from the box; the sample average is like the average of the draws. The SD of the box is estimated as 4.5 years, the SE for the sum of the draws is $\sqrt{900} \times 4.5 = 135$ years, the SE for the average is $135/900 = 0.15$ years.
   (a) Estimate is 22.3 years, off by 0.15 years or so.
   (b) The interval is $22.3 \pm 0.3$ years.

7. (a) The interval is \$568±\$24. Even though the data don't follow the normal curve, the probability histogram for the average of the draws does.
   (b) False: \$24 is the SE for the average of the draws, not the SD of the box.

8. False. The SE for the average gives the likely size of the difference between the sample average and the population average, not the difference between two sample averages. So \$18 is the wrong margin of error. See exercise 7 on p. 387.

9. The probability histogram is about chances for the sample average; it is not about data. Here, the probability histogram is given. Part (a) asks for +1 in standard units, relative to the probability histogram. We need the center and spread of this histogram. The center is the expected value for the sample average, which equals the average of the box. This is given: it is \$61,700. The spread is the SE for the sample average. This can be worked out exactly, because the problem gives the SD of the box. This is \$50,000. So the SE for the sum of the draws is $\sqrt{625} \times \$50{,}000 = \$1{,}250{,}000$. The SE for the average of the draws is $\$1{,}250{,}000/625 = \$2{,}000$. And +1 in standard units is $\$61{,}700 + \$2{,}000 = \$63{,}700$. That is the answer to (a).

   In part (b), you are being asked to see where \$58,700 fits, on the axis of the probability histogram. It comes in below the expected value: \$58,700 is below \$61,700. So, \$58,700 is on the negative part of the axis. In fact, this value is \$3,000 below the expected value. And 1 SE is \$2,000. So \$58,700 is $-1.5$ in standard units. That is the answer to (b).

   *Comments.* (i) The key point: in this problem, the average and SD of the box are given.

   (ii) A typical sample average is around 1 SE away from the population average. Our sample average was 1.5 SE too low. We didn't get enough rich people in the sample.

   (iii) Look at figure 1 on p. 411. The histogram is about the process of drawing at random and taking the average; it is not about any particular set of draws. If you draw 25 tickets and their average happens to be 3.2, that doesn't change the histogram. This exercise illustrates the same point, in a more complicated setting.

   (iv) You would use the SD of \$50,000 to convert to standard units relative to a data histogram—for the incomes of all 25,000 families in the town. The SD of

$49,000 works relative to another data histogram—for the incomes of the 625 sample families.

Set C, page 423

1.

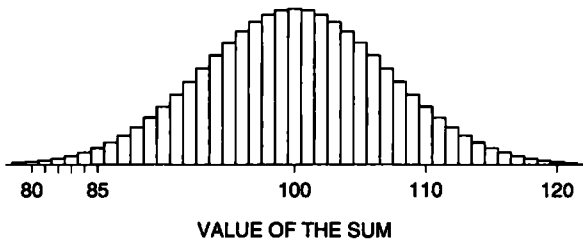
| *Number of draws* | *EV for sum of draws* | *SE for sum of draws* | *EV for average of draws* | *SE for average of draws* |
|---|---|---|---|---|
| 25 | 75 | 10 | 3.0 | 0.4 |
| 100 | 300 | 20 | 3.0 | 0.2 |
| 400 | 1,200 | 40 | 3.0 | 0.1 |

2. (a) True. The expected value for the average of the draws equals the average of the box (p. 410).
   (b) Can't tell; you need the SD of the box.

3. (a) Estimated from the data as; you would need the average of the box to compute the expected value exactly.
   (b) To compute the SE exactly, you need the SD of the box; even to estimate it, you would need the SD of the draws.

   *Comment.* The expected value applies to the process of drawing at random, rather than any particular set of draws. For example, suppose you draw 25 times at random with replacement from the box [0 2 3 4 6]. The expected value for the average of the draws is 3. The average of your draws could be 3.1, which is 0.1 above the expected value; or, the average of the draws could be 2.6, which is 0.4 below. There are many other possibilities. But the expected value only depends on the box, and stays the same no matter how the draws turn out.

4. (a) The SE for the <u>sum of the draws</u> is 7.1, and the SE for the <u>average of the draws</u> is 0.18.
   (b) The expected value of 100 is at the center; the next tick mark to the right is 10 boxes over, that must be 110, and so forth.




5. You can't estimate the SD of the box, so you can't get margins of error.

6. For all three boxes, the EV for the sum of 100 draws is 200. The SE for the average of the draws is

   1 from box A  1.4 from box B  2 from box C.

   (a) 203.6 is very unlikely to come from box A—it is 3.6 SEs away from the expected value for the average of 100 draws from box A. It is also quite unlikely to come from box B, because $3.6/1.4 \approx 2.6$ is too many SEs. So it comes from box C. Similarly, 198.1 comes from box B, leaving 200.4 for box A by elimination.
   (b) It could be otherwise, but that would be pushing things.

Set D, page 424

1. The 95%-confidence interval is $1.86 \pm 0.06$.
2. This is qualitative data, use the method of chapter 21. The interval is $60.1\% \pm 5.4\%$.
3. Can't be done with the normal curve. Suppose the sample reflects the population exactly. Then the company is drawing from a box which has 99.87% $\boxed{1}$'s and 0.13 of 1% $\boxed{0}$'s. This box is so skewed that with 750 draws, the probability histogram for the sum won't be anything like the normal curve. See exercises 3 and 4 on p. 383.
4. This is not a simple random sample of people: either you get everybody in a household, or nobody. So the SE can't be estimated by the methods of this chapter. See exercises 3 and 4 on p. 404.

   *Comments.* (i) This is a cluster sample of people—the household is the cluster; the half-sample method could be used to get the SE (p. 402), but more information would be needed.

   (ii) People in a household tend to be similar with respect to TV-watching, so this sample will be less informative than a simple random sample of the same size. Cluster samples are less accurate than simple random samples, but much cheaper to take.

   (iii) The usual problem with cluster samples is chance error, rather than bias; the sampling method in this exercise is unbiased.
5. (a) This is not a probability sample of any kind. It is a sample of convenience.
   (b) Same as (a).

   *Comment.* A cluster sample is a special kind of probability sample (pp. 340, 342).
6. The average of the box is estimated by the average of the sample: $297/100 \approx 3.0$; for the SE, you need the SD.
7. The two procedures are the same: simple random sampling means drawing at random without replacement (p. 340).

# Part VII. Chance Models

## Chapter 24. A Model for Measurement Error

Set A, page 444

1. Use the SE for the sum; this was figured as 60 micrograms.
2. The estimate is the average of the measurements, 82,670 pounds. This is likely to be off by the SE for the average, 100 pounds.
3. (a) 800 microns (b) 80 microns (c) 91.4402 cm $\pm$ 160 microns
4. (a) False. This range is 2 SEs, not 2 SDs, either way from the average.
   (b) False. Same reason as in (a).
   (c) True. See p. 384.
   (d) False. This is just like exercise 8, p. 421.
5. The factor is 5.

Set B, page 449

1. You would have to toss the thumbtack many times, and see whether the percentage of times it landed point down was closer to 50% or to 67%. (This will depend on the surface: in one experiment, the tack landed point down 66% of the time when tossed on linoleum, but only 50% of the time when tossed on a carpet.)
2. No. The rainy days all come close together in the rainy season. If it rains one day, it is more likely to rain the next.
3. Last digits, yes. First digits, no. For instance, in the San Francisco phone book the first digit cannot be 0. Also, many more phone numbers start with 9 than with 2.
4. No, the letters come out in alphabetical order. No box will do that.
5. Like a shot. You have a 50–50 chance to win \$5 or lose \$4.

Set C, page 452

1. In both cases, the measurement is 504 micrograms above ten grams.
2. No, as the previous exercise shows.
3. Six micrograms is the SD of the 100 measurements reported in table 1 on p. 99. This is used to estimate the SD of the error box. So, "estimated from the data as."
4. (a) Chance variation—the investigators get different sample averages.
   (b) Chance variation again—the investigators get different sample SDs.
   (c) About 95% of the 50 intervals should cover the exact weight, that is, about 48 intervals.
   (d) 48. (One of the intervals is off by quite a bit—chance variation at work.)
5. The SD of the error box is estimated as 50 micrograms.
   (a) 5 micrograms—the SE for the average.
   (b) 50 micrograms—the estimated SD of the error box.
   (c) 95%—two SEs.
6. The answer is 1.2 micrograms. See example 5, p. 451.
7. (a) 300,007 (the average); 2 (the SE for the average).
   (b) False: the average is 300,007 exactly.
   (c) True: each number on a list is off the average of the list by an SD or so.
   (d) True: the interval is "average $\pm$ 2 SEs."
   (e) False: the average of the 25 measurements is 300,007 exactly.
   (f) False: 2 is the SE, not the SD.
8. The answer is 2 inches. Here is the reason. Each measurement equals the exact length, plus a draw from the error box. The estimated distance AE is the sum of the 4 measurements, and is off the exact length AE by the sum of 4 draws from the box. The average of the error box is 0. So the sum of 4 draws will be around 0, give or take an SE or so. It is the SE for the sum which is the right give-or-take number. The SD of the box is 1 inch, so the SE for the sum is $\sqrt{4} \times 1$ inch $= 2$ inches.

   *Comment.* Finding the length AE involved adding the measurements, not averaging them.
9. The chance errors for different people could have different SDs. Also, if the same

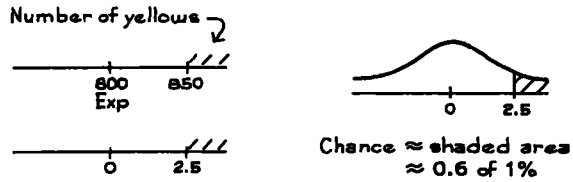
person takes the test several times, the errors may be dependent. The Gauss model does not seem to apply.

## Chapter 25. Chance Models in Genetics

Set A, page 461

1. Each seed has a 50% chance to get $y$ from the $y/g$ parent, and a 50% chance to get $g$. It is bound to get $g$ from the $g/g$ parent. So the seed has a 50% chance to be $y/g$, and yellow in color; it has a 50% chance to be $g/g$, and green in color. About 50% of the seeds should be yellow.

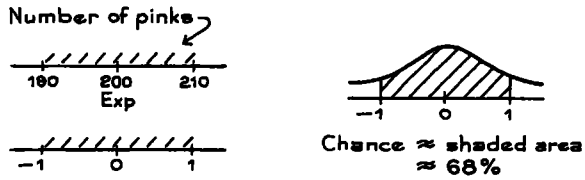
   The number of yellows among 1,600 seeds is like the sum of 1,600 draws from the box $\boxed{0}\boxed{1}$. The expected number of yellows is $1{,}600 \times 1/2 = 800$. The SE for the number is $\sqrt{1{,}600} \times 1/2 = 20$. Now the normal approximation can be used:




2. (a) white × red → 100% pink
   white × pink → 50% white, 50% pink
   pink × pink → 25% red, 50% pink, 25% white.
   *Work for pink × pink:* Each parent is $r/w$, so the offspring's flower color is determined by choosing a row and column at random from the table below.

| | $r$ | $w$ |
|---|---|---|
| $r$ | red | pink |
| $w$ | pink | white |

   (b) The expected number of pinks in 400 plants is 200, with an SE of 10. Use the normal approximation:




3. (a) One gene-pair controls leaf width, with variants $w$ (wide) and $n$ (narrow). The rules: $w/w$ makes wide, $w/n$ and $n/w$ make medium, and $n/n$ makes narrow.
   (b) narrow × narrow $= n/n \times n/n \rightarrow 100\%\, n/n =$ narrow
   narrow × medium $= n/n \times n/w \rightarrow$

$$50\%\, n/n = \text{narrow}, \quad 50\%\, n/w = \text{medium}.$$

4. $B$ = brown, $b$ = blue. Husband is $B/b$, wife is $b/b$. Each child has 1 chance in 2 of having brown eyes. The three children are independent, so the chance that all three will be brown-eyed is $(1/2)^3 = 1/8$.

# Part VIII. Tests of Significance

## Chapter 26. Tests of Significance

Set A, page 476

1. (a) estimated from the data as
   (b) observed
2. The observed value is only 1.3 SEs below expected, Dr. Null is looking good.
3. Again, the observed value is about 1.3 SEs below expected, and Dr. Null is looking good. Moral: results depend on the observed value, the expected value, and the SE. The SE depends on the SD and the sample size.
4. If the die is fair, the total number of spots is like the sum of 100 draws from the box $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}\boxed{6}$. The average of the box is 3.5; the SD is 1.7. So the expected value for the sum is 350, and the SE is 17. The number of spots is a little over 1 SE above its expected value, which looks like chance variation.
5. The problem can be set up like exercise 4, but this time the number of spots is over 3 SEs above its expected value. This doesn't look like chance variation.
   *Comments.* (i) Sample size matters; compare exercises 4 and 5.
   (ii) A more complete test for the fairness of a die will be presented in chapter 28.

Set B, page 478

1. (iii)
2. The <u>null</u> hypothesis says that the difference is due to chance but the <u>alternative</u> hypothesis says that the difference is real.
3. Choose (ii). Dr. Null and Dr. Alt both knew the data, they didn't know what was in the box. The null hypothesis is a statement about the box, and the test tells you whether this statement is plausible.
4. <u>box</u>. The null hypothesis is about the box.
5. The SD of the box can be estimated as 10, so the SE for the average of 100 draws is estimated as 1. If the average of the box is 20, then the average of the draws is 2.7 SEs above its expected value. This isn't plausible.

Set C, page 481

1. (a) $P = 32\%$ is best for the null.
   (b) $P = 0.1$ of 1% is best for the alternative.
   Big $P$ is good for null; small $P$ is bad for null.
2. (a) True. (b) False. See pp. 480–81.
3. (a) True, see pp. 480–81.
   (b) False, see pp. 480–81.
4. SE for average $\approx 1.25$, so $z \approx (52.7-50)/1.25 \approx 2.16$ and $P$ is approximately the area to the right of 2.16 under the normal curve. From the table, this is about 1.6%. The difference is hard to explain as chance variation. The alternative hypothesis is looking good.

5. box. The null hypothesis is about the box.
6. No. With 10 draws, the probability histogram for the sample average may not look like the normal curve, and the SD of the data will not be a good estimate for the SD of the box.
7. The sample is like 100 draws made at random from a box which has one ticket for each employee, showing the number of days that employee was absent. Null hypothesis: the average of the box is 6.3 days. Alternative hypothesis: the average of the box is less than 6.3 days. The SD of the box is estimated as 2.9 days, so the SE for the average is 0.29 days, and $z \approx (5.5 - 6.3)/0.29 \approx -2.8$, so $P \approx 0.3$ of 1%. This is strong evidence against the null; chance variation will not explain the drop in absenteeism.
8. Now $z \approx (5.9 - 6.3)/0.29 \approx -1.4$, so $P \approx 8\%$. The null hypothesis looks more plausible.

Set D, page 482

1. (a) False. Even if the null hypothesis is true, 1% of the time the experiment will give a result which is "highly significant."
   (b) False; pp. 480–81. (c) False; pp. 480–81.
2. (a) True. Big $P$ is good for null.
   (b) True. Small $P$ is bad for null.
3. (a) True; p. 482.
   (b) False; $P$ has to be less than 1%.
   (c) True; p. 482.
   (d) True; p. 479.
   (e) True; $z = (\text{obs} - \text{exp})/\text{SE}$, and "exp" is computed on the null.
4. About 2%.
5. (a) True; $z = (\text{obs} - \text{exp})/\text{SE}$; "obs" is the average of the draws; "exp" is the average of the box, given as 50.
   (b) About 50.
   (c) About 2; and 3 of them do.
   (d) About 2%. See exercise 4.

Set E, page 486

1. Option (i) is right: "like" means, "as far as the chances are concerned." Each guess has one chance in four to be right, and each draw has one chance in four to be 1. Then the number of correct guesses is like the sum of the draws, and the square root law applies.

   Option (ii) is wrong: if there is no ESP, the chance of a correct guess is 1/4 not 1/3. Option (iii) is worse: the 2,006/7,500 is the fraction of 1's in the sample, not in the box. Option (iv) is wrong: the fraction of 1's in the sample is known, there is no argument about that. Option (v) is way off: the null corresponds to the idea that there is no ESP.
2. (a) student registered at Berkeley that term. Reason: the box corresponds to the population.
   (b) 1 = man, 0 = woman. Reason: you're counting the men. (If you want to count the women, that's fine too, but be consistent.)
   (c) There are 25,000 tickets in the box, and 100 draws. Reason: the sample is like the draws.

(d) 100 draws.
(e) 67%. Reason: You know the percentage of men in the population.

3. (a) 53 (b) 67 (c) sum
(d) $\sqrt{100} \times \sqrt{0.67 \times 0.33} \approx 4.7$
(e) $z \approx (53 - 67)/4.7 \approx -3$ and $P \approx 1/1{,}000$.

4. No. He got too many women. $P$ is very small, so chance won't explain the difference. Taking people haphazardly isn't like a simple random sample (chapter 19).

5. (a) Computed from the null hypothesis: $100 \times 0.67$. The expected is always computed from the null hypothesis.
(b) Computed from the null hypothesis. Here, the null tells you the composition of the box. Otherwise, you might have to estimate the SD of the box from the data (p. 485).

6. (a) Null hypothesis: the number of correct guesses is like the sum of 1,000 draws from a box with one ticket marked 1 and nine 0's.
(b) $\sqrt{0.1 \times 0.9}$. The null hypothesis tells you what's in the box. Use it.
(c) $z \approx (173 - 100)/9.5 \approx 7.7$, and $P$ is tiny.
(d) Whatever it was, it wasn't chance variation.

7. (a) Tossing the coin is like drawing at random with replacement 10,000 times from a 0–1 box, with 0 = tails and 1 = heads. The fraction of 1's in the box is unknown. Null hypothesis: this fraction equals 1/2. Alternative: the fraction is bigger than 1/2. The number of heads is like the sum of the draws.
(b) $z = 3.34$, $P \approx 4/10{,}000$.
(c) There are too many heads to explain as chance variation.

8. (a) Same as 7(a). (b) $z = 1.34$, $P \approx 9\%$.
(c) The coin looks to be fair.

9. (a) box. The null hypothesis is about the box.
(b) False; see pp. 480–81.

10. The data consist of the 25 weights. The null hypothesis says that the data are like 25 draws made at random from a box. There is one ticket in the box for each animal in the colony, showing its weight. So the average of the box is 30 grams. And its SD is 5 grams, so the SE for the average of 25 draws is 1 gram. Now $z = (33 - 30)/1 = 3$, and $P \approx 1/1{,}000$.

*Comments.* (i) Here, the null tells you the SD of the box, so you don't need to estimate it from the data. The SD of the data is not used in working the problem. See p. 485.

(ii) Choosing haphazardly is not like taking a simple random sample (chapter 19). When you reach into the cage to pick up an animal, probably it is the tamer ones who come to your hand, and they are a bit heavier than the others.

11. The null hypothesis says that the reduced price had no effect on sales volume. So in each pair of stores, the one with the regular price is just as likely to sell more as its partner with the reduced price. In terms of a box model, the null hypothesis says the data are like 25 draws from the box $\boxed{1}\boxed{0}$, where 1 means the regular-price store sold more and 0 means the regular-price store sold less. The expected number of 1's is 12.5, and the SE is 2.5, so $z = (18 - 12.5)/2.5 = 2.2$ and $P \approx 1.4\%$. The evidence against the null is quite strong.

*Comment.* This procedure is called "the sign test." See exercise 6 on p. 258 (kangaroos) and exercise 11 on p. 262 (smokers). If the continuity correction is made, the normal approximation gives $P \approx 2.28\%$, compared to 2.16% from the binomial formula.

Set F, page 493

1. (a) 5% (b) 5% (c) 90% (d) 95%
2. From the table, the area to the right of 2.92 is 5%, and the area to the right of 6.96 is 1%. Since 4.02 is between 2.92 and 6.96, the area to the right of 4.02 is between 1% and 5%.
3. No, 3 degrees of freedom.
4. (a) degrees of freedom = 2, ave $\approx 72.7$, $\text{SD}^+ \approx 5.7$, SE $\approx 3.3$,
$$t \approx (72.7 - 70)/3.3 \approx 0.8,$$
$P$ is about 25%. Inference: the calibration is fine.
   (b) $P$ is about 2.5%, recalibrate.
   (c) One measurement is never enough.
   (d) $P$ is about 25%.

   *Comment.* Two measurements are better than one; more would be even better.
5. In (a), the number 93 is an outlier, so the errors do not seem to follow the normal curve. In (c), the numbers are just switching back and forth between 69 and 71. This is not good for the Gauss model.
6. According to the Gauss model, each of the 10 new measurements equals the exact weight, plus bias, plus a draw from the error box. The null hypothesis says that the bias is zero; the alternative hypothesis says that there is some bias. The SD of the error box is estimated as $\sqrt{10/9} \times 9 \approx 9.5$. (The errors belong to the rebuilt scale, so the old SD of 7 micrograms is irrelevant.) The SE for the average $\approx 3$ micrograms, $t \approx -2.67$. The area to the left of $-2.67$ under Student's curve with 9 degrees of freedom is about 1%, strong evidence against the null.
7. (a) According to the Gauss model, each of the 100 measurements equals the exact weight plus a draw from the error box. The tickets in the error box average out to 0. The unknown parameter is the exact weight. The null hypothesis says that this is still 512 micrograms above a kilogram. The alternative says that the exact weight is less.
   (b) The SD of the error box can be estimated from the past data as 50 micrograms: the error box belongs to the equipment. (The new SD of 52 micrograms is irrelevant.)
   (c) With 100 measurements, use $z$ not $t$. The SE for the average of 100 measurements is 5 micrograms, so $z = (508 - 512)/5 = -0.8$ and $P \approx 21\%$.
   (d) The drop in weight looks like a chance variation.

## Chapter 27. More Tests for Averages

Set A, page 503

1. True, now the numbers are independent, so the square root law applies.
2. The expected value is $100 - 50 = 50$, and the SE is $\sqrt{2^2 + 3^2} \approx 3.6$. The square root law applies because the draws are all independent.
3. The expected value for each percent is 50%; the SEs are 2.5 and 5 percentage points. The expected value for the difference is 0, and the SE is $\sqrt{2.5^2 + 5^2} \approx 5.6$ percentage points. The square root law applies because the two percents are independent.

4. (a,b) True.
   (c) False. The percentages are dependent: if the coin lands heads, it can't land tails. The square root law does not apply.

   *Comment.* The difference "number of heads − number of tails" is like the sum of 500 draws from the box $\boxed{-1}\boxed{+1}$, so the SE for the difference in the two numbers is about 22, and the SE for the difference in percentages is

$$(22/500) \times 100\% = 4.4\%.$$

5. True. If the draws are made with replacement, the two averages would be independent: the SE for the difference would equal $\sqrt{3^2 + 3^2}$ exactly. The box is so large that there is no practical difference between drawing with or without replacement.

6. The SD of box F can be estimated as 3, so the SE for the average of 100 draws from box F is 0.3; similarly, the SE for the average of 400 draws from box G is estimated as 0.4; the averages are independent, so the SE for the difference is $\sqrt{0.3^2 + 0.4^2} = 0.5$. If the two boxes have the same average, the observed difference $51 - 48 = 3$ is 6 SEs away from the expected value of 0. Not a likely story.

Set B, page 506

1. Two-sample $z$-test.

2. (a) Two-sample $z$-test: you're comparing two samples.
   (b) The setup is as in the text, with a 1990 box and a 2004 box. There are oodles of tickets in each box, and 1,000 draws from each. The tickets show test scores. The null hypothesis says the average of the boxes are the same. The alternative says the averages are different.
   (c) The SE for the difference is 1.37, so $z = 2/1.37 \approx 1.46$. This could easily be chance variation.

3. This difference is big, and highly significant—both practically and statistically.

4. The difference between the two sample averages is 2.1 hours, and the SE for the difference is 0.35 hours. So $z \approx 6$, and $P \approx 0$. The difference is very hard to explain away as a chance variation. Students in private universities generally come from wealthier families, and have more support from home.

5. The numerator is in percent, and the denominator is a decimal. That's a mistake. In fact, $z = (41 - 17)/4.8 = 5$. If you prefer decimals, $z = (.41 - .17)/.048 = 5$.

   *Comment.* Forgetting to convert the denominator to percent is a common slip. You can do the whole problem in percents or in decimals, but don't change in the middle.

6. There are two samples, so you need to make a two-sample $z$-test. The data consist of 1,600 0's and 1's for the men (1 = illiterate), and another 1,600 0's and 1's for the women. The model has two boxes, M and F. Box M has a ticket for every male youth in the country, marked 1 for the illiterates and 0 for the literates. Box F is similar, for the females. The data for the men are like 1,600 draws from box M, and similarly for the women. Null hypothesis: the percentage of 1's is the same in the two boxes. Alternative: the percentage of 1's is bigger in box M. The SE for the percentage of 1's in the male sample can be estimated as 0.64 of 1%; for the female sample, the SE is 0.43 of 1%. So the SE for the difference is $\sqrt{0.64^2 + 0.43^2} \approx 0.77$ of 1%. Then $z \approx (7 - 3)/0.77 \approx 5.2$ and $P$ is almost 0. This difference is almost impossible to explain as a chance variation.

7. The SE for the difference of the two averages can be estimated as $\sqrt{0.5^2 + 0.5^2} \approx 0.7$. So $z = (26 - 25)/0.7 \approx 1.4$, and $P \approx 8\%$. The difference could well be due to chance.

8. $z = 1/0.45 \approx 2.2$ and $P \approx 1.4\%$.

   *Comment.* The observed significance level depends on the sample size. With large samples, even small differences will be highly statistically significant. More about this in chapter 29.

9. The treatment and control averages are dependent, because the rats came in pairs from the same litter, so if one rat has a heavy cortex, the other one in the pair is likely to also. The SE calculation does not take this pairing into account.

   *Comment.* See review exercise 12 in chapter 26 for a better analysis. In each pair, take the difference "treatment − control." Make the $z$-test on the differences.

Set C, page 511

1. (a) Two numbers. The B-number is not observed; it says what his score would have been, if he had been assigned to the control group.
   (b) Yes. The A-number says what her score would have been, if she had been assigned to the coaching group. This number is not observed, because she was in the control group. The investigators do not know what the A-number was.
   (c) Take the conservative route. The SE for the coaching average is 9.8 points. The SE for the control average is 10.3 points. The SE for the difference is $\sqrt{9.8^2 + 10.3^2} \approx 14.2$ points. The difference in average scores was 9, so $z \approx 9/14.2 \approx 0.65$, and $P \approx 26\%$. This could easily be chance variation.

   *Comment.* Exercise 1 is not like comparing NAEP test scores in 1990 and 2004 (section 2), because we do not have two independent samples. But it is like the vitamin C experiment (example 4). Each of the 200 students has two possible responses—one if coached and one if not coached. The investigators get to see only one of the two responses, and make their choices at random. That is why the calculation of the SE is legitimate (pp. 509–511).

2. (a) The difference is $66 - 59 = 7$ points, and the SE is 1.8 points. So $z \approx 3.9$, and $P \approx 0$. This difference is hard to explain as a chance variation. Wheaties work!
   (b) The students will know what cereal they are eating, so it is hard to blind that aspect of the study. The grading of the final could be done blind. Consent to the study should be obtained before randomization, not after, to reduce selective drop outs.

3. (a) The difference is 1 point, and the SE is 1.75 points. This looks like chance variation. The two groups are comparable—the randomization worked.
   (b) Now the difference is 9 points, with the same SE of 1.75 points. So $z \approx 5$, and $P \approx 0$. Something went wrong in the randomization.

   *Comment.* The difference in (b) can't be explained as the result of eating Wheaties, because the cereal feeding did not start till after the midterm. See exercise 7 on p. 22. This exercise was hypothetical; for a real study on cornflakes, see N. Vaisman et al., "Effect of breakfast timing on the cognitive functions of elementary school students," *Archives of Pediatric and Adolescent Medicine* vol. 150 (1996) pp. 1089–92; eating breakfast improves your test scores.

4. (a) The difference between the two sample averages is 0.1, and the SE is 0.13. So $z \approx 0.8$ and $P \approx 21\%$. This looks like chance variation.

(b) There is a new batch of random numbers, and other factors might be at work too—weather, new cold viruses, etc. After all, the studies involve two different groups of people, at two different times.

5. (a,b) True.
   (c) False. The sample averages are dependent, so the square root law does not apply (section 1).

Set D, page 514

1. (a) $\boxed{0}\boxed{1}$
   (b) Form A, prefers surgery; form B, prefers radiation.
   (c) Only (ii).
   (d) The number of students who got form A was $84 + 112 = 196$; of these, $112/196 \times 100\% \approx 57\%$ favored surgery. Of the students who got form B, about 83% favored surgery. The difference between the percents is 26%, and the SE is about 5.2%. So $z \approx 5$, and $P \approx 0$. The difference is hard to explain as a chance variation.

2. "Percent" means "per 100," but the rates in this problem are so small that it is more convenient to express them per 100,000. The rate in the vaccine group was 57/200,000, or 28.5 per 100,000. The SE for the number of cases is

$$\sqrt{200{,}000} \times \sqrt{\frac{57}{200{,}000} \times \left(1 - \frac{57}{200{,}000}\right)} \approx 8$$

(See section 4 of chapter 17 for the shortcut method.) So the SE for the rate is 8/200,000 or 4 per 100,000. In the placebo group, the rate was 71 per 100,000, and the SE for the rate is 6 per 100,000. The SE for the difference in rates is

$$\sqrt{4^2 + 6^2} \approx 7 \text{ per } 100{,}000$$

The difference in rates is $28.5 - 71 = -42.5$ per 100,000. On the null hypothesis, the expected difference in rates is 0. So $z \approx -42.5/7 \approx -6$. The difference in rates cannot be explained as a fluke in the randomization. The vaccine works.

3. (a) $z \approx -2.4$, $P \approx 1\%$, significant. The difference is hard to explain as chance variation. Screening prevents death from breast cancer.
   (b) $z \approx -1$, $P \approx 16\%$, not significant. Breast cancer is rare: you don't see the impact of screening on the total death rate.

4. In the treatment group, 6.9% of the women experienced at least one event, compared to 7.1% in the control group. The difference is 0.2 of 1%. The SE for the difference is 0.7 of 1%. The difference is is not significant. The difference could easily be due to chance. The diet was not protective.

5. This question cannot be answered from the information given. The investigators do not have two independent samples, with one sample answering the question about Great Britain and the other the question about France. So the method of example 3 (p. 507) does not apply. The investigators have only one sample, and there are two responses for each student in the sample:

| | |
|---|---|
| 1 1 | found Great Britain and France on the map |
| 1 0 | found Great Britain; could not find France |
| 0 1 | could not find Great Britain; found France |
| 0 0 | could not find either country |

The investigators observe both responses when they score the test; that makes it different from the experiment in section 4, where only one of the two responses can be observed.

*Comment.* The question can be answered by using more advanced statistical methods, if you know the percentages in each of the 4 categories listed above.

6. (a) This is a straightforward two-sample $z$-test, as in section 2, because there are two independent simple random samples. The SE for the 2005 percentage is estimated as 1.6%; so is the SE for the 2000 percentage. The SE for the difference is computed from the square root law (section 1) as $\sqrt{1.6^2 + 1.6^2} \approx 2.2\%$. The observed difference is $41 - 50 = -9\%$. On the null hypothesis, the expected difference is 0%. So $z = (\text{obs} - \text{exp})/\text{SE} = -9/2.2 \approx -4.1$, $P \approx 2/100{,}000$. The difference is real. People are losing faith in the Supreme Court.
   (b) You can't tell. The method of section 2 does not apply, because you do not have two independent samples. The method of sections 3–4 does not apply, because you observe two responses for each person. See exercise 5 above.

7. (a) The difference in the two sample percents is 0.6% and the SE is about 3.6%. This looks like a chance variation. Withholding supplementation has no effect on breast feeding later.
   (b) The difference is 20.9 ml/day and the SE is 3.1 ml/day. This is almost impossible to explain as chance variation. Feeding patterns do seem to have been affected by different treatments in the nurseries.
   (c) The difference is 0.9% and the SE is 0.14%. So $z \approx 6.4$. Withholding supplementation increases weight loss: a bad side-effect.
   (d) The difference between the two sample averages is 27 grams and the SE is about 31 grams. This is chance variation: the randomization was successful.

*Comments.* (i) There is a tricky point in (c). Weight loss for each infant is measured in percent, relative to the birth weight. These percents are quantitative data, for which averages and SDs are computed.

(ii) The experiment shows that withholding supplementation does not promote breast feeding, and has a bad side effect—weight loss. The observational studies got it wrong. The explanation: there is an important confounding variable. Nurturing mothers are more likely to breast feed in the hospital, and their babies get less supplement. These mothers are also more likely to be breast feeding later, so there is a negative association between bottle feeding in the hospital and breast feeding later. But this association is driven by a third factor—the mother's personality.

## Chapter 28. The Chi-Square Test

Set A, page 531

1. (a) 90% (b) 10% (c) 1%

2. About 10%.

   *Comment.* Compare this with 1(c). As the degrees of freedom go up, the curve shifts to the right and spreads out, so there is more area to the right of 15.09 with 10 degrees of freedom than with 5. See figure 1.

3. $\chi^2 = 13.2$, $d = 5$, $1\% < P < 5\%$; actually, $P \approx 2.2\%$.

   *Comment.* $d$ = degrees of freedom. The data do not fit the model so well.

4. $\chi^2 = 1.0$, $d = 5$, $95\% < P < 99\%$.

5. $\chi^2 = 10.0$, $d = 5$, $5\% < P < 10\%$; actually, $P \approx 7.5\%$.

   *Comment.* Compare exercises 4 and 5. The observed frequencies just got multiplied by 10; this doesn't change the percents. But the result of the $\chi^2$-test depends on the sample size. With large samples, the $\chi^2$-test will reject very reasonable models. More about this in the exercises below and in chapter 29.

6. $\chi^2 \approx 18.6$, $d = 5$, $P < 1\%$—although for most purposes, the die is as fair as could be wanted; more about this in chapter 29.

7. (a) False; the $\chi^2$-test is preferred, see p. 524.
   (b) $\chi^2$, see p. 525.
   (c) True.
   (d) Expected; for instance, in line 1, the expected is $0.42 \times 66 \approx 27.7$; see p. 524.
   (e) The work for the $\chi^2$-test:

| *Age* | *Observed* | *Expected* |
|---|---|---|
| 21 to 40 | 5 | 27.7 |
| 41 to 50 | 9 | 15.2 |
| 51 to 60 | 19 | 10.6 |
| 61 and up | 33 | 12.5 |

   $\chi^2 \approx 61$, $d = 3$, $P \approx 0$. With simple random sampling, it is almost impossible for a jury to differ this much from the county age distribution. The inference is that grand juries are not selected at random.

   *Comments.* (i) The expected frequencies need not be whole numbers.
   (ii) Grand juries are nominated by judges, who prefer older jurors.

8. This is not a good method. The formula for $\chi^2$ involves frequencies—numbers not percents. Compare exercises 4 and 5 above.

9. (a) 12
   (b) You use $\chi^2$. The $\chi^2$ statistics: A) 15.2, B) 26.7, C) 7.5, D) 16.5. With 9 degrees of freedom, the 10% level is 14.68, the 5% level is 16.92, and the 1% level is 21.67. So A is marginal, B is way out of line, C is fine, D is marginal.
   (c) On retest, the $\chi^2$ for set A was 14.5, and for D it was 18.8. Reject D, and maybe A as well.

10. (a) The $\chi^2$-test will do the job.

(b) Can't be done: both boxes have the same fractions of 1's, 2's, and so forth; the test can't tell the difference.

Set B, page 534

1. Pooled $\chi^2 = 13.2 + 10 = 23.2$, $d = 5 + 5 = 10$, $P \approx 1\%$.
2. No, dependent experiments.
3. $\chi^2 \approx 0.5$, $d = 3$, $P \approx 8\%$. Inconclusive, but points to fudging.

Set C, page 539

1. It's fine. The method in the text is $(28/2{,}237) \times 1{,}170$. The method in the exercise is $(1{,}170/2{,}237) \times 28$. The result is the same, because $28 \times 1{,}170 = 1{,}170 \times 28$.
2.

| *Observed* | | | *Expected* | | *Difference* | |
|---|---|---|---|---|---|---|
| 2,792 | 3,591 | 6,383 | 2,730.6 | 3,652.4 | 61.4 | −61.4 |
| 1,486 | 2,131 | 3,617 | 1,547.4 | 2,069.6 | −61.4 | 61.4 |
| 4,278 | 5,722 | 10,000 | | | | |

$$\chi^2 \approx 6.7,\ d = 1,\ P \approx 1\%.$$

The expected frequencies are computed as on p. 538: for instance, the expected number of men who voted is $(4{,}278/10{,}000) \times 6{,}383 \approx 2{,}730.6$.

*Comments.* (i) 65% of the men voted, compared to 63% of the women. This is a small difference, but with a large sample it is accurately estimated. All $P$ tells you is whether the difference can be explained by chance. More about this in chapter 29.

(ii) A $2 \times 2$ table can be handled either by the $\chi^2$-test or by the $z$-test: note 3 to chapter 27.

3. Choose option (iv); the $z$-test is inappropriate because there are multiple categories, and the null hypothesis doesn't tell you what's in the box.

| *Observed* | | | *Expected* | | *Difference* | |
|---|---|---|---|---|---|---|
| 45 | 30 | 75 | 35.9 | 39.1 | 9.1 | −9.1 |
| 86 | 105 | 191 | 91.3 | 99.7 | −5.3 | 5.3 |
| 12 | 21 | 33 | 15.8 | 17.2 | −3.8 | 3.8 |
| 143 | 156 | 299 | | | | |

$$\chi^2 \approx 6.8,\ d = 2,\ P \approx 3\%$$

The expected frequencies are computed as on p. 538: for instance, the expected number of never-married men is $(143/299) \times 75 \approx 35.9$. In general, women marry earlier than men; and in the age group 25–34, more women than expected are married. ("Expected" means, on the null hypothesis that men and women have the same distribution of marital status.) The extra husbands are in higher age groups, like 35–44.

4. The Current Population Survey is not a simple random sample, the formulas do not apply, the clustering would have to be taken into account.
5. You are looking at averages, so it is time for the $z$-test not the $\chi^2$-test. There are two samples not just one, so option (ii) is right: $z \approx (\$50{,}000 - \$30{,}000)/\$3{,}000 \approx 6.7$, $P \approx 0$, the difference looks real. College grads make more money.

6. You are comparing a sample percent to an external standard, so option (i) is right: $z \approx (568 - 550)/15.7 \approx 1.15$, $P \approx 25\%$ (two-sided), the demographers' theory looks fine.

   You can also work this problem by method (iii): the box has 55 1's and 45 0's; there are 1,000 draws at random with replacement; make a $\chi^2$-test.

   *Comment.* When there are only two kinds of tickets in the box, you can use either the $z$-test or the $\chi^2$-test. The $\chi^2$-test will give the same result as two-sided $z$-test because $\chi^2 = z^2$.

7. Choose option (iii). Just because the data are laid out in a $2 \times 2$ table doesn't mean you're testing independence. The $\chi^2$-test is done below, there is only weak evidence against the null.

| | *Ways* | *Chance* | *Expected* | *Observed* |
|---|---|---|---|---|
| Even, large | 4, 6 | 2/6 | 200 | 183 |
| Even, small | 2 | 1/6 | 100 | 113 |
| Odd, large | 5 | 1/6 | 100 | 88 |
| Odd, small | 1, 3 | 2/6 | 200 | 216 |

$$\chi^2 \approx 6, \ d = 3, \ P \approx 10\%.$$

## Chapter 29. A Closer Look at Tests of Significance

Set A, page 546

1. (a) True. (b) True. See p. 482.
2. (a) False. (b) False. See p. 480–481.

Set B, page 550

1. (a) About 5. (b) 8. (c) About 1.

   *Comment.* If you toss 100 coins, you expect to get around 50 heads. If the null hypothesis is true, the chance of getting a "significant" result is 5%; so you can expect this to happen about 5 times in 100.

2. (a) 25 (b) $\boxed{0\,0\,1\,1}$

   (c) The sum of the ranks is like the sum of 25 draws made at random with replacement from the box $\boxed{1\,2\,3\,4}$.

3. (a) About 3. On the null hypothesis, the number of hits is like the sum of 25 draws from the box $\boxed{0\,0\,0\,1}$, so the chance of a "significant" result is about 3%. (The chance is given just before the exercise, on p. 551.)

   (b) About 5.

   (c) About 5.

4. Somewhat more than. The first test by itself has about a 3% chance of getting a "significant" result; the second and third each have about a 5% chance. But the chance that at least one of the three tests finds something is bigger than 5%.

   *Comment.* The trouble with data snooping is that it makes significance levels close to meaningless. Those who snoop, find—even when nothing is going on.

5. Data snooping again. If 25 different hypotheses are tested, some results are likely to be significant.

6. Two-tailed.

7. One-tailed.

8. (a) Yes; $P \approx 4\%$.
   (b) No; $P \approx 96\%$.
   (c) No; $P \approx 8\%$.

9. Doctors are more likely to write a journal article if they have an unusually high fatality rate, and that is more likely with a small sample—which leaves more room for flukes. As Chalmers says, "Physicians have a tendency to report the unusual."

Set C, page 554

1. (a) False. (b) False. See pp. 552–554.

2. The question makes sense, because we are dealing with simple random samples, and it can be answered by a two-sample $z$-test:

$$\text{SE for men's average} \approx 1, \quad \text{SE for women's average} \approx 1$$

$$\text{SE for difference} \approx \sqrt{1^2 + 1^2} \approx 1.4, \quad z \approx 1.4, \quad P \approx 8\% \text{ (one-sided)}$$

   This could be a chance variation.

3. The SE for each average is 0.5, so the SE for the difference is 0.7, $z \approx 2.8$, and $P$ drops to 1/4 of 1%.

   *Comment.* The observed significance level depends on the size of the sample. With the smaller sample, the difference was estimated as $2 \pm 1.4$ points; with the larger, $2 \pm 0.7$ points.

4. The second sentence is right. However, the null hypothesis can be rejected on the basis of a trivial difference—if the sample is large (p. 553).

5. $P = 27\%$. Big $P$ is good for null, small $P$ is bad.

6. (a) The test is legitimate. This is like the radiation-surgery example (section 4 of chapter 27).
   (b) If $P \approx 2\%$ (one-tailed) then $z \approx 2$. The difference is $71.5 - 25 = 46.5$ percentage points, so the SE must have been around 23 percentage points.
   (c) The difference between 71.5% and 25% is huge.
   (d) To see what the $P$-value adds, imagine the editors of the journal saying,

   > Look. Some of our reviewers are more critical than others. By the luck of the draw, too many critical ones were chosen to get the negative version.

   The $P$-value tells you that the editors cannot use the luck-of-the-draw defense with a straight face. The $P$-value does not help you compare 71.5% and 25%.
   (e) This study demonstrates publication bias. Reviewers are more likely to find mistakes in articles they disagree with—which is only human.

   *Comment.* The observed difference was 46 percentage points. The SE puts a give-or-take number of 23 percentage points on the estimate. The difference is big, but poorly estimated. (To get better accuracy, a larger sample would have been needed, and that might have been hard to arrange: there are only so many reviewers.) The $P$-value tells you that the difference would be hard to explain away as a chance variation.

7. The $P$-value does not measure the size of the difference, so there is no way of telling just from $P$ whether the impact is weak or strong: what determines $P$ is size relative to the SE.

8. A 99%-confidence interval is $-6.0 \pm 2.6$ SEs, that is, $-6 \pm 6.5$. The estimate is not very accurate. The $P$-value suggests that the elasticity is not exactly 0; nobody said it was. The use of tests seems questionable, and so is the model.

Set D, page 558

1. There are no probability samples here, so caution is in order. The students did rather well by comparison with the TAs.
2. Statistical significance does not make much sense here. The two inner planets do not constitute a random sample of size 2 from the population of inner planets. They *are* the inner planets. Similarly for the outer ones.
3. A test of significance is not appropriate here, unless a box model can be specified for the data.
4. The question makes sense, because we are dealing with a probability sample. However, it cannot be answered on the basis of the information given. This is a cluster sample, so the simple random sample formulas do not apply: section 4 of chapter 21 and section 5 of chapter 22.

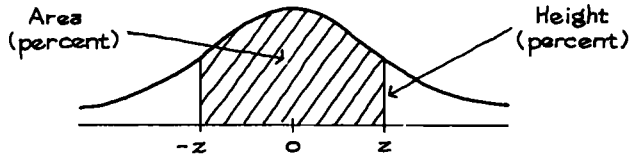
   *Comment.* In this study, like many others, children's performance on intelligence tests goes up with family income.
5. The sample is so large that unimportant differences are likely to be highly significant.

   *Comment.* The statistical procedures in this study may be open to question too.
6. A test of significance is being done on data for a whole population—the "elites." A box model does not make much sense here.

Set E, page 561

1. In the study, absenteeism was compared to that in previous years. But this year may be different from the last (milder weather, more interesting work, etc.). It would be better to compare the amount of absenteeism among workers on flex-time and among contemporary controls. To avoid resentment by those not given flex-time, it might be a good idea to assign whole work units to treatment or control.
2. This experiment was very well designed. It is fair to conclude that the vaccine protected the children against polio. Other explanations (like the placebo effect) are ruled out by the design of the experiment.
3. No. The $P$-value tells you that the increase is not a fluke in the random assignment of animals to treatment or control. The $P$-value does not help you extrapolate from high doses in rats to low doses in humans.
4. Where's the model? Why are lower salaries evidence of discrimination? (You might have to look at experience, education, productivity, etc.) And if this expert insists on doing a test, the pairs are very dependent—for instance, there could be one highly paid man who turns up in 16 of the pairs.

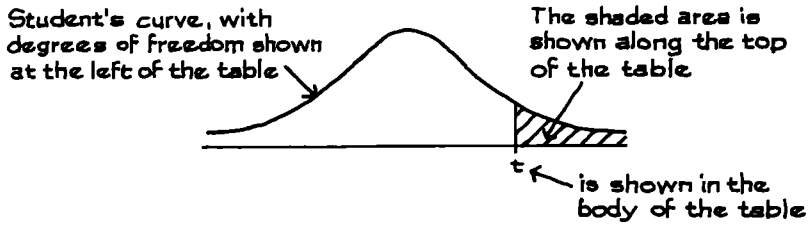
# Tables




A NORMAL TABLE

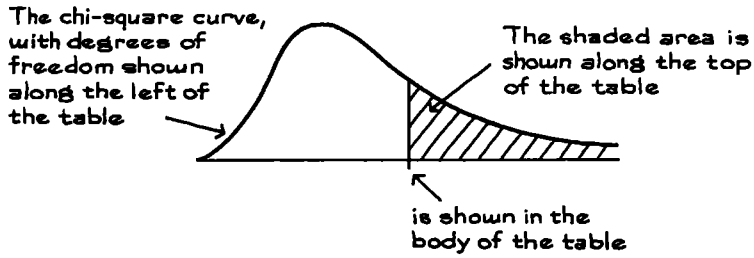
| *z* | *Height* | *Area* | *z* | *Height* | *Area* | *z* | *Height* | *Area* |
|---|---|---|---|---|---|---|---|---|
| 0.00 | 39.89 | 0 | 1.50 | 12.95 | 86.64 | 3.00 | 0.443 | 99.730 |
| 0.05 | 39.84 | 3.99 | 1.55 | 12.00 | 87.89 | 3.05 | 0.381 | 99.771 |
| 0.10 | 39.69 | 7.97 | 1.60 | 11.09 | 89.04 | 3.10 | 0.327 | 99.806 |
| 0.15 | 39.45 | 11.92 | 1.65 | 10.23 | 90.11 | 3.15 | 0.279 | 99.837 |
| 0.20 | 39.10 | 15.85 | 1.70 | 9.40 | 91.09 | 3.20 | 0.238 | 99.863 |
| 0.25 | 38.67 | 19.74 | 1.75 | 8.63 | 91.99 | 3.25 | 0.203 | 99.885 |
| 0.30 | 38.14 | 23.58 | 1.80 | 7.90 | 92.81 | 3.30 | 0.172 | 99.903 |
| 0.35 | 37.52 | 27.37 | 1.85 | 7.21 | 93.57 | 3.35 | 0.146 | 99.919 |
| 0.40 | 36.83 | 31.08 | 1.90 | 6.56 | 94.26 | 3.40 | 0.123 | 99.933 |
| 0.45 | 36.05 | 34.73 | 1.95 | 5.96 | 94.88 | 3.45 | 0.104 | 99.944 |
| 0.50 | 35.21 | 38.29 | 2.00 | 5.40 | 95.45 | 3.50 | 0.087 | 99.953 |
| 0.55 | 34.29 | 41.77 | 2.05 | 4.88 | 95.96 | 3.55 | 0.073 | 99.961 |
| 0.60 | 33.32 | 45.15 | 2.10 | 4.40 | 96.43 | 3.60 | 0.061 | 99.968 |
| 0.65 | 32.30 | 48.43 | 2.15 | 3.96 | 96.84 | 3.65 | 0.051 | 99.974 |
| 0.70 | 31.23 | 51.61 | 2.20 | 3.55 | 97.22 | 3.70 | 0.042 | 99.978 |
| 0.75 | 30.11 | 54.67 | 2.25 | 3.17 | 97.56 | 3.75 | 0.035 | 99.982 |
| 0.80 | 28.97 | 57.63 | 2.30 | 2.83 | 97.86 | 3.80 | 0.029 | 99.986 |
| 0.85 | 27.80 | 60.47 | 2.35 | 2.52 | 98.12 | 3.85 | 0.024 | 99.988 |
| 0.90 | 26.61 | 63.19 | 2.40 | 2.24 | 98.36 | 3.90 | 0.020 | 99.990 |
| 0.95 | 25.41 | 65.79 | 2.45 | 1.98 | 98.57 | 3.95 | 0.016 | 99.992 |
| 1.00 | 24.20 | 68.27 | 2.50 | 1.75 | 98.76 | 4.00 | 0.013 | 99.9937 |
| 1.05 | 22.99 | 70.63 | 2.55 | 1.54 | 98.92 | 4.05 | 0.011 | 99.9949 |
| 1.10 | 21.79 | 72.87 | 2.60 | 1.36 | 99.07 | 4.10 | 0.009 | 99.9959 |
| 1.15 | 20.59 | 74.99 | 2.65 | 1.19 | 99.20 | 4.15 | 0.007 | 99.9967 |
| 1.20 | 19.42 | 76.99 | 2.70 | 1.04 | 99.31 | 4.20 | 0.006 | 99.9973 |
| 1.25 | 18.26 | 78.87 | 2.75 | 0.91 | 99.40 | 4.25 | 0.005 | 99.9979 |
| 1.30 | 17.14 | 80.64 | 2.80 | 0.79 | 99.49 | 4.30 | 0.004 | 99.9983 |
| 1.35 | 16.04 | 82.30 | 2.85 | 9.69 | 99.56 | 4.35 | 0.003 | 99.9986 |
| 1.40 | 14.97 | 83.85 | 2.90 | 0.60 | 99.63 | 4.40 | 0.002 | 99.9989 |
| 1.45 | 13.94 | 85.29 | 2.95 | 0.51 | 99.68 | 4.45 | 0.002 | 99.9991 |

# A $t$-TABLE




| *Degrees of freedom* | *25%* | *10%* | *5%* | *2.5%* | *1%* | *0.5%* |
|---|---|---|---|---|---|---|
| 1 | 1.00 | 3.08 | 6.31 | 12.71 | 31.82 | 63.66 |
| 2 | 0.82 | 1.89 | 2.92 | 4.30 | 6.96 | 9.92 |
| 3 | 0.76 | 1.64 | 2.35 | 3.18 | 4.54 | 5.84 |
| 4 | 0.74 | 1.53 | 2.13 | 2.78 | 3.75 | 4.60 |
| 5 | 0.73 | 1.48 | 2.02 | 2.57 | 3.36 | 4.03 |
| 6 | 0.72 | 1.44 | 1.94 | 2.45 | 3.14 | 3.71 |
| 7 | 0.71 | 1.41 | 1.89 | 2.36 | 3.00 | 3.50 |
| 8 | 0.71 | 1.40 | 1.86 | 2.31 | 2.90 | 3.36 |
| 9 | 0.70 | 1.38 | 1.83 | 2.26 | 2.82 | 3.25 |
| 10 | 0.70 | 1.37 | 1.81 | 2.23 | 2.76 | 3.17 |
| 11 | 0.70 | 1.36 | 1.80 | 2.20 | 2.72 | 3.11 |
| 12 | 0.70 | 1.36 | 1.78 | 2.18 | 2.68 | 3.05 |
| 13 | 0.69 | 1.35 | 1.77 | 2.16 | 2.65 | 3.01 |
| 14 | 0.69 | 1.35 | 1.76 | 2.14 | 2.62 | 2.98 |
| 15 | 0.69 | 1.34 | 1.75 | 2.13 | 2.60 | 2.95 |
| 16 | 0.69 | 1.34 | 1.75 | 2.12 | 2.58 | 2.92 |
| 17 | 0.69 | 1.33 | 1.74 | 2.11 | 2.57 | 2.90 |
| 18 | 0.69 | 1.33 | 1.73 | 2.10 | 2.55 | 2.88 |
| 19 | 0.69 | 1.33 | 1.73 | 2.09 | 2.54 | 2.86 |
| 20 | 0.69 | 1.33 | 1.72 | 2.09 | 2.53 | 2.85 |
| 21 | 0.69 | 1.32 | 1.72 | 2.08 | 2.52 | 2.83 |
| 22 | 0.69 | 1.32 | 1.72 | 2.07 | 2.51 | 2.82 |
| 23 | 0.69 | 1.32 | 1.71 | 2.07 | 2.50 | 2.80 |
| 24 | 0.68 | 1.32 | 1.71 | 2.06 | 2.49 | 2.80 |
| 25 | 0.68 | 1.32 | 1.71 | 2.06 | 2.49 | 2.79 |

# A CHI-SQUARE TABLE




| *Degrees of freedom* | *99%* | *95%* | *90%* | *70%* | *50%* | *30%* | *10%* | *5%* | *1%* |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00016 | 0.0039 | 0.016 | 0.15 | 0.46 | 1.07 | 2.71 | 3.84 | 6.64 |
| 2 | 0.020 | 0.10 | 0.21 | 0.71 | 1.39 | 2.41 | 4.60 | 5.99 | 9.21 |
| 3 | 0.12 | 0.35 | 0.58 | 1.42 | 2.37 | 3.67 | 6.25 | 7.82 | 11.34 |
| 4 | 0.30 | 0.71 | 1.06 | 2.20 | 3.36 | 4.88 | 7.78 | 9.49 | 13.28 |
| 5 | 0.55 | 1.14 | 1.61 | 3.00 | 4.35 | 6.06 | 9.24 | 11.07 | 15.09 |
| 6 | 0.87 | 1.64 | 2.20 | 3.83 | 5.35 | 7.23 | 10.65 | 12.59 | 16.81 |
| 7 | 1.24 | 2.17 | 2.83 | 4.67 | 6.35 | 8.38 | 12.02 | 14.07 | 18.48 |
| 8 | 1.65 | 2.73 | 3.49 | 5.53 | 7.34 | 9.52 | 13.36 | 15.51 | 20.09 |
| 9 | 2.09 | 3.33 | 4.17 | 6.39 | 8.34 | 10.66 | 14.68 | 16.92 | 21.67 |
| 10 | 2.56 | 3.94 | 4.86 | 7.27 | 9.34 | 11.78 | 15.99 | 18.31 | 23.21 |
| 11 | 3.05 | 4.58 | 5.58 | 8.15 | 10.34 | 12.90 | 17.28 | 19.68 | 24.73 |
| 12 | 3.57 | 5.23 | 6.30 | 9.03 | 11.34 | 14.01 | 18.55 | 21.03 | 26.22 |
| 13 | 4.11 | 5.89 | 7.04 | 9.93 | 12.34 | 15.12 | 19.81 | 22.36 | 27.69 |
| 14 | 4.66 | 6.57 | 7.79 | 10.82 | 13.34 | 16.22 | 21.06 | 23.69 | 29.14 |
| 15 | 5.23 | 7.26 | 8.55 | 11.72 | 14.34 | 17.32 | 22.31 | 25.00 | 30.58 |
| 16 | 5.81 | 7.96 | 9.31 | 12.62 | 15.34 | 18.42 | 23.54 | 26.30 | 32.00 |
| 17 | 6.41 | 8.67 | 10.09 | 13.53 | 16.34 | 19.51 | 24.77 | 27.59 | 33.41 |
| 18 | 7.00 | 9.39 | 10.87 | 14.44 | 17.34 | 20.60 | 25.99 | 28.87 | 34.81 |
| 19 | 7.63 | 10.12 | 11.65 | 15.35 | 18.34 | 21.69 | 27.20 | 30.14 | 36.19 |
| 20 | 8.26 | 10.85 | 12.44 | 16.27 | 19.34 | 22.78 | 28.41 | 31.41 | 37.57 |

Source: Adapted from p. 112 of Sir R. A. Fisher, *Statistical Methods for Research Workers* (Edinburgh: Oliver & Boyd. 1958).

# Index


*Abilities of Man, The* (Spearman), 48–49
abortion and exercise, 23
accuracy:
  in estimating averages, 409–37, 441–57
  in estimating percents, 333–408, 436
  in measurement work, 97–109, 441–57
  in probability samples, 339–40, 355–94, 402–5, 415–37
  related to size of sample and size of population, 367–70, 373, 394, 553–54
  *see also* estimates
addition rule for calculating chances, 241–47, 254, 256
  *see also* multiplication rule
additivity, genetic, 461, 466–68, 471
advertising, 350, 351
age:
  and causes of death, 52
  census data on, 556
  cross-sectional vs. longitudinal data and, 77
  digit preference in reporting, 53–54
  distribution of, 50
  education and, 155
  effect of pill and, 45–48
  handedness and, 106
  health and, 566
  height and, 68–69
  height and weight related to, 59–60
  of husbands and wives, 168, 194, 567
  income and, 74, 95, 265, 366
  marital status and, 543
  of students at CUNY, 566
  as a variable, 13, 42, 43
AIDS, sexual behavior and, 570
air pollution and death rates, 568
air sampling for CO concentrations, 488–94
Almer and Jones, measurements of NB 10 by, 99
  *see also* NB 10
Alpert, Frank, 486–87
alternative hypothesis, 476–78, 489, 500, 504–6, 533, 534, 547–52, 560–61, 576
  defined, 477–78
  dependence as, 536
  *see also* null hypothesis; tests of significance; z-test
American Housing Survey, 50–51
areas under normal curve, 79, 82–85
asbestos and lung cancer, 574–76
associated with, 206
association:
  causation vs., 12–13, 16, 28, 149, 150–53, 206, 211–13
  linear, 126, 211
  negative, 128, 129, 131, 139–40
  nonlinear, 162–63, 165, 189, 195, 211
  positive, 120, 126, 127, 131, 139–40
  strength of, 121, 125–28, 139
  *see also* correlation; regression effect; scatter diagrams
average(s), 57–66, 76–77, 92–93, 109
  accuracy of, 409–37, 441–55, 456–57
  histogram and, 61–66, 75, 76, 96
  indicating center of histogram, 57–58, 61–65, 76
  of list of numbers, 59
  long-run, effect of bias on, 103–4
  median and, 64–65
  of sample, *see* sample averages
  SD and, 67–71
  treatment vs. control, 522
  weighted, 19–20
  *see also* law of averages
average of box, 289, 307, 326, 436–37
  defined, 289
  estimated from average of draws, 415–23, 436–37
  expected value for average of draws equals, 410, 436
average of draws, 409–28, 436–37, 450
  average of box estimated from, 415–23, 436–37
  chance variation in, 409–15, 475–78
  expected value for, 409–10
  expected value for difference of two, 501–3
  SE for, 410–19, 436–37, 476, 478–79, 508–11, 522
  SE for difference of two, 502, 521–22
  SE indicates how far from average of box, 416–17
average of measurements:
  estimating accuracy of, 441–55, 456–57
  model for, 450–52, 454–55
  SE for, 441–45, 451, 454–55, 500
  SE indicates likely size of chance error in, 441–45

baby boom, Great Blackout and, 498
baseball salaries, 429
beer consumption, 215

behavior:
as influenced by prediction, 520
rational, 512–14
Belmont and Marolla, 575
bends, the, 199
Berkeley, graduate admissions, 17–20, 556
Berkeley Institute of Human Development, 137
Berkson, J., 13, 199
Bernoulli, James, 308
Bertrand, Joseph, 273
bias:
confounding factors and, 17–20, 28
in measurements, 103–4, 109, 452, 509
in nonrandomized experiments, 4–8, 10
nonresponse, 336, 344, 353
panel, 398
publication, 521, 554–55
response, 348, 521, 554–55
in sampling, 333–48, 353–54, 355, 404–5
selection, 335, 353–54
*see also* confounding; samples, sampling
"Big Spin" game show, 529–30
binomial coefficients, 255–63, 268–69, 309
binomial formula, 259–61
bioassays, 552
biology teachers, 350–51
birth order and intelligence, 521, 575
birth weight, 16
blackjack, winning at, 307
blood, contaminated, 552
blood pressure, 61, 177, 201, 216, 264, 573
by age and pill use, 45–48
histogram for, 51–52, 75
body fat, 27
bookstores, 497
boot camp and recidivism, 27, 565
bootstrap, 377–79, 416
Bouman case, 264
box models (chance models), 278–84, 287, 288–304, 305–7, 330, 439–72, 555–60, 562–63
appropriate use of, 387–88, 402–3, 422, 445, 448, 454–55, 457, 509, 555–60, 562–63
chances defined by, 555–56, 562
defined, 279
making them, 281–85
for measurement error, 441–57
for net gain, 281–84, 289, 295–96
for null hypothesis, 476–78, 480–81, 486–87, 500, 555–63
and parapsychology experiments, 484–86, 523, 560
for randomized controlled experiments, 503–11
for sampling, 339–41, 348, 355–71, 373–74, 375–80, 387–88, 402–4, 415–22, 436–37
SD, short-cut for, 298–99
statistical inference and, 457
in tests of significance, 475–576
*see also* coin-tossing; dice; draws from a box; error box; Gauss model; SD; SE; sum of draws; tests of significance; zero-one boxes
Boyd, L. M., 407
Brahé, Tycho, 97
brain:
effect of psychological environment on, 498–500, 508
handedness and, 536
breast cancer, screening for, 22–23, 107, 515
breast-feeding, effects of, 516
Bureau of Labor Statistics, 395–408
Bureau of Standards, *see* National Bureau of Standards
Bureau of the Census, 50–51, 53–54, 366, 372, 395–408, 421, 445–46, 556, 564, 570
burglaries, 26
Bush, George H. W., 346, 391
Bush, George W., 367, 369, 371

calculators, effects of using, 519
calibration, 97–104
CALTRANS, freeway route choices, 555
Canadian National Breast Cancer Study, 107
cancers:
chemicals and, 552, 561–62, 574–76
diet and, 515
fat in diet and, 152
*see also specific forms of cancer*
*Canterbury Tales, The* (Chaucer), 392
capital punishment, opinion surveys on, 506
carbon monoxide (CO) concentrations, 488–94
cardiovascular disease, 515
Carnegie Commission, 427
cartesian coordinates, 111
car theft rates, 24–25
causation:
association vs., 12–13, 16, 28, 149, 150–53, 206, 211–13
*see also* association
CBS–*New York Times* poll, 559
Census Bureau, *see* Bureau of the Census
center and spread, 57, 76–77
central limit theorem, 325
certainty (100% chance), 222–23, 236
cervical cancer, 16, 23, 26
Chalmers, Thomas, 10, 21
chance(s), 221–69
addition rule, 241–47, 254, 256
are in procedure, not parameter or thing measured, 384, 417, 443, 457, 465
conditional, 226–27
empirical determination of, 222–23, 308–9
listing ways for, 237–40

long-run argument and, 222
multiplication rule in, 228–30, 249, 269
probability histogram and, 310
product rule for, 232, 254
of something, related to chance of opposite, 223, 248–50
tests of significance and, 480, 500, 555–61, 562–63, 576
used in controlled experiments, 5–6
*see also* frequency theory; probability theory
chance error, 172–73, 287, 302–3
law of averages and, 273–78
SE and, 288, 290–93, 307, 359–66, 376–77
chance errors in measurement, 97–109
in average of series, 441–45
bias and, 103–4, 452
box model for, 450–55
confidence intervals and, 443
SD indicates size of, for one measurement, 100–101, 442
SE indicates size of, for average of series, 442–43
chance errors in sampling, 348–49, 354, 355–74, 437
confidence intervals and, 381–83, 415–22, 443
in percentages, 375–80
SE indicates size of, 359–66, 373, 377–80, 403, 422–23
chance models, *see* box models
chance processes, 278–79
*see also* chance(s); frequency theory; probability theory
chance variability, 271–330
in average of draws, 409–15, 475–78
chi-square ($\chi^2$) and, 525–26, 535–40
regression lines and, 162
in repeated measurements, 97–101, 441–55
in sampling, 355–62, 409–22
in sampling compared to coin-tossing, 356–57
in sum of draws, 279–81, 287, 290–93
*see also* box models; SE
change of scale, 92–93
changing the box, 300–304, 359–66, 483–88, 505–6
*see also* classifying and counting; data; zero-one boxes
Chaucer, Geoffrey, 392
cheating in civil service exams, 54–55
cheating in high schools, 353
check weight, 100
chemicals, cancer and, 552, 561–62, 574–76
Chern, W. S., 564
chest diseases, 350
children, 128
big-city vs. rural, 553
"controlling" mothers and, 27
family income and, *see* income
heights of, 70
intelligence of and regression fallacy, 169
in Project Follow Through, 557–58
*see also* intelligence
chi-square ($\chi^2$) curves, 526–29, 531, 544
chi-square ($\chi^2$) statistic, 482, 525–44, 556
chi-square ($\chi^2$) table, 527, 531–32
chi-square ($\chi^2$) test, 482, 523–40, 542, 556
degrees of freedom for, 526–27, 531, 537–38, 544
for independence, 535–40, 544
for model, 523–31
*P*-value for, 526–29, 538
structure of, 530–33
*z*-test vs., 523, 529, 539–40
*see also* tests of significance
cholesterol and heart disease, 429, 550, 573
chromosomes, 461, 468–69
homologous, 468
randomness and, 468–69
civilian labor force, 399
SE in estimated size of, by half-sample method, 402–3
classifying and counting, 299–304, 359–62, 365, 483–88
*see also* changing the box; data; zero-one boxes
class intervals, 32, 35, 36–37, 43, 56
Clinton, Bill, 346
clofibrate trial, 13–14
cluster samples, 402–4, 408, 547
multistage, 340–41, 559
SEs for, 402–4
*see also* half-sample method; SE
coin-tossing, 222, 226, 230–31, 255, 432, 573
box model used for, 301–4, 487, 548
Kerrich's experiment on, 273–78, 302, 356–57, 449
normal curve and, 315–27
probability histogram and, 315–27
in randomized controlled experiment, 5
sampling and, 352, 356–57, 361, 371, 373, 393–94
*z*-tests for bias in, 548–50
college students:
hours worked by, 506–7
opinions of, 506
in reading tests, 554
Collins case, 233–34
colon cancer and diet, 26
comparison:
in controlled experiments, 3–11
in observational studies, 12–24, 45–47
computer simulation, *see* confidence intervals; dice; draws from a box; Gauss model

conditional probabilities, 226–27, 269
"confidence," reason for use of new word, 383–87, 416–18
confidence intervals, 381–86, 416–17, 437
  for average of box, 416–17, 437
  computer simulation of, 386, 418–19
  for exact value of thing measured, 457
  frequency theory and, 384
  interpretation of, 383–87
  normal curve used for, 381, 415–22, 437, 443, 457
  for percents, 381–83, 394
  SE and, 330
confidence levels, *see* confidence intervals
confounders, *see* confounding
confounding, 429, 566
  in controlled experiments, 4–5, 11
  in controlled experiments vs. observational studies, 19
  controlling for, 17–20, 28
  cross-tabulation for control of, 47–49, 56
  defined, 4, 20
  in observational studies, 13, 16, 20, 28, 46, 60, 206
  *see also* association; controlling for a variable
constitutional hypothesis, 20, 262
continuity correction, 318
continuous data, 43
Contraceptive Drug Study, 45–48
contradiction, argument by, in testing, 480
control groups:
  in controlled experiments, 3–11, 498–500, 504–11
  in observational studies, 12–24, 27–28, 45–48
  in $z$-test, 504–8
controlled experiments, 3–11
  defined, 11, 12
  historical controls vs., 8–10
  observational studies compared to, 12, 27–28, 217
  regression and, 207
controlling for a variable, 12–24, 45–47, 150–53
  by cross-tabulation, 47–49, 56
  by regression, 212–13, 217
  *see also* confounding; observational studies
controls, defined, 3, 12
controls, historical, 8–10
coronary bypass surgery, 9–10
Coronary Drug Project, 13–14, 22
coronary heart disease, *see* heart disease
correction factor, 367–70, 374, 412–13
correlation, 119–57
  association vs. causation and, 150–52, 157, 206
  computation of, 132–34
  graphical interpretation of, 125–29, 133–34, 144–48
  as misleading by itself, 147, 157
  as misleading when based on rates or averages, 149, 157
  nonlinearity and, 147, 148, 157
  outliers and, 147, 148, 157
  perfect (+1), 126
  positive vs. negative, 128, 134, 139–40
  as pure number, 141, 143, 157
  scale and, 141–42
  standard units for, 132–34, 140, 141
correlation coefficient ($r$), *see* correlation
Correns, Karl Erich, 458
craps, 310–13
  *see also* dice
Crossley poll, 337
cross-sectional surveys, 60
  longitudinal surveys compared to, 58, 77
cross-tabulation, 47–49, 56
  of unemployment numbers, 398–400
Current Population Survey, 31, 42, 76, 106, 156, 168, 395–408, 437, 539, 541, 543, 559, 564, 572, 574
  bias in, 398, 404–5, 407–8
  chance error in, 402–3
  chance of getting into sample, 398
  classifications and definitions in, 398–400
  current design of, 396–98
  function and scope of, 395–96
  households in, 397–98
  quality of data in, 404
  reinterview program in, 405
  rotation groups in, 398
  SEs for, 402–3, 407–8

data:
  discrete vs. continuous, 43–44
  qualitative vs. quantitative, 42–43, 364, 416, 485
  *see also* classifying and counting; variables
data-snooping, 547–50
death rates and air pollution, 568
decision making, rational, tests of theory, 512–16
degrees of freedom:
  in chi-square ($\chi^2$) tests, 526–27, 531, 537–38, 544
  in $t$-tests, 490–91, 493–94, 495
Deighton, Len, 251
de Méré, Chevalier, paradox of, 245, 248–50
Democrats, political polls and, 334–39, 361, 366, 367–70, 375, 382, 387, 390–91
de Moivre, Abraham, 78, 89, 221–24, 237, 308–10, 316

density scale, 38–42, 56
crowding shown by, 39–40
percents figured with, 40–41
dependent events, 230–33
*see also* independent events
dependent variable, in regression, 122, 158–61, 165–67, 196–97, 202–7
changes in, related to changes in the independent variable, 158–61, 205–6
DES (diethylstibestrol), 10
Descartes, René, 111
descriptive statistics, 31–116
contrasted with statistical inference, 333–34, 455
*see also* average(s); correlation; histograms for data; interquartile range; measurement error; median; normal curve; percentiles; regression effect; r.m.s.; scatter diagrams; SD; slope
design of study, test of significance and, 555–61, 562–63, 576
de Vries, Hugo, 458
Dewey, Thomas E., and the 1948 election polls, 337–39
dice, 222, 226, 242, 244, 259–60, 362, 371, 407, 431–32, 569
box models for, 279, 287, 300–301
computer simulation of, 310–13
in ESP experiment, 560–61
listing ways, 237–40
Paradox of Chevalier de Méré and, 248–50
probability histograms for, 310–12
real vs. ideal, 252
testing to see if loaded, 477, 523–24, 534, 540, 542–43
diet:
colon cancer and, 26
lung cancer and, 26
diethylstibestrol (DES), 10
discount pricing, effect on sales, 488
discrete data, 43–44
distribution tables, histograms and, 35–38, 43
DNA fingerprinting, 234, 565
*Doctrine of Chances, The* (de Moivre), 221–22
Doll, Sir Richard, 13, 148–50
dominance, genetic, 460, 471
double-blind experiments, 3, 6, 11, 13–14, 21–22
defined, 5
double-counting, 254
draws from a box (made at random), 223–25
average of, *see* average of draws
classifying and counting, 299–304, 359–62, 365, 483–88, 503–8
computer simulation of, 279–80, 292, 310–13, 355, 409, 418
square root law and, 446, 455
with or without replacement, 223–25, 230–33, 278–81, 360, 361–62, 363–66, 367–70, 372, 373, 374, 410–13, 481, 487, 503–4, 540–41
*see also* box models; probability histograms; sum of draws
dual-income families, 130
Dukakis, Michael, 391

ecological correlations, 148–50, 157
economic behavior, test of theory, 512–14, 519
"Ecstasy" (drug), 352
Edison, Thomas, 392
educational level:
age and, 155
blood pressure and, 201
distribution of, 38–40
histogram, 419
of husbands and wives, 156–57
income and, 126, 150–51, 161, 200, 202–7, 266, 559, 567–68
number of children and, 44, 128
occupation and, 574
Educational Testing Service, 156
Einstein, Albert, 458, 461
elasticity, 555, 564
election forecasting, presidential, 333–48, 389–91
in 1936, 334–36
in 1948, 337–39, 344
in 1952, 389–91
in 1984, 343–47
in 1988, 391
in 1992, 346, 390
in 2000, 371
in 2004, 367–70
elevation of earth's surface above sea level, 57–58, 455
empirical histograms, 311–13, 326–27, 418–19
employment, 153, 395–408, 541, 562, 563, 564
*see also* Current Population Survey; unemployment rates
*Employment and Earnings*, 396, 400
endpoint convention for class intervals, 35, 38–39, 43
Equal Opportunity Employment Commission, 264
error box, 382, 445–49, 450–52, 457, 489–90
amount of past vs. current data when estimating SD of, 451–52
description of, 450
SD of, 450–52
square root law and, 455
zero average of, 450, 457
*see also* Gauss model; measurement error
ESP experiments, test of significance in, 484, 487, 523, 551–52, 560–61

estimates:
for average of box, 415–37
in measurement work, 97–109, 441–57
parameters vs., 333–34
for percentages, 333–408
for SD of box, 375–79, 415–19, 437, 451, 476, 485, 489–92
tests vs., 486
*see also* accuracy; bootstrap; box models; inference in sampling; samples, sampling; SE; statistical inference
exact value of thing measured, 101, 450–52
bias and, 103–4
expected frequencies, 524, 525–26, 531, 534, 537–39, 542, 544
expected values, 288–307
for average of draws, 409–10
compared to observed values, 273–78, 292, 293, 378, 416
defined, 288
formula for, 288–89
and probability histograms, 315–16, 326, 330
for sample average, 410
for sample percentage, 359–62
SE and, 359–66
simple random sample and, 359
and standard units, 315–16
for sum of draws, 289, 290–93, 416
in z-tests, 478–80, 484–85, 548, 553
experiments, design of, 3–28, 555–60, 562–63, 576
*see also* controlled experiments; cross-sectional surveys; observational studies; randomized controlled experiments
extradition, 543
extrapolation in regression method, 166, 178
eye color, 463

"factorial" ($n!$), 257
families, dual-income, 130
family members, resemblances between:
and the regression fallacy, 170–72
shown in scatter diagram, 119–22
family size, 349, 575
Farley, James A., 391
fat in diet, cancer and, 152
Fermat, Pierre de, 248–49
Fisher, Sir R. A., 12, 13, 20, 180, 463–67, 533–34, 546
$\chi^2$-test used by, 533–34
fixed-level significance-testing, 492–93, 545–46
*see also* $P$-values; statistical significance; tests of significance
flex-time, testing reduction of absenteeism with, 482, 561
follow-back studies, 429
Follow Through, 557–58
framing of decisions, 514–16
France, kings and prime ministers of, 268
frequency tables, 524–25
frequency theory, 221–33, 236
chances are in the procedure, not the parameter, 384, 417, 457
confidence intervals and, 384, 394
games of chance and, 221–22, 236, 310–13, 530
null hypothesis and, 480
*see also* chance(s); draws from a box; probability theory

"g" (general intelligence factor), 48–49
Galileo, 239, 240
Gallup, George, 335, 336
Gallup poll, 335, 337, 338–39, 343–48, 389–91, 515–16, 574
biases in, 344, 348
interviewer control in, 346
1984 questionnaire used by, 343–47
nonvoters in, 344
probability methods and accuracy of, 342, 370–71
simple random sample compared with, 340, 389–90, 518
undecided in, 344
Galton, Sir Francis, 57, 119, 122
regression effect and, 170, 172, 465–68, 471
gamblers and gambling, 238–41, 248–50, 281–86, 306–7, 310, 330, 449
games of chance, *see* frequency theory; *individual games*
Ganzfeld experiments, 551–52
Gauss, Carl Friedrich, 202, 208, 451
Gauss model, 450–57, 472, 489, 572
applicability of, 451–52
computer simulation of, 448
square root law and, 446
genes, Mendel's discovery of, 458–61, 468, 533–34, 535
height and, 466–68
seed color and, 459–61, 463
genetics:
additivity, 461, 466–68, 471
chance model for, 458–72
chance variability in, 458–63, 468–71
regression effect in, 466–67
selective breeding and, 48–49, 572
geographic mobility, 429–30
Goldberger, Joseph, 16
Gore, Al, 371
Gossett, W. S., 488
grade point average, histogram for, 53
graduate admissions, bias in, 17–20, 556

graph of averages, 162–65, 178
in regression effect, 172
Gray-Donald, K., 516
Guinness Brewery, 488

half-sample method, 402–3, 408
Hammond, E. C., 13
handedness:
age and, 106
and average age at death, 429
death rates and, 566
sex and, 535–38
haphazard sampling, random sampling vs., 487
Health and Nutrition Examination Study (HANES), 58–62, 67–71, 80, 85, 87, 88, 94, 106, 158–62, 165, 169, 175, 180–81, 188, 210, 429–30, 535, 536
Health Examination Survey, 559
Health Insurance Plan of Greater New York (HIP), 22–23, 515
heart disease, 573
anticoagulants and, 497
cholestyramine and, 550
clofibrate trial and, 13–14
exercise and, 107
Multiple Risk Factor Intervention Trial on, 573
and smoking, 12–13, 22, 25, 42, 153, 262–63, 429, 573
surgery for, 9–10, 263–64
*see also* blood pressure; smoking, effects of
heat disease, diet and, 515
height:
age and, 68–69
average, 59–60, 67–68
genetics of, 466–68
histogram for, 52
histogram for, vs. normal curve, 80–81, 85–88, 94, 106
of husbands and wives, 430
SD of, 67–71
secular trend in, 60
weight and, 59–71
heights of fathers and sons:
regression effect in, 158–61, 165, 169, 170–72, 175, 180–83, 188, 193, 198, 207, 466–68
scatter diagram for, 119–22, 190–91, 568
height-weight relationship, 152–53, 154
graph of averages, 162
individual estimates, 165
regression method, 158–61
r.m.s. error, 180–87
scatter diagram, 159, 181, 188
heredity:
intelligence and, 48–49
Mendel's experiments on, 458–72, 533–34, 535
*see also* genetics; height; heights of fathers and sons
heteroscedasticity, 192, 197, 201
*see also* homoscedasticity; residuals, residual plot; scatter diagrams
HIP (Health Insurance Plan of Greater New York), 22–23, 515
histograms for data, 31–96, 109
average and, 61–66, 75, 76
balancing at the average, 62–64
center and spread of, 57–58
class intervals in, 32, 35, 36–37, 43–44, 56
cross-tabulation and, 47–48
crowding represented by height of, 39–40
defined, 31–32
density scale for, 38–42
empirical, 311–13
horizontal scale in, 31–32
median and, 64–65
normal curve and, 79, 80–81, 85–88, 95, 96, 102, 327, 419, 421
outliers in, 102–3, 109
percentiles of, 88–92
plotting of, 35–37, 43, 54–55
SD and, 57, 68–69, 96
symmetry and, 62–64
tails of, 64–65
variables and, 42–47
vertical scale for, 31, 38–41
vs. probability histograms, 310–15, 326–27, 418–19
*see also* normal curve; probability histograms; standard units
Holmes, Oliver Wendell, 545
Holmes, Sherlock, 333, 362, 375
homelessness, 569
homoscedasticity, 190, 201
*see also* heteroscedasticity
Hooke, Robert, 208–12
horizontal axis ($x$-axis), 35, 41, 110–16, 119
Horn, D., 13
household size, 427
hybrid seeds:
first-generation, 458–60, 464
second-generation, 459, 463, 464

idiopathic hypoguesia, 25
impossibility (0% chance), 222–23, 236
income:
age and, 74, 95, 265, 366
education and, 126, 150–51, 161, 200, 202–7, 216, 266, 366, 539, 559, 568
family, 75
of husbands and wives, 163, 184–85, 214–15, 430–31

income (*continued*)
intelligence and, 215
poverty and, 76
sex discrimination and, 562
income histograms, 31–33, 35–37, 40–41, 52–53, 65, 88–90, 106–7, 421
income tax, 371–72
independence, chi-square ($\chi^2$) test of, 535–40
independent events, 230–33, 234, 236, 241–47, 269
*see also* box models
independent variable in regression, 122, 158–61, 165–67, 196, 205
changes in dependent variable related to changes in, 158–61, 205–6, 217
inference, statistical, *see* inference in sampling; statistical inference
inference in sampling, 333, 375–90, 400, 415–37
*see also* statistical inference
intelligence, 137–38
birth order and, 521, 575
family size and, 349, 507–8, 575
of husbands and wives, 134–35, 175, 176–77
near-sightedness and, 266
selective breeding and, 48–49
test-retest of, 169, 172–73, 176
intelligence quotient (IQ), 215
intelligence testing:
of big-city vs. rural children, 553
of rats, 48–49
intercept, 113–14
of regression line, 202–7, 210, 216–17
International Bureau of Weights and Measures, 98
International Prototype Kilogram (The Kilogram), 98, 104
International Rice Research Institute, 207, 542
interquartile range, 57, 89
interview procedures, 335–36, 340–41
in Current Population Survey, 398, 405, 407–8
in Gallup poll, 340, 343–48
in quota sampling, 337–38
IQ (intelligence quotient), 215

Jablonski, David, 151
Japanese-Americans, 351
Jimmy the Greek, 282, 288, 449
jury studies, 496, 532, 541, 567
Just, R. E., 564

Kahn, H. A., 13
Kanarek, M. S., 574
Keno, 288, 289, 393
Kerrich, John, coin-tossing experiment, 273–78, 302, 356–57
Kerry, John, 367
Keynes, John Maynard, 221
Kilogram, The, 98, 104
Kolmogorov, A. N., 237
Kopans, Daniel, 107
Kramer, M. S., 516

Lanarkshire free milk experiment, 428
Landon, Alf (*Literary Digest* poll), 334–36
Laumann, E. O., 569–70
law of averages, 273–78, 287
chance error and, 275–78, 407
chance processes and, 278–79
compensation and, 287
Kerrich's experiment and, 273–78
normal curve and, 308–19
square root law and, 300–304
sum of draws and, 279–84, 300–304
lead levels, 215
least squares, method of, 208–11
least squares estimates, 211–13, 216–17
least squares line (regression line), 208
length of spring related to weight, 208–9, 210
Lévy, P., 237
Lewontin, R. C., 569–70
Lindberg, David, 151
line, algebraic equation for, 115–16
linear association, 126, 206, 211
*see also* nonlinear association
Lippman, G., 308
literacy rates, 507
*Literary Digest* poll, 334–36
liver cancer, 547
and smoking, 22
longitudinal surveys, cross-sectional surveys vs., 58, 77
long-run argument, 222, 241–47
long-tailed distributions, 64–65, 102–3, 419, 490–92, 527–29
lotteries, $\chi^2$ used in, 529–30, 532–33
low-fat diets, 515
LSAT scores, 88, 195–96
related to first-year scores, 195–96, 267
lung cancer:
asbestos and, 574–75
cigarette smoking and, 12–13, 148–50, 262–63
diet and, 26
surgery vs. radiation for, 512–15

mammography, 22–23, 107
marital status:
age and, 543
employment and, 541
mathematics tests, 506
math tests, 90–92, 94, 95, 106, 156, 176, 211, 267, 430, 569, 571

measurement error, 97–109
model for, 441–57
*see also* chance errors in measurement
measurements:
accuracy in, 97–109, 353, 441–57, 572
bias in, 103–4, 452, 509
equations for, 101, 103–4, 452
outliers in, 102–3, 109
replication of, 100, 108, 441–55, 456–57
SD of, 100–103, 109, 442–43, 451, 457, 489–92
median, 57, 77
average and, 64–65
histograms and, 64–65
Mendel, Gregor, 458–72, 533–34, 535
midparent height, 467
Montaigne, Michel de, 475
multiphasic checkups, 45
multiple regression, 212–13
Multiple Risk Factor Intervention Trial, 573
multiplication rule, 228–30, 269
*see also* addition rule
multistage cluster sampling, *see* cluster samples; SE
mutually exclusive events, 241–47

National Assessment of Educational Progress (NAEP), 392, 426, 436, 503–5, 517, 521, 571
National Bureau of Standards, 97–104, 449, 451, 453
*see also* NB 10
National Crime Survey, 573
National Foundation for Infantile Paralysis, 3–6, 21
NB 10, 98–103
average of measurements on, 106, 442–55
confidence interval for exact weight of, 444
Gauss model for measurements on, 450–55
negative association, 128, 139–40
neon signs and measurement error, 443–44
net gain, box models for, 281–84, 289, 295–96
Newton, Isaac, 255–56
Nielsen ratings, 350
nonlinear association, 147, 148, 157, 162–63, 165, 189, 195, 206, 211
*see also* linear association
nonrandomized experiments, bias in, 4–8, 10, 28
*see also* observational studies; randomized controlled experiments
nonresponse bias, 336, 344, 353
non-sampling error, 354
*see also* bias
normal approximation for data, 78–96
defined, 82
method for, 85–88
scope of, 319–24
*see also* histograms for data; normal curve; standard units
normal approximation for probability histograms, 294–96, 315–30
*see also* normal curve; probability histograms; standard units
normal curve, 78–82
area under, 79, 82–85
confidence levels and, 381, 415–22, 437, 457
equation of, 79
graph of, 79
histograms for data and, 79, 80–81, 85–88, 95, 96, 102, 327, 419, 421
law of averages and, 308–19
outliers and, 102–3, 109
percentiles for, 90–92
probability histograms and, 310–12, 315–30, 415–19, 437, 476
standard units and, 79–82, 86–87, 317–18, 325–26, 330, 481
sum of draws and, 294–96, 307, 322–27
symmetry, 79
used to figure chances, 294–97, 315–30, 418–19
used to figure percentages, 78–96
used in vertical strips of a scatter diagram, 195–98
used in z-tests, 479, 504–8, 548
validity depends on model, 556–60
vs. Student's curve, 490–91, 493
*see also* histograms for data; normal approximation for data; probability histograms; standard units
normal table, 82–84, A105
Northern Ireland:
religious discrimination in, 543
schools in, 566
null hypothesis, 477–88, 547–48
about average of box, 476–78, 480–81, 483, 500
about difference between the averages of two boxes, 504–8
box models and, 476–78, 480–81, 486–87, 499–500, 505–6, 513–14, 555–63
defined, 477–78
frequency theory and, 480, 533, 534
independence as (chi-square test), 536–37
model as (chi-square test), 523–31
*P*-values computed on basis of, 479–82, 500, 550
small *P*-values leading to rejection of, 480, 482–83, 484, 546
*see also* alternative hypothesis; tests of significance

observational studies, 12–28
association and causation in, 12–13, 16, 28, 150–53, 206, 211–13
comparison in, 12–24, 45–47
controlled experiments vs., 12, 27–28, 217
controlling for variables in, 17–20, 45–47, 150–53
controls in, 12–13
regression used in, 206, 212–13, 217
slope and intercept of regression line in, 206, 212–13, 216–17
observed frequencies, 524, 525–26, 531–32, 534, 536–38, 542, 544
observed significance levels, *see* *P*-values
observed values:
compared to expected values, 273–78, 292, 293, 378, 416
of test statistics, 478–81, 500, 504–5, 513–14, 521–22, 523, 542, 544
one-tail vs. two-tail tests, 547–50, 552, 576
opinion surveys, 393, 417–18, 518, 574
oral cancer, 428–29
oral contraceptives, 26, 45–48
outliers, 102–3, 109
in scatter diagrams, 147, 148, 157

*P*, *see* *P*-values
pain killers, placebos as, 5
panel bias, 398
parameters, 333–34, 346, 348, 353, 384–86, 394
*see also* accuracy; estimates; statistical inference
Pascal, Blaise, 248–50, 255
Pearson, Karl, 119, 175, 176, 190, 468
chi-square ($\chi^2$) and, 523–24
peas, Mendel's experiments with, 458–66
pellagra, 15–16
percentages, 4, 17, 24
accuracy of estimates for, 333–408
chances as, 222–23, 236
for confidence intervals, 381–83
in histograms, 32, 35–40, 56, 88–90, 96, 364, 365
in law of averages, 276, 277
normal curve and, 78–96
sample, *see* sample percentages
SEs for, 359–70, 373, 377–79, 381–83, 389–90, 402–3, 408
percentiles (percentile ranks), 88–92, 96, 166–69
Perot, Ross, 346
Phillips curve, 153
pizza consumption, 215
placebos, 5, 11, 14, 21–22, 25–26
planets, 558–59, 563
point of averages, 125, 126, 130, 140
regression line and, 160
polio epidemic, 3–4
*see also* Salk vaccine field trial
political cultures, differing, 559
political polls, 334–39, 367–71, 375, 382, 389–91, 407, 559–60
political repression and public opinion, 559–60
population, U.S.:
Census figures on, 421, 445–46, 556, 564
educational level in, *see* educational level
employed women in, 564
incomes of, *see* income histograms
population, vs. sample, 333, 353
population average, 422–23
population percentage:
confidence interval for, 381–86
defined, 378–79
*see also* parameters; percentages; sample percentages
portacaval shunt, 7–8, 428
positive association, 120, 126, 139–40
poverty, 76, 570
precincts, in election polls, 340–41
prediction:
association between variable and, 121–22
behavior and, 520
with regression estimates, 165–67, 180–85, 201
pretrial conferences, 108
price elasticity, 555, 564
Primary Sampling Units (PSUs), 397
prisoner recidivism rates, 27, 519–20, 565
probability histograms, 308–30, 352
area in, 311, 314–15, 330
for average of many draws, 411, 418–19, 421–22, 423, 428, 433–34, 436, 437, 476
chance represented by, 310
chi-square ($\chi^2$) curve and, 528–29
data histograms vs., 310–15, 326–27, 418–19
empirical histograms converging to, 311–13, 326–27, 329, 418–19
endpoints in, 317–18
as ideal histograms, 311–12
normal curve and, 310–12, 315–30, 364, 365, 366, 411, 415–19, 437, 476
for products, 323
reading, how to, 312–15
for sample averages, as converging to normal curve, 411–12, 418–19
for sample percentages, as converging to normal curve, 365–66
scaling and, 315–17
SE and, 315, 326, 330
*see also* normal approximation for data; normal curve; standard units; sum of draws

probability methods in sampling, 339–42, 354, 437
accuracy of, 339–40, 355–94, 402–5, 415–37
bias in, 342, 354, 404
chance error in, 354, 373
Current Population Survey, 402–4, 407–8
SEs and, 359–70, 373, 375–94, 402–4, 407–8, 415–37
*see also* cluster samples; Gallup poll; samples, sampling; simple random samples
probability samples, *see* probability methods in sampling
probability theory, 221–69, 437
binomial coefficients in, 255–63, 268–69
conditional probabilities in, 226–27, 269
independence vs. dependence in, 230–33
long-run argument in, 222, 241–47
*see also* chance processes; chance(s); frequency theory
probability waves, 322
product rule for calculating chances, 232, 254
*see also* addition rule for calculating chances; conditional probabilities
Project Follow Through, 557–58
publication bias, 521, 554–55
Public Health Service, 3, 58–59
*P*-values (in tests of significance), 479–82, 495–96, 500, 563
for chi-square ($\chi^2$) test, 526–29, 538
as inadequate summaries, 546, 576
meaning of, 480–81, 545–63, 576
normal curve and, 476, 556, 557
for one-sample *z*-test, 479–82
for one-tailed and two-tailed *z*-tests, 547–50
and rejecting the null, 480, 482–83, 547, 554, 560
sample size and, 553–54, 576
samples of convenience and, 556, 576
Student's curve and, 490–92, 493
for two-sample *z*-test, 504–8, 553
validity of model and, 555–60, 562–63
*see also* fixed-level significance-testing; statistical significance

qualitative vs. quantitative data, 42–43, 364, 416, 485
Quetelet, Adolph, 78
quota sampling, 337–38, 353
bias in, 338, 340, 353
probability methods vs., 339–42
ratio estimation vs., 346

*r*, *see* correlation
randomized controlled experiments, 5–8, 10, 11, 17, 516
box model for, 504–11
double-blind, 5–6, 13–14
historical controls vs., 8–10
in Salk vaccine field trial, 5–6, 508, 515
*z*-test and, 517
random number generator, 356, 484, 485
random sampling, 339–42, 346, 348, 355–59
haphazard sampling vs., 487
*see also* box models; chance errors in sampling; probability methods in sampling; samples, sampling; simple random samples
ratio estimation, 346, 401–5
rational behavior, tests of theory, 512–14
rational decision making, tests of theory, 512–16
rats:
effect of psychological environment on, 498–500
effect of selection for intelligence on, 48–49
Raven's progressive matrices, 349, 575
Rayleigh, Lord, 444
reading tests, 503–4, 521, 554
recidivism, 27, 519–20, 565
regression, law of, 465–68, 471
*see also* regression effect
regression, multiple, 212–13
regression coefficient, 466–67
regression effect:
defined, 169
graphical explanation of, 169–72
in test-retest situations, 169, 172–73, 179
regression equation, 205
regression estimates, 158–61, 165–69, 196–97, 202–6
extrapolation and, 166, 178
for individuals, 165–69
for percentile ranks, 166–69
*see also* regression lines
regression fallacy, 169–74, 179
defined, 169
regression lines, 159–65, 196, 202–17
chance variation smoothed away by, 162
defined, and compared to average, 160
graph of averages and, 162–65
and least squares, 208–11, 216–17
logarithmic transformation and, 197
nonlinear association and, 162–63, 165, 189, 211
and residual plot, 187–90, 201
r.m.s. error of, 180–87, 192, 201, 216
slope and intercept of, 202–7, 210, 216–17
there are two regression lines, 174–75
regression method, 158–61, 165–69, 178–79, 196–97
*see also* r.m.s. error for regression
"regression to mediocrity" (Galton), 170, 467

"reject the null," 480
religious discrimination in Northern Ireland, 543
replacement, drawing with or without, 367–70, 374, 410–13, 423–26, 428, 432–34, 436, 449, 463, 483, 500, 518, 551–52
replication of measurements, 100, 108, 441–55
Republican bias in polls, 338–39
Residential Energy Consumption Survey, 391–92
residuals, residual plot, 187–90, 201
response bias, 344, 521, 554–55
riots and temperature, 55
rise vs. run, 113, 203
r.m.s. (root-mean-square), 66–67, 71–72, 77
r.m.s. error for regression, 180–92, 208–10, 216
  and normal approximation, 195–98
  SD vs., 183
Rookies of the Year, 199–200
Roosevelt, Franklin D., 334–36
Roper poll, 337
Rosenzweig, Mark, 498, 499
roulette, 277, 304, 495–96
  bets at, 281–82
  box models for, 281–85
  chance of winning at, 281–83, 295–96, 432
  diagram of table, 282
  net gain in, 281–85
  "surrender' at, 569
Royal Oak, 250–51, 290
run vs. rise, 113, 203
Russell, Michael, 151

saccharin, bioassays of, 561–62
sale prices and volume, 488
Salk, Jonas, 3
Salk vaccine field trial, 3–6, 21, 25, 561
  bias in NFIP design for, 5–6, 21
  double-blind in, 5–6
  as randomized controlled, 5–6, 508, 515
Samaritans and suicide, 17
sample averages, 415–23
  difference between two, 501–8, 552–55
  observed vs. expected value of, 475–80
  probability histograms for, 411–12, 418–19
sample percentages:
  confidence intervals from, 383–87
  defined, 378
  expected values for, 359–62
  probability histograms for, 365–66
  SEs for, 359–70, 373, 375–80, 387–90, 394, 402–4, 408
samples, sampling, 278, 333–54
  absolute size of, 373
  accuracy of, 333–34, 342, 367–70, 373, 375–94, 402–5
  bias in, 333–48, 353–54, 355, 404–5, 408
  box model for, 339–41, 348, 355–71, 373–74, 375–80, 387–88, 402–4, 415–22, 436–37
  chance error in, *see* chance errors in sampling
  chance variability in, 355–62, 409–22
  confidence intervals and, 381–83, 416–17, 437
  of convenience, 424, 437, 556–58
  Current Population Survey, 395–408
  inference in, 333, 375–90, 400, 415–37
  multistage, 340–41, 397–98
  nonrespondents vs. respondents in, 336
  random, 339–42, 346, 348, 355–59
  size of, and accuracy, 367–70, 373, 394, 553–54
  size of, defined, 359
  splitting of (half-sample method), 402–3, 408
  terminology for, 333–34
  weighting of, 346, 401–5
  *see also* Bureau of the Census; cluster samples; Current Population Survey; Gallup poll; Health Examination Survey; Health and Nutrition Examination Study; National Assessment of Educational Progress; probability methods in sampling; quota sampling; ratio estimation; SD; SE; simple random samples; Student's curve
sampling error, 354
SAT scores, 90–92, 94, 95, 105–6, 156, 175, 176, 189, 211, 267, 430, 569
  first-year scores and, 165–67
  percentile ranks, 90–92
scatter diagrams, 119–40, 141–43, 148
  changing SDs in, 144–47
  as football-shaped clouds, 120–21, 125–26, 170, 196
  graph of averages for, 162–65, 172
  as heteroscedastic, 192, 197, 201
  as homoscedastic, 190, 201
  how to read, 119–22
  logarithmic transformation, 197
  nonlinear association in, 147, 148, 157, 162–63, 165, 189, 195, 211
  normal curve and, 195–98
  outliers in, 147, 148, 157
  plotting of, 119–20
  predictions and, 180–81
  regression effect and, 169–72
  regression line and, *see* regression lines
  residual plot and, 187–90
  r.m.s. error and, 180–82
  rough sketch of, 121
  SD line in, 130–31, 140, 158–59, 169–73
  summary statistics for, 125–28, 139

vertical strips in, 119–22, 158–59, 190–97, 201
*see also* correlation; regression effect; regression lines
school performance, changes over time, *see* National Assessment of Educational Progress
schools in Northern Ireland, 566
screening for breast cancer, 22–23, 107
SD (standard deviation), 67–77, 298–99
of box, and SE, 290–304, 307, 326, 359–60, 409–15, 437
of box, estimated from data, 375–79, 415–19, 437, 451, 476, 485, 489–92, 495
change of scale and, 92–93
computation of, 71–74
of error box, 451–54, 457, 476, 485, 489–92
and histograms, 57, 68–69, 96
horizontal, 125, 126, 144
of measurements, 100–103, 109, 442–43, 451, 457, 489–92
of measurements, vs. SE for average, 442–43
outliers and, 102–3, 109
regression method and, 158–61, 165–67, 169, 178–79, 196–97
r.m.s. and, 66, 71–72, 77, 183
of sample, 415–22, 437, 476–77, 506
of sample, vs. SE for average, 415–17
SD+ and, 74, 457, 493
SE contrasted with, 291, 416–17, 442–43
spread and, 57, 67–77
standard units and, 79–82
statistical calculator used for, 74
vertical, 125, 126, 144
of zero-one box, formula for, 292
*see also* box models; normal approximation for data; normal approximation for probability histograms; normal curve; SE; standard units
SD line, 130–32, 144–47, 158–59, 169–73
regression effect and, 164, 169–73
SE (standard error), 288, 290–307, 326, 330
for average of draws, 410–19, 436–37, 476–77, 478–79, 495, 508–11, 521–22
for average of measurements, 441–45, 451, 454–55, 500
for average of measurements vs. their SD, 442–43
for cluster samples (half-sample method), 402, 408
correction factor for, 367–70, 374, 412–13
for Current Population Survey, 402–5, 407–8
for difference of two averages, 501–22
for difference of two independent quantities, 501–3, 522
for percentages, 359–70, 373, 375–80, 387–90, 402–4, 408
and probability histograms, 315, 326, 330
for sample average vs. SD of sample, 415–17, 422–23
for sample percentage, square root law and, 360
SD contrasted with, 291, 416–17, 442–43
standard units and, 315
for sum of draws (square root law), 288–93, 307, 326, 330, 363
for sum vs. average, 410–12
in test statistics (z-tests), 475–523
validity of calculation depends on model, 387–88, 394, 402–3, 407–8, 457, 555–60, 562–63
*see also* box models; chance error; chance errors in measurement; chance errors in sampling; confidence intervals; normal approximation for data; normal approximation for probability histograms; normal curve; SD; standard units
Secchi depth, 430
secular trend in height, 60
seed color, genetics of, 458–62, 463, 464, 465, 470, 496
selection bias, 335, 353–54
selection ratio in Title VII litigation, 264
selective breeding, 48–49
sex bias:
in graduate admissions, 17–20, 556
in labor market, 562
sex cells, random pairing of, 468–71
sexual behavior, 569
AIDS and, 570
shortcut for SD of box, 298–304
sickle cell anemia, 471
"significance," *see* statistical significance
significance, tests of, *see* tests of significance
significance levels, *see* *P*-values
simple random samples:
accuracy of averages estimated from, 409–25, 441–57
accuracy of percentages estimated from, 375–94, 539
cluster samples vs., 340, 402–4
defined, 340, 354
formula for SEs depends on assumptions, 387–88, 394, 402–4, 407–8, 437
Gallup poll compared to, 340, 389–90, 518
SEs for, 359–70, 375–80, 387–90, 402–4, 407–8, 409–28, 436–37, 501–8
Simpson, Thomas, 441
simulation, *see* confidence intervals; draws from a box
slavery, 349–50, 351
slope, 113–16
of regression line, 202–7, 210, 216–17
of SD line, 131

smoking, effects of, 12–13, 22, 25, 42, 148–50, 153, 262–63, 573
snapdragons, genetics of, 461
social class, questionnaire response and, 336
Spearman, Charles, 48–49
spectrophotometers, 488
speed of light, 454, 456
Spock, Benjamin, 496
square root law, 291, 296, 300, 301–3, 307, 326, 360, 412, 502, 517
  applicability of, 388, 446, 517
  formula, 291–93
  measurement error and, 374
  and SE for sample percentages, 360
  statistical inference and, 446, 455
standard deviation, *see* SD
standard error, *see* SE
standard units, 79–82, 92–93, 96, 294–95, 315–17, 330, 419
  correlation coefficient and, 132–34, 140, 141
  normal curve and, 79–82, 86–87, 315–18, 325–26, 330
  regression method and, 166
  tests of significance and, 489–92
Stanford Research Institute (SRI), 557–58
statistical calculators, 74
statistical inference:
  chance model required for, 457
  defined, 333–34, 455
  *see also* inference in sampling
statistical significance, 478–82, 553–55, 562–63, 564, 576
  practical significance vs., 552–53, 555
statistical tables, testing and, 546–47, 555–56
  *see also* chi-square ($\chi^2$) table; normal table; $t$-table
statistics, sample, 333–34, 353
Stirling, James, 309
stochastic models, *see* box models
strata, in Current Population Survey, 397
Student's curve, 330, 490–92, 493–94, 500
suicide, Samaritans and, 17
sum of draws, 279–81, 287
  chance variability and, 279–81, 287, 290–93
  computer simulation of, *see* draws from a box
  expected value and SE for, 288–307, 326, 330, 416
  *see also* box models; classifying and counting; normal approximation for data; normal approximation for probability histograms; normal curve; probability histograms
*Swain v. Alabama*, 435
systematic error, *see* bias
$t$, *see* Student's curve; $t$-distribution; $t$-statistic; $t$-table; $t$-tests
Tart, Charles, 484, 486, 561
taxes, income, 371–72
$t$-distribution, 443, 488–95
Teasdale, T. W., 266
telephone surveys, bias and, 346, 348
test-retest situations, regression effect in, 169, 172–73, 179
tests of significance, 473–576
  alternative hypothesis and, *see* alternative hypothesis
  applicability of, 517
  argument by contradiction in, 480
  defined, 476
  for difference between average of box and external standard, *see* $t$-test; $z$-test
  for difference between averages of two boxes, *see* $z$-tests
  experimental design and, 555–60, 562–63, 576
  of independence, *see* chi-square ($\chi^2$) tests
  limitations of, 545–65, 576
  main idea of, 475–82
  for many hypotheses, 547–50
  meaning of chance and, 480, 500, 555–61, 562–63, 576
  of model, *see* chi-square ($\chi^2$) tests
  normal curve and, 479, 504–8, 548
  null hypothesis and, *see* null hypothesis
  observed significance levels, *see* $P$-values
  one-tailed vs. two-tailed, 547–50, 552, 576
  popularity of, 562–63
  $P$-values, *see* $P$-values
  questions answered by, 555–60, 562–63, 576
  for randomized controlled experiments, 504–11, 516
  role of model in, 555–60, 562–63, 576
  in Salk vaccine field trial, 508, 515
  sample size and, 553–54
  significance levels, *see* fixed-level significance-testing; $P$-values
  for small samples, *see* $t$-tests
  steps in making of, 482–83
  validity of model and, 555–60, 562–63, 576
  zero-one boxes in, 483–88, 504
test statistics, 476, 478–82, 490, 500, 504, 525–27, 544
3 × 2 tables, 535–38
Title VII litigation, 563
  *see also* Bouman
treatment groups:
  in controlled experiments, 3–11, 27–28, 498–500, 504–11
  in observational studies, 12–24, 27–28, 45–48
Treaty of the Meter (1875), 98

Truman, Harry, and the 1948 election polls, 337–39
Tryon, Robert, 48–49
Tschermak, Erich, 458
*t*-statistic, 482, 490
*t*-table, 546
*t*-test, 330, 488–95, 500
  *see also* tests of significance
Tversky, Amos, 512
Twain, Mark, 3
twins, identical, correlation between heights of, 126
twin studies, 262–63

Ullyot, Daniel, 263–64
Ultimate Sampling Units (USUs), 397, 402
unemployment rates, 395–408
  cross-tabulation of, 398–400
  SEs estimated for, 402–4

variables, 42–44
  continuous and discrete, 43–44, 56
  dependent, in regression, 158–61, 205–6, 217
  independent, in regression, 122, 158–61, 165–67, 196, 205
  qualitative and quantitative, 42–44, 56
  *see also* association; correlation; histograms for data; regression estimates
vertical axis (*y*-axis), 110–16, 119
  *see also* histograms for data
vertical strips in scatter diagrams, *see* scatter diagrams
vitamin C for prevention of colds, 21–22, 508–11
vitamins as cancer preventives, 26

Walker, Francis A., 395
wards, in sample surveys, 340–41
water clarity, 430
weighing, precision, 97–104, 442–55
weight:
  age and, 59–60
  at birth, 16, 516
  of brain, 498–500, 508
  of mice, 487
  relationship to height, 152–53, 154, 162
  of women, 61–62
weight-height relationship, *see* height-weight relationship
weights in sampling, *see* ratio estimation
*Who's Who Among American High School Students*, 353
WISC, 507, 553, 559
Wittgenstein, L., 523
Women's Health Initiative (WHI), 515
Wright, I. S., 497

*x*-coordinates, 110–16, 119

*y*-coordinates, 110–16, 119
Ylvisaker, Don, 529–30

*z*, *see* *z*-statistic; *z*-test
zero-one boxes, 300–304, 307, 359–66, 373, 377–78, 394
  SD of, 301, 307, 484
  tests of significance and, 483–88, 504
*z*-statistic, 479–82, 483, 500, 504, 522
*z*-test, 475–522, 548–50
  applicability of, 517
  chi-square ($\chi^2$) test compared to, 523, 529, 539–40
  for difference between average of box and external standard (one-sample *z*-test), 475–500
  for difference between averages of two boxes (two-sample *z*-test), 504–8, 509–11, 521–22, 552, 553, 556, 558
  formula, how to use, 484–86
  limitations of, 545–63, 576
  null hypothesis and, 479–81
  one-tailed vs. two-tailed, 547–50
  role of model and, 555–60, 562–63, 576
  for small samples, 488–95
  *t*-test compared to, 488–92, 522
  *see also* tests of significance