'avid Spiegelhalter is probably the greatest ving statistical communicator ... Read *The rt of Statistics* and learn. I did'
im Harford, author of *The Undercover conomist* and *Messy*

you want to develop the skills to see the orld as it is, and to tell it how it is – honestly nd seriously – this is the book'
lichael Blastland, co-author of *The Tiger That n't: Seeing Through a World of Numbers*

I. G. Wells is frequently quoted as having said hat "Statistical thinking will one day be as ecessary for efficient citizenship as the ability read and write." That day has certainly ome. This wonderful book provides a non-echnical and entertaining introduction to the asic tools of statistical thinking. Wells would ave approved'
rofessor Sir Adrian Smith FRS,
irector, Alan Turing Institute

ven those with expertise in statistics will find uch within these pages to stimulate the mind, nce Spiegelhalter combines clarity of thinking ith superb communication skills and a wealth f experience. A real tour de force which eserves to be widely read'
rofessor Dorothy Bishop, University of Oxford

wonderfully accessible introduction to odern statistics'
rofessor David J. Hand, author of
he Improbability Principle

BN 978-0-241-39863-0

63599

The Art of Statistics David Spiegelhalter

A PELICAN BOOK

The Art of Statistics
Learning from Data

David Spiegelhalter

CHAPTER 7

How Sure Can We Be About What Is Going On? Estimates and Intervals

How many people are unemployed in the UK?

In January 2018 the BBC News website announced that over the three months to the previous November, 'UK unemployment fell by 3,000 to 1.44 million'. The reason for this fall was debated, but nobody questioned whether this figure really was accurate. But careful scrutiny of the UK Office of National Statistics website revealed that the **margin of error** on this total was ± 77,000 – in other words, the true change could have been between a fall of 80,000 and a rise of 74,000. So although journalists and politicians appear to believe this claimed decline of 3,000 was a fixed, immutable tally of the entire country, it was in fact an imprecise estimate based on a survey of around 100,000 people.* Similarly, when the US Bureau of Labor Statistics reported a seasonally adjusted rise in civilian unemployment of 108,000 from December 2017 to January 2018, this was based on a sample of around 60,000

* When I once suggested to a group of journalists that this should be clearly stated in their articles, I was met with blank incomprehension.

households and had a margin of error (again rather difficult to find) of ± 300,000.*[1]

Acknowledging uncertainty is important. Anyone can make an estimate, but being able to realistically assess its possible error is a crucial element of statistical science. Even though it does involve some challenging concepts.

Suppose that we have collected some accurate data, perhaps with a well-designed survey, and we want to generalize the findings to our study population. If we have been careful and avoided internal biases, say by having a random sample, then we should expect the summary statistics calculated from the sample to be close to the corresponding values for the study population.

This important point is worth elaborating. In a well-conducted study, we expect our sample mean to be close to the population mean, the sample inter-quartile range to be close to the population inter-quartile range, and so on. We saw the idea of population summaries illustrated with the birth-weight data in Chapter 3, where we called the sample mean a statistic, and the population mean a parameter. In more technical statistical writing, these two figures are generally distinguished by giving them Roman and Greek letters respectively, in a possibly doomed attempt to avoid confusion; for example m often represents a sample mean, while the Greek μ (mu) is a population mean, and s generally represents a sample standard deviation, σ (sigma) a population standard deviation.

* Changes in unemployment derived from payroll data is based on employer returns and is somewhat more accurate, with a margin of error of around ± 100,000.

Often just the summary statistic is communicated, and this may be enough in some circumstances. For example, we have seen that most people are unaware that unemployment figures for the UK and US are not based on a full count of those officially registered as unemployed, but instead on large surveys. If such a survey finds that 7% of the sample are unemployed, national agencies and the media usually present this value as if it is a simple fact that 7% of the whole population are unemployed, rather than acknowledging that 7% is only an estimate. In more precise terms, they confuse the sample mean with the population mean.

This may not matter if we just want to give a broad picture of what is going on in the country, and the survey is huge and reliable. But suppose, to take a rather extreme illustration, that you hear that only 100 people were asked if they were unemployed, and seven said they were. The estimate would be 7%, but you probably wouldn't think it was very reliable, and you would not be very happy at this value being treated as if it described the whole population. What if the survey size were 1,000? 100,000? With a large enough survey, you may start feeling more comfortable with the fact that a sample estimate is a good enough summary. The sample size should affect your confidence in the estimate, and knowing exactly how much difference it makes is a basic necessity for proper statistical inference.

Numbers of Sexual Partners

Let's revisit the Natsal survey in Chapter 2, in which participants were asked how many sexual partners they had had in their lifetime. In the age band of 35–44 there were 1,125 female and 806 male respondents, so it was a large survey from which the sample summary statistics shown in Table 2–2 were calculated, such as the median number of reported partners being 8 for men and 5 for women. Since we know the survey was based on a proper random-sampling scheme, it is fairly reasonable to assume that the study population matches the target population, which is the adult British population. The crucial question is: how close are these statistics to what we would have found had we been able to ask everyone in the country?

As an illustration of how the accuracy of statistics depends on sample size, we shall pretend for the moment that the men in the survey in fact represent the population in which we are interested. The bottom panel of Figure 7.1 shows the distribution for the 760 men who reported up to 50 partners. For illustration, we then take successive samples of individuals from this 'population' of 760 men, pausing when we reach 10, 50 and 200 men. The data distributions of these samples are shown in Figure 7.1 – it is clear that the smaller samples are 'bumpier', since they are sensitive to single data-points. The summary statistics for the successively larger samples are shown in Table 7.1, showing that the rather low number of partners (mean 8.4) in the first sample of ten individuals gets steadily overwhelmed, as the statistics get closer and closer to those of the whole group of 760 men as the sample size increases.

Let's now go back to the actual problem at hand – what can we say about the mean and median number of partners in the entire study population of men between 35 and 44, based on the actual samples of men shown in Figure 7.1? We could estimate these population parameters by the sample statistics

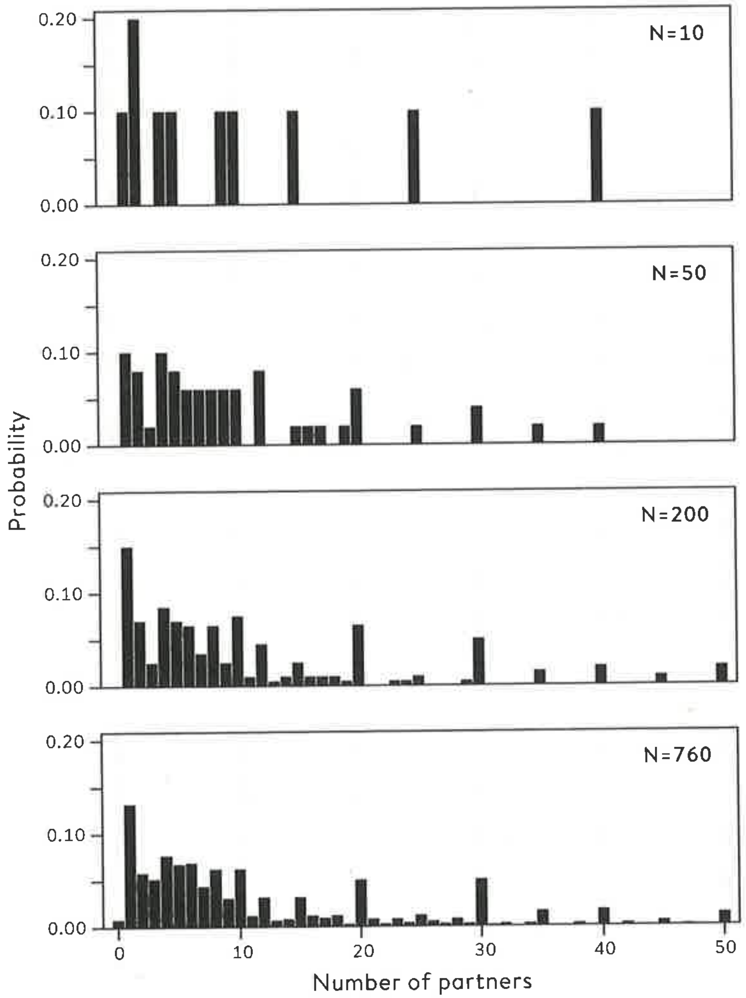

Figure 7.1
The bottom panel shows the distribution of responses of all 760 men in the survey. Individuals are successively sampled at random from this group, pausing at samples of size 10, 50, 200, producing the distributions in the top three panels. Smaller sample sizes show a more variable pattern, but the shape of the distribution gradually approaches that of the whole group of 760 men. Values above 50 partners are not shown.

Size of sample	Mean number of partners	Median number of partners
10	8.3	9
50	10.5	7.5
200	12.2	8
760	11.4	7

Table 7.1
Summary statistics for the lifetime number of sexual partners reported by men aged 35–44 in Natsal-3, for successively larger random samples and the complete data on 760 men.

of each group shown in Table 7.1, presuming that those based on the bigger samples are somehow 'better': for example the estimates of the mean number of partners are converging towards 11.4, and with a big enough sample we could presumably get as close as we wanted to the true answer.

Now we come to a critical step. In order to work out how accurate these statistics might be, we need to think of how much our statistics might change if we (in our imagination) were to repeat the sampling process many times. In other words, if we repeatedly drew samples of 760 men from the country, how much would the calculated statistics vary?

If we knew how much these estimates would vary, then it would help tell us how accurate our actual estimate was. But unfortunately we could only work out the precise variability in our estimates if we knew precisely the details of the population. And this is exactly what we do not know.

There are two ways to resolve this circularity. The first is to make some mathematical assumptions about the shape of the population distribution, and use sophisticated probability theory to work out the variability we would expect in our estimate, and hence how far away we might expect, say, the average of our sample to be from the mean of the population. This is the traditional method that is taught in statistics textbooks, and we shall see how this works in Chapter 9.

However, there is an alternative approach, based on the plausible assumption that the population should look roughly like the sample. Since we cannot repeatedly draw a new sample from the population, we instead repeatedly draw new samples from our sample!

We can illustrate this idea with our previous sample of

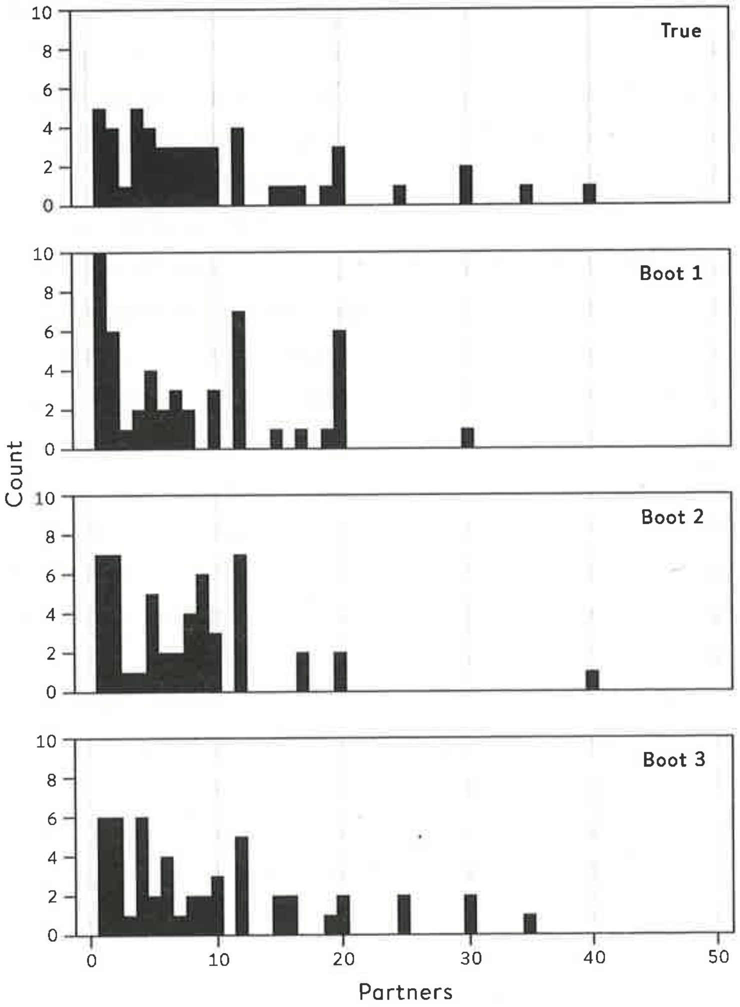

Figure 7.2
The original sample of 50 observations, and three 'bootstrap' resamples, each based on sampling 50 observations at random from the original set, replacing the sampled data-point each time. For example, an observation of 25 partners occurs once in the original data. This data-point was not sampled in the first or second bootstrap sample, but was sampled twice in the third.

50, shown in the top panel of Figure 7.2, which has a mean of 10.5. Suppose we draw 50 data-points in sequence, each time replacing the point we have taken, and get the data distribution shown in the second panel, which has a mean of 8.4.* Note that this distribution can only contain data-points taking on the same values as the original sample, but will contain different numbers of each value and so the shape of the distribution will be slightly different, and give a slightly different mean. This can then be repeated, and Figure 7.2 shows three such resamples, with means of 8.4, 9.7 and 9.8.

We therefore get an idea of how our estimate varies through this process of resampling with replacement. This is known as **bootstrapping** the data – the magical idea of pulling oneself up by one's own bootstraps is reflected in this ability to learn about the variability in an estimate without having to make any assumptions about the shape of the population distribution.

If we repeat this resampling, say, 1,000 times, we get 1,000 possible estimates of the mean. These are displayed as histograms in the second panel of Figure 7.3. Other panels show the results of bootstrapping the other samples shown in Figure 7.1, with each histogram showing the spread of bootstrap estimates around the mean of the original sample. These are known as **sampling distributions** of estimates, since they reflect the variability in estimates that arise from repeated sampling of data.

* Think of a bag of 50 balls, each labelled as one data-point from our sample of 50; for example, two would be labelled '25', four would be labelled '30', and so on. We pick one ball at random from the bag, record its value, and then replace it, restoring the number of balls in the bag to 50. We repeat this process of picking, recording and replacing a total of 50 times, producing a distribution of data-points such as 'Boot 1'.

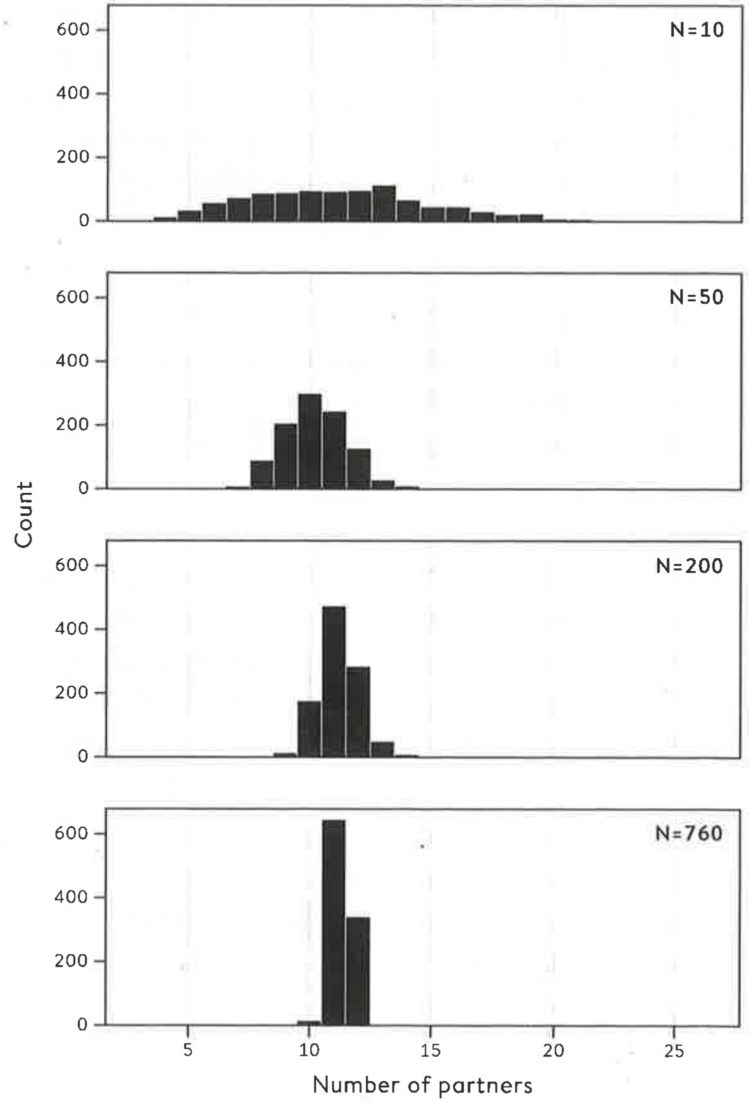

Figure 7.3
Distribution of sample means of 1,000 bootstrap resamples, for each of the original samples of size 10, 50, 200 and 760 shown in Figure 7.1. The variability of the sample means of the bootstrap resamples decreases as the sample size increases.

Figure 7.3 displays some clear features. The first, and perhaps most notable, is that almost all trace of the skewness of the original samples has gone – the distributions of the estimates based on the resampled data are almost symmetric around the mean of the original data. This is a first glimpse of what is known as the Central Limit Theorem, which says that the distribution of sample means tends towards the form of a normal distribution with increasing sample size, *almost regardless of the shape of the original data distribution*. This is an exceptional result, which we shall explore further in Chapter 9.

Crucially, these bootstrap distributions allow us to quantify our uncertainty about the estimates shown in Table 7.1. For example, we can find the range of values that contains the central 95% of the means of the bootstrap resamples, and call this a 95% uncertainty interval for the original estimates, or alternatively they can be called margins of error. These are shown in Table 7.2 – the symmetry of the bootstrap distributions means the uncertainty intervals are roughly symmetric around the original estimate.

The second important feature of Figure 7.3 is that the bootstrap distributions get narrower as the sample size increases, which is reflected in the steadily narrower 95% uncertainty intervals.

This section has introduced some difficult but important ideas:

- the variability in statistics based on samples
- bootstrapping data when we do not want to make assumptions about the shape of the population

Size of sample	Mean number of partners	95% bootstrap uncertainty interval
10	8.3	5.3 to 11.5
50	10.5	7.7 to 13.8
200	12.2	10.5 to 13.8
760	11.4	10.5 to 12.2

Table 7.2
Sample means for the lifetime number of sexual partners reported by men aged 35–44 in Natsal-3, for nested random samples of size 10, 50, 200 and complete data on 760 men, with 95% bootstrap uncertainty intervals, also known as margins of error.

- the fact that the shape of the distribution of the statistics does not depend on the shape of the original distribution from which the individual data-points are drawn

Rather remarkably, this has all been accomplished without any mathematics except the idea of drawing observations at random.

I now show that the same bootstrap strategy can be applied to more complex situations.

In Chapter 5 I fitted regression lines to Galton's height data, enabling predictions to be made of, say, a daughter's height based on her mother's height, using a regression line with an estimated gradient of 0.33 (Table 5.2). But how confident can we be about the position of that fitted line? Bootstrapping provides an intuitive way of answering this question without making any mathematical assumptions about the underlying population.

To bootstrap the 433 daughter/mother pairs shown in Figure 7.4, a resample of 433 is drawn, with replacement, from the data, and the least-squares ('best-fit') line fitted. This is repeated as many times as desired: for illustration, Figure 7.4 shows the fitted lines arising from just twenty resamples in order to demonstrate the scatter of lines. It is clear that, since the original data set is large, there is relatively little variability in the fitted lines and, when based on 1,000 bootstrap resamples, a 95% interval for the gradient runs from 0.22 to 0.44.

Bootstrapping provides an intuitive, computer-intensive

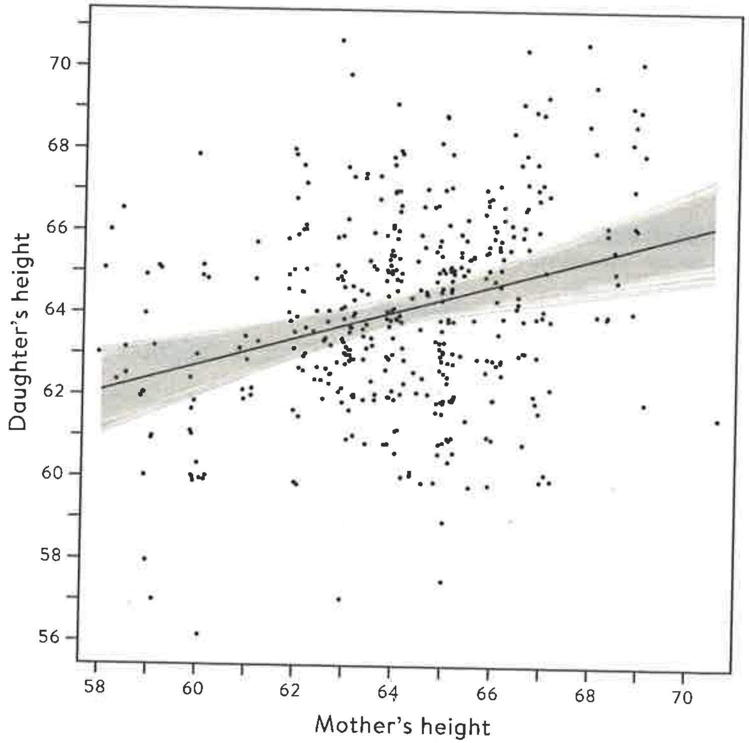

Figure 7.4
Fitted regression lines for twenty bootstrap resamples of Galton's mother–daughter height data superimposed on original data, showing the relatively small variability in gradient due to the large sample size.

way of assessing the uncertainty in our estimates, without making strong assumptions and without using probability theory. But the technique is not feasible when it comes to, say, working out the margins of error on unemployment surveys of 100,000 people. Although bootstrapping is a simple, brilliant and extraordinarily effective idea, it is just too clumsy to bootstrap such large quantities of data, especially when a convenient theory exists that can generate formulae for the width of uncertainty intervals. But before demonstrating this theory in Chapter 9, we must first face the delightful, but challenging, theory of probability.