CHAPTER 1

How Are We Doing? Data-Driven Views of Business Performance

1.0. INTRODUCTION: WHAT IS THE ISSUE?

Very few business outcomes are completely predictable. If they were, decision making would be so much easier. But as in life, we have to look at what the world throws at us and try to make sense of it.

If we collect data on a **business variable** then, because of the many factors that are not under our control, we will see a range of values. Here are some examples:

- Weekly **sales figures** over a year. The 52 numbers will vary greatly. Some of the variation we might attribute to calendar events like Christmas. But much of the variation is not predictable or easily explicable.
- **Daily returns** of a stock. In a year there might be 250 trading days, and so there will be 249 daily growth figures. These numbers are almost completely unpredictable. Typically just over half of them are positive and just under half of them are negative. You might think that good trading days occur in runs that can be used to predict tomorrow's result. You would be wrong.
- The **sale prices** of a large number of apartments sold in a particular area. Certainly we can anticipate that some properties will be worth more than others, but it all depends on who attends the auction and the available stock, and we will often be surprised at the price that is actually achieved.
- **Satisfaction ratings** of a surveyed list of customers. Every customer has a different experience with your business. Even if the experience is the same, every customer will respond differently.

Data-Driven Business Decisions, First Edition. Chris J. Lloyd.

When we have data on a business variable that we are interested in, the first thing we need is an idea of the range of likely values. Unless the data set is very small, it is quite impossible to understand the numbers without the aid of graphical or other summary techniques. In this chapter we look at producing a chart that describes the business variable and its so-called **distribution**. These charts are useful for quickly grasping the overall shape of a set of numbers, including what is a typical value, the overall spread of the data, and whether or not there are some unusual data values that are inconsistent with the others.

We also look at two simple numbers that can be used to describe this distribution—the **sample mean** and the **sample standard deviation** (SD). Breaking your data up into different segments and seeing how the mean and SD change is one simple way of getting a basic understanding of what drives the business variable.

M-TEL CALL CENTER: CASE BRIEFING

Esther Ching manages the complaints section of a call center, located in India. The call center is owned and run by M-Tel, a Malaysian telco company. Her section handles all billing complaints and service difficulties. The main products are landlines, mobile telephones, and Internet access.

She has a worksheet in front of her that contains a bewildering array of numbers, describing 1120 calls automatically stored by an automatic monitoring system. Each row describes a call. The variable *hold* records the number of seconds the customer was kept on hold before reaching the operator. The column headed *service* records the duration of the call in seconds. The column headed *rating* gives a satisfaction rating from very dissatisfied (VD) to very satisfied (VS). In addition, the gender of the customer and the consultant is recorded.

Esther has to prepare a report and one of the issues that has been previously raised is the role of gender in handling complaints. She is interested in contrasting the behavior of males. Are male customers less easy to please and less patient than female customers? Do customers respond better or worse to a consultant of the same gender, or doesn't it matter? What kinds of tables and charts might shed light on these questions?

As explained in the full case, there are also data on how long the consultant has been working for the call center. This could also have an effect on customer experience. But before we start, what is the main business outcome that Esther should care about?

1.1. SETTING OUT BUSINESS DATA

Most business data can be thought of this way. You have a large number of individuals or business items or **observational units** that have been included in the database. These might be different customers or different days of the year, or different apartments that we sold through an estate agent. For each of these, we have measured or recorded a set of numbers that we think are relevant to the business. These are the **business variables**. For instance, we might measure customer satisfaction and waiting time of customers; we might measure the day of the week and revenue for the dif-

ferent days of the year; or we might measure the sale price and floor area of the different apartments.

Typical Arrangement of Data

Each row is an observational unit and each column is a variable.

In Excel, it is common to arrange the data so that each row represents a different observational unit, and that each column represents a different variable. The first 10 rows of the call center data are displayed below. The observational units are different calls that were handled through the center. The variables are ID (just a label for the call) and five other variables as described in the case briefing. So for the first call, a male (M) customer called, waited 90s, was served by a female (F) consultant for 280s, and delivered a satisfaction rating of 3.

ID	Hold (secs)	Service	Rating	Customer	Consultant
84493	90	280	3	M	F
84496	88	305	4	F	F
84497	77	237	1	M	F
84499	94	261	3	M	M
84500	56	285	5	M	F
84501	59	267	4	M	F
84503	134	340	1	M	M
84504	108	315	4	M	F
84505	203	346	2	M	M
84507	71	213	1	M	M
..	..	..	..	..	..
..	..	..	..	..	..

This is the most convenient form for data to arrive. In practice, real data will often be spread over several sources and will contain wrong or missing data. For instance, you might have four different data sets depending on the gender of the customer and consultant. The spreadsheets FF, FM, MF, and MM contain these different subsets of the full data. For most purposes, it is best to have all the data on a single spreadsheet.

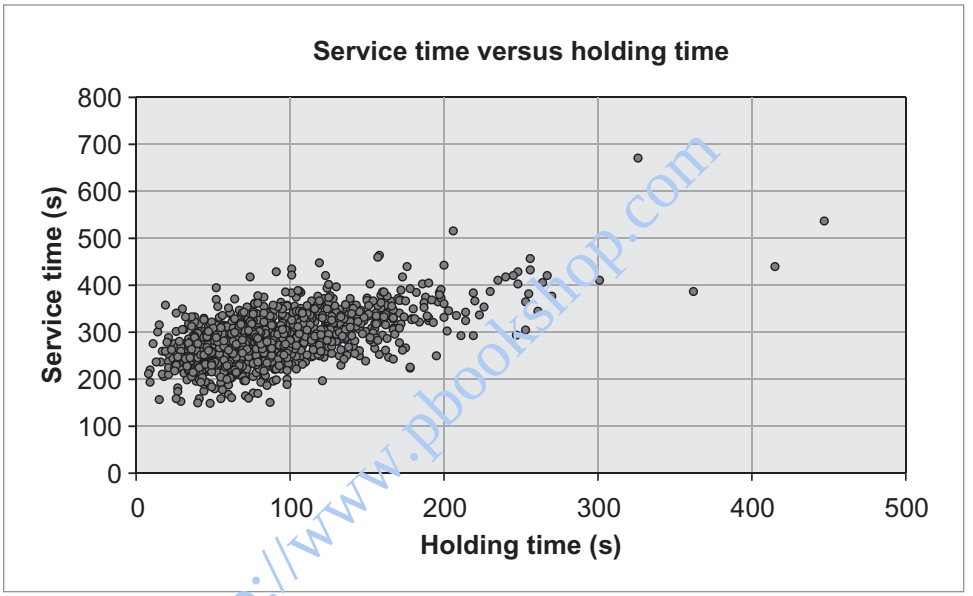
Finding Errors with Sorts

Sorting the data is a quick way of finding errors, because the erroneous values will often be extreme. To sort on a particular column, click on any cell in that column, go to the data menu, and then click either the AtoZ or ZtoA button. This will quickly show you the largest and smallest values in each column (which erroneous values will often be), as well as rogue characters like blanks and stray text.

If you sort on the consultant gender (column F), you will find that call 86428 has the gender recorded as D. This was obviously an error and since D is next to F on the keyboard, we might infer it should have been an F. Change it into an F now and store the file.

Finding Errors with Plots

Erroneous values can often be detected by plotting the data. The idea is that such points will probably stand out from the bulk of the data. For instance, to examine the holding and service times you might start by plotting one against the other. To do this, you select the two columns (including the headings), go to the Insert ribbon, and select the scatter plot chart. The resulting chart is below. Each point represents one of the 1120 calls. The horizontal axis is the holding time and the vertical axis is the service time. Excel always puts the leftmost columns on the horizontal axis and puts the rightmost column/s on the vertical axis.



The first thing you notice is that there appears to be some sort of relationship between the holding time and the service time. We might speculate why this would be so but it would only be speculation. Perhaps those who have a more important problem are prepared to hold for longer before hanging up and, when they finally reach the consultant, it takes longer to resolve. Or perhaps those who are forced to wait longer to reach the consultant are determined to get the most that they can out of their call and take the opportunity to resolve other minor problems and questions that may occur to them.

We can also see some unusual points to the right of the plot. These calls involved very long waiting times of the order of 5–8 min. These could possibly be misrecorded—for instance, 420 s might actually have been 42 s.

However, since the times were recorded electronically in this instance, they are unlikely to be wrong. Misrecorded data are probably less of a problem these days with electronic recording, but not all data are recorded electronically. When you have survey data that have been entered into a spreadsheet by hand, it is common to find wrong data values—often 10 times too large or 10 times too small because of a misplaced decimal point! Such errors are often easily spotted with a graph.

1.2. Different Kinds of Variables

Not all data can be taken at face value. A number is sometimes not even a number. Data come in different "flavors." Just as important, variables can be classified by their function in the business. The call center data nicely illustrate the different kinds of measurement scales and functions that arise in practice.

Different Flavors of Data

Ordinal Data The most interesting variable is Rating, which can be any whole number from 1 to 5. It is important to realize that these numbers are nothing more than labels. In particular, a rating of 4 does not mean that the customer was twice as satisfied as a rating of 2. The numbers 1–5 were actually just a numerical translation of the responses VD, dissatisfied (D), neutral (N), satisfied (S), and VS, as described in the case. Data of this kind are called rank or **ordinal** data. You can pretend these numbers are numbers if you like—for instance, you can calculate an average rating across all customers—but you need to bear in mind that the measurement scale is not objective. The practical difference between VS and S might be quite different to that between S and N. A more practically relevant analysis would assign appropriate numbers to the five ratings, perhaps reflecting the costs in terms of future customer retention and profits (which could perhaps be obtained from a follow-up study of these customers).

Categorical Data Gender is a **categorical** or **nominal** variable. It makes no sense at all to talk about average gender. Often variables like gender are recorded as 0 and 1 instead of as F and M. In this case, you can calculate the average of the 0's and 1's. This will give you the proportion of customers who are male, assuming M was coded as 1. But, generally speaking, you will not calculate averages of categorical data. Rather, such variables are used to break the data up into different business segments.

Interval and Numeric Data The holding and service variables are genuine numbers—measured in seconds. Such data are called **interval** data. Not only are these data interval, but zero does mean zero (even though there were no such calls). Such data are called **numeric**.

There is no ambiguity at all about what these numeric data mean, and you can calculate their average as well as other more complex statistics: you can rescale them, for instance, by dividing them by 60, which would change the units to *minutes*, and you can even do strange things like calculate their square! We will do this in the final chapters when we are trying to detect nonconstant growth in a business.

Different Business Functions of Data

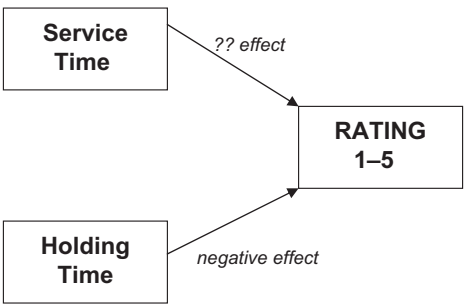
The key variables are the ones that measure the performance of the business. These are the ones you want to influence and manage. The point of collecting and analyzing the data is to understand the range of performance and to see what factors might be driving the performance. To begin building a conceptual model to analyze the data you need to think of what variables you really care about and what factors you think might affect these variables.

Performance Measures What constitutes a **performance measure** is subjective and depends on who is looking at the data. To illustrate, the customer's main performance measure is the Rating. Customers would presumably be happy with any changes that increased their satisfaction rating. Is holding time a performance measure? Certainly, the customer would prefer a shorter holding time, but this may already be accounted for in the rating. In fact, if you look at the case you will see that the rating does implicitly suggest that the customer should include their holding time in their assessment. So holding time is probably not a performance measure in itself, but it does affect the performance measure.

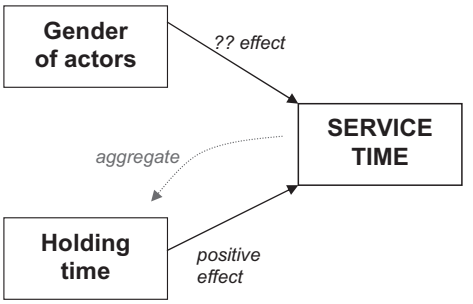
To look at performance from the call center manager's point of view, we would need to know her key performance indicators (KPIs). These would surely include maintaining an acceptable level of service to customers but will also surely involve controlling costs. So Esther would care about rating because it is one way to monitor if she is maintaining customer service. She would also care about service time, because this is a direct cost input. If consultants spend unnecessarily long periods of time on a call then they will not be as productive in terms of calls handled per hour. So from Esther's viewpoint there are two performance measures here. When there are multiple performance measures, then we often need to combine them somehow into a single measure of performance. It is not obvious how to do this in the present case. We talk about combining performance measures in Chapter 2.

Input or Driver Variables A simple metaphor for a business is as an industrial process that takes various inputs and produces outputs. This is a completely inadequate metaphor in many respects! But it is a helpful way to explain what I mean by **input variables**. These are variables that come logically before the performance measures that you care about and which we suspect will influence performance. You will see in the later chapters of this book that we typically use the symbol X for input variables and Y for performance measures.

Looking at Rating as the performance Y variable, the most obvious input variable is *holding time*. We suspect that longer holding times will reduce the Rating. Service time may also affect the Rating but it is not at all clear how. Customers will not want a longer or a shorter call for its own sake. They will want a call that takes precisely as long as is required to solve the problem. So it is not clear that service time will have a simple effect on rating. We will have to look at the data to see what it is. Certainly though, holding time and service time are inputs into the Rating. They come prior to the Rating and influence the Rating.



Looking at Service time now as a performance variable (remember it is a cost from Esther's viewpoint), what might influence it? Perhaps the gender of the main actors could have an effect; that is, the gender of the consultant and of the customer. What combinations of gender are associated with the longest and shortest service times? The data will tell us.



What else might influence service time? We already saw in the plot that longer holding times are associated with longer service times. So holding time is an input into service time. It comes before the service time and, perhaps surprisingly, seems to influence the service time. Looking at a different level though, an excessively long service time will directly affect the holding time—not of the present customer but of the next customer. So service time also affects holding time. We will not be able to see this in the present data because it is on the level of an individual customer. However, we might see it if we had aggregated data, say at the hourly level. We might expect to see longer holding times in those hours when the service times were systematically longer and the queue grew longer.

Modifying or Segmenting Variables Often we want to consider different market segments separately. For many products, gender is a natural segmentation. For segmenting variables like gender, we are often interested in first asking whether the performance measure is different for the different segments/genders. In this case, gender is really just a special kind of input variable—it is something that affects the performance.

More subtly though, we are often interested in seeing how the segmenting variable might itself affect the relationships between other variables. For instance, is the relationship between service time and rating different for the two genders? Segment variables are most commonly handled by simply breaking the data up into the different segments, though there are more sophisticated ways of measuring segment differences. These are explained and illustrated in Chapters 18–20.

Missing Variables It is essential to think whether you have forgotten to measure or collect important variables before you start the analysis. If you do not measure it, you can't account for it. Failure to account for critical factors can completely invalidate your analysis and lead to not only wrong but completely absurd conclusions.

As a simple example, you might look at how your sales vary from week to week as you increase and decrease your price. You could plot your sales against your price

and try to see the relationship. But if you did not measure your competitor's price, which may well be moving in response to yours, then you will get completely the wrong impression of how pricing policy affects your sales. There is an example of exactly this nature in Chapter 17.

1.3. THE IDEA OF A DISTRIBUTION

It is very easy to forget that nearly all business variables are uncertain and to mistakenly conceptualize a variable like *next month's sales* as a single figure. Of course next month's sales *will* be a single figure when you finally see it. But from the viewpoint of a month prior, it is an uncertain number and is best conceptualized quite differently. The **data distribution** gives you a tabular or graphical representation of the spread of likely values of the business variable. It is also visually helpful in understanding the sample mean and the sample SD, which will be introduced a little later.

You start with a column of observed values of your business variable—possibly a very long column! Several of the data sets in this book involve thousands of observational units. It is very difficult to understand a long column of numbers. The data distribution is a simple tabular summary. For simple variables, you can look at the table and see exactly what it says. However, most people can only hold half a dozen numbers in their head at a time. So when the variable is more complicated, even the data distribution can usefully be further summarized as a chart. Human minds can hold and process charts much more efficiently than tables of numbers.

What Is the Data Distribution?

Suppose we have a set of repeated observations on an unpredictable business variable, typically sitting in a column of an Excel spreadsheet.

DEFINITION

The **data distribution** or **frequency distribution** of a business variable is a list of *all observed values* of the variable together with the *number of times* these values are observed.

The relative frequencies will add up to 100%.

The number of times the value occurs is called its *frequency*. For instance, 193 customers might give you a satisfaction rating of 1 out of 5 on a particular survey. You would immediately then ask—193 out of how many? So commonly, we divide the frequency by the number of observational units—which is then called the *relative frequency*. If there were 1120 customers in the survey, the relative frequency would be 193/1120 = 0.171 or 17.1%. The data distribution would be a table of the five relative frequencies for the five possible satisfaction ratings.

The data distribution can be represented as a table with observed values ordered from smallest to largest in one column and frequencies in the adjacent column. When there are too many observed values the table itself starts to become complicated and hard to digest. In this case, it can be represented as a chart.

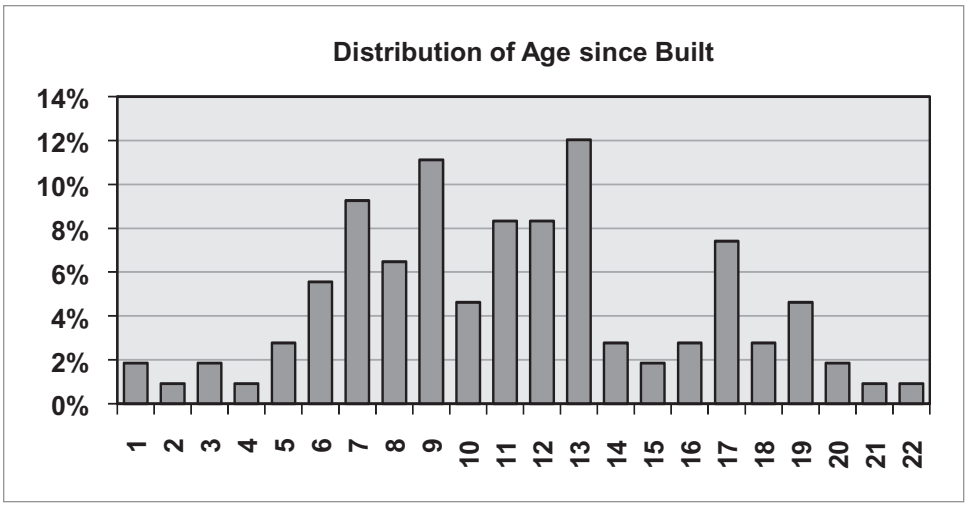
EXAMPLE 1 *Deepwater Bay real estate*

This data set, which is in the file ex2 (HKRE).xls, is analyzed extensively in later chapters. It lists the characteristics of 108 apartments that were sold in Deepwater Bay, Hong Kong, in 1995. Ultimately, we will be interested in explaining the sale price of these apartments in terms of their characteristics. In other words, the characteristics will be input variables and the performance measure will be sale price. For now, however, we are just interested in getting on overview of the characteristics themselves. The four characteristics listed are the number of bedrooms, the number of bathrooms, the number of dedicated car parking spaces, and the age of the apartment since construction.

Below is a tabular summary of the data distribution of the Beds, Baths, and Cars characteristics. Most of the apartments have two or three bedrooms, one or two bathrooms, and one car parking space. It is easy enough to look at this table and get the main message, so there is probably no reason to display it as a chart.

Number	Beds	Baths	Cars
0	0	0	1
1	0	42	96
2	54	62	8
3	44	4	0
4	9	0	0
5	1	0	0
Total	108	108	108

The age variable is a bit more complicated because there are more distinct ages represented in the data set. The youngest apartment was 1 year old, the oldest was 22 years old, and every age in between is represented by at least one apartment in the data set. This is best displayed as a chart.



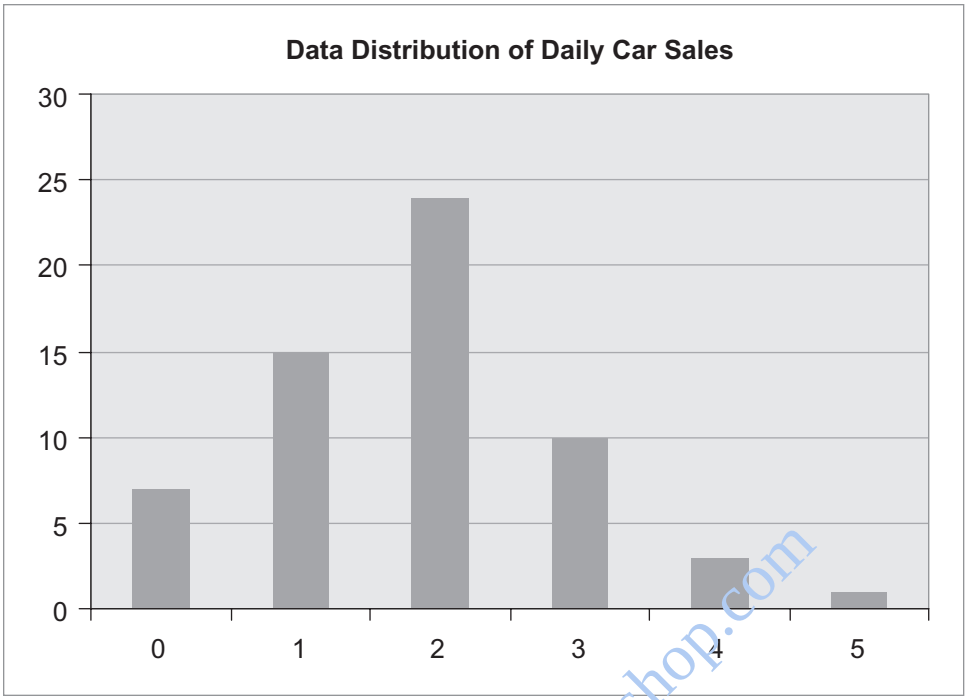
You can very quickly see from this chart what age most apartments are—from about 5 to 15 years. It is much more difficult to see the pattern in the table below, which is what the chart is representing in a visual form. The relative frequencies are just the frequencies divided by 108, and then displayed as a percentage. It is, to some extent, a matter of taste whether you use a table or a chart to summarize the data distribution. Very simple distributions like that of beds, baths, and cars can even be put together into a single table—as I have done above. The larger the number of distinct values in the data set, the more complicated and space taking the table will be, and you will start to think about using a chart instead.

Age	Freq	Rel Freq
1	2	1.9%
2	1	0.9%
3	2	1.9%
4	1	0.9%
5	3	2.8%
6	6	5.6%
7	10	9.3%
8	7	6.5%
9	12	11.1%
10	5	4.6%
11	9	8.3%
12	9	8.3%
13	13	12.0%
14	3	2.8%
15	2	1.9%
16	3	2.8%
17	8	7.4%
18	3	2.8%
19	5	4.6%
20	2	1.9%
21	1	0.9%
22	1	0.9%

CHECK YOUR UNDERSTANDING

Q1. The chart below describes daily car sales at a car yard over a typical 10-week period. Suppose that we had a full year's data.

(a) Roughly, how do you think this chart would change based on 52 weeks of data rather than 10 weeks? In other words, what view of *daily sales* would you get with an entire year of data?



(b) Consider *weekly sales*? This is a different performance measure. To get the data distribution you would look back over the year and see how many cars you sold each week. You would have 52 numbers and then you would convert them into a data distribution and then into a chart. What would this chart look like?

Producing the Table in Excel

The first step is to create the table of frequencies. There are several ways to achieve this using ordinary Excel. The simplest is to use the Excel function **COUNTIF**, which has the Syntax

=COUNTIF(data, criterion).

Here, "data" is the range of the spreadsheet containing the column of data values and "criterion" is the value or text string you are looking for. The function counts up how many of the values match the criterion. For numerical criteria, just enter the number you are looking for. For text strings, for instance if you want to count up how many males are in the data set and these are labeled "M" for males, then the text is entered with the double quotes.

Open the file ex2 (HKRE).xls and click on the tables tab at the bottom. You will see the frequency tables for Beds, Baths, and Cars displayed earlier and also below.

	D	E	F	G	I	J	K	L
row 1	BEDS	BATHS	CARS					
row 2	3	1	1		Number	Beds	Baths	Cars
row 3	2	1	0		0	0	0	4
row 4	2	1	1		1	0	42	96
row 5	4	1	1		2	54	62	8
row 6	2	1	1		3	44	4	0
row 7	2	1	1		4	9	0	0
row 8	3	1	1		5	1	0	0
row 9	2	1	0		Total	108	108	108

To create the frequencies for Beds, click on cell J3. This counts up how many 0's there are in the Beds data in column D. The data range is D2:D109 and the criterion you are searching for is the 0 in cell I3. So the formula is

=COUNTIF(D$2:D$109,$I3).

The reason for the "$" symbols is to fix the data range so that it will not move when the formula is dragged down and to fix the search value in cell I3 so that this will not move when we drag to the right. Your instructor will explain how to drag a formula and the use of the "$" symbol. The total in cell J9 is calculated using the internal Excel function SUM. The formula is

=SUM(J3:J8).

This should equal the number of apartments (here 108) and serves as a check that the frequencies have been calculated correctly. To calculate the frequencies for baths and cars, you simply drag the formulas in column J to the right. The data ranges will automatically move across to columns E and F. Because there is a "$" symbol before the "I" in the criterion field, these cell references will remain in column I, which is what we want because column I contains the numbers that we are searching for in the data.

A template, data_distribution.xls, has been provided, which will give both the tabulated data distribution and a chart when the raw data is copied into column A.

Producing the Chart in Excel

Because there are 22 different ages represented in the data set, the data distribution table for Age is probably too complicated to digest. To create the chart, open the worksheet Age and mark out the data in columns I and K from rows 2 to 24. To mark out disconnected ranges like this you need to mark out the first section, then hold down the control key, and then mark out the next section.

To create a chart you either click on the chart wizard (in XL2003) or go to the Insert ribbon. First, select the Scatter plot chart. This will plot the numbers 1–22 on

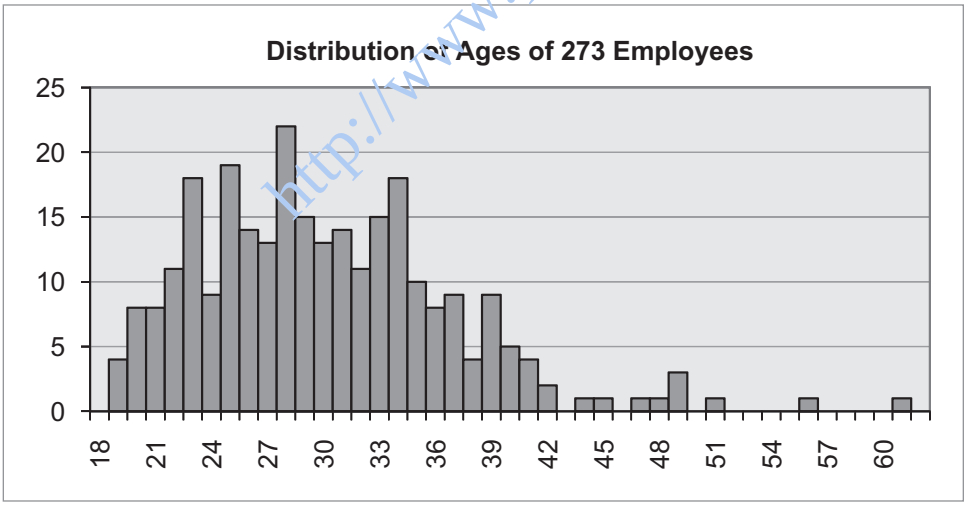
the horizontal axis and the relative frequencies on the vertical axis. Right click on any part of the chart and select "Change chart type" and change it into a column chart. This will produce the chart displayed earlier. All the cosmetic aspects of the chart can be customized by right clicking on the part of the chart you want to change—for instance, the axes, the external chart area, the internal chart color, the color of the bars, and how separated they are.

1.4. TYPICAL PERFORMANCE: THE SAMPLE MEAN

EXAMPLE 2 *Salary data*

The data in ex3 (salaries).xls describe 273 employees in a firm. The background of these data is in the case NTC Pay Equity in Chapter 18, but for the present it is enough to know that these employees were within a single business unit of the company, and that there was interest in relating their annual salaries to their age, experience, and other variables.

Below is a chart of the age distribution, measured to the nearest year. To summarize this distribution, it would be enough to say what a typical or central age was, and how much variation there is around this central value. These two features do not give you the chart exactly, but they give you the broad outline.



Apart from a few much older employees, the vast majority of the work force is between 20 and 40 years old. A rough central value looks to be around 30 years old and the vast majority of the workforce is within 10 years of this. There is an objective way of calculating the central figure and the typical deviation figure that we have just estimated by eye.

The **sample mean** or **sample average** of a variable is, as almost every one knows, the total of the observations divided by how many there are. It is denoted by the variable symbol with a "bar" above it. A general formula for the sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The Greek Σ sign means the "sum" of whatever is to the right of it. In this case, it means add up the individual observed values of the variable, and then multiply by the reciprocal of the number of observation, which we are calling *n*.

You can judge the sample average quite accurately when you have the data distribution in front of you as a chart. It is *the point where the chart would balance* if the bars were weights. Looking at the distribution of ages, the balancing point looks close to the earlier quoted central value of 30. When it is calculated, we find that the total of the 273 ages is 8221 so the sample mean is 8221/273 = 30.1.

Interpretation

The role of the sample mean is mainly as a single summary measure of the middle of a distribution. It gives you a typical or central outcome for the unpredictable business variable.

The sample average is not necessarily a value that the variable is likely to take. For instance, nobody in the data set actually had an age of 30.1! Even rounding the mean to the nearest whole year—30 in this case—we find that this is not the most common value in the data set. The sample mean measures the middle by the point where the data distribution would *balance*. Another way to think of it is that it is the amount per individual if the total of the data was evenly distributed between all the individuals. This does not make much sense for the present example of Age, but it would make sense if you calculate the average weekly revenue over a year.

Calculating the Sample Mean in Excel

To calculate the sample mean in Excel from a list of raw data values use

=AVERAGE(raw data values).

When the data come to you already expressed as a data distribution rather than in raw form, the mean is calculated using

=SUMPRODUCT(values, rel frequencies).

A template called frequency_distribution.xls has been provided, which will automatically calculate the average when you enter the values and the frequencies.

One way to see how the input variables affect the performance measure is to break the data up by different values of the input variable and see how the sample

mean performance changes. It is quite tedious to break the data set up into different segments and then calculate the average of each segment. There is a simple tool in Excel that does this automatically, which is explained in Section 2.3.

Other Location Measures: Median and Mode

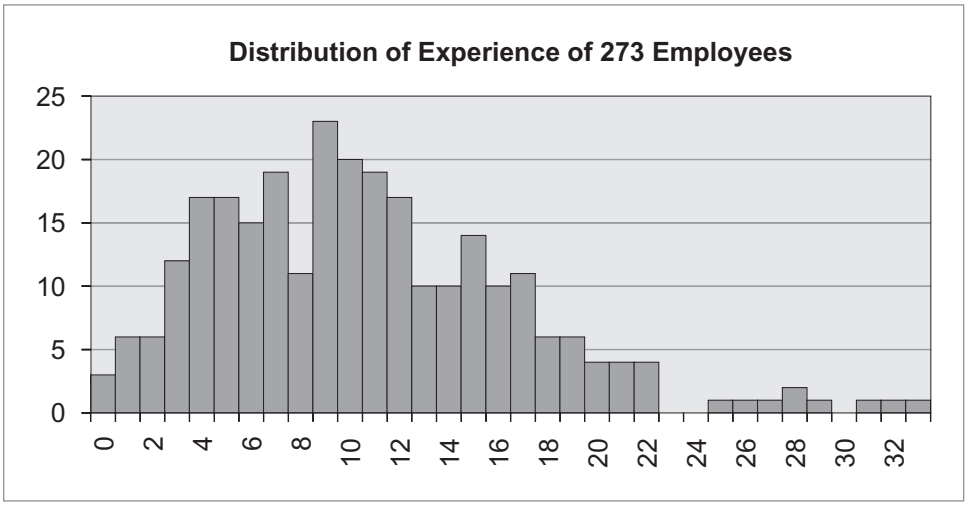
The **median** of a distribution is the point where the data are divided into the bottom 50% and upper 50%. It is the probabilistic middle of the data. In the case of the age data, there were 273 employees, so if we arranged them from youngest to oldest, the middle one would be the 137th. You can find the median by sorting the data and scrolling down to row 137. This person has an age of 29 so this is the median. However, there were several employees aged 29 so it is not quite right to say that 50% of the employees were older and 50% were younger than 29. But that is the rough interpretation of the median.

The **mode** of a distribution is the most commonly observed value. In contrast, the mean is not necessarily even a possible value of the variable! The mode of the age data is 28. There were 22 employees of this age and this is the largest frequency in the data distribution. You can see it clearly in the earlier chart.

You can calculate the median of a large data set by using the Excel function MEDIAN. There is no direct Excel function for the mode.

1.5. UNCERTAINTY IN PERFORMANCE: SD

Below is a chart of the data distribution of the *years of experience* of the same 273 employees mentioned in the previous section. You can see that the mean is around 10 years—it is actually 10.6 if you calculate it. But there is a great deal of variability around this figure.



What would be an objective way to measure the level of variability in a set of numbers? How about we look at how far each individual deviated from 10.6 years of experience? Some are very close to 10.6 years. Some are very far. The average of this deviation figure would be a sensible measure of variability. This is roughly what the

> The **sample SD** of a variable is a measure of the spread or variability of numbers in a sample. The square of the SD is computed by
>
> $$s^2 = \left(\frac{n}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
>
> The SD is the square root of this and denoted by s.

sample SD measures. It measures typical variation from the mean. Looking at the chart, you might guess that a typical deviation around the mean of 10.6 years is somewhere around 5 years. The SD is harder to judge from a chart than from a sample mean. This does not make it any less useful or valid.

The factor $n/(n-1)$ is a bit annoying. It makes the answer slightly larger than it would otherwise be. It also means that we will not be able to calculate the SD of a single number—which makes sense. When the number of data points n is large, the factor makes hardly any difference at all.

In Excel the function STDEV(raw data values) gives the sample SD. The SD of the experience data is 6.2. When the data come to you already as a data distribution, you can use the template frequency_distribution.xls as mentioned earlier.

Together, the sample mean and SD give you the main features of the data distribution—the middle and the spread. As soon as I tell you that the mean experience is 10.6 years and the SD is 6.2 years, you can imagine in your head what the data distribution will roughly look like.

Interpretation of SD

One SD on either side of the mean may be considered the *typical* range of the data. There is absolutely nothing unusual about a data point being 1 SD above or below the mean. In most cases you will find that between 65% and 75% of the data are within this range.

Two SDs on either side of the mean may be considered the *reasonable* range of the data. In most cases you will find that between 90% and 98% of the data are within this range. Data outside this range are very unusual and may merit some investigation. If the variable is a performance measure, you would want to know why performance is more than 2 SDs above or below the mean performance level.

Other Measures of Variability

A common alternative measure of variability or spread is the **interquartile range** (IQR). This is based on the idea of breaking the data up into the bottom 25%, the

top 25%, and hence the middle 50%. The IQR is the length of the middle 50% of the data. For instance, if the median salary is $28,500 per annum (pa), if 25% of people earn less than $16,300 pa and 25% earn more than $65,700 pa, then the IQR is $65,700 − $16,300 = $49,400.

For the experience data, sort the employees from least experienced to most experienced. Scroll down 25% of the way, which is about the 68th least experienced. This person has 6 years experience. Now look 68 rows from the bottom. The 68th most experienced person has 14 years experience. The difference is 14 − 6 = 8, so the IQR = 8.

In Excel you can calculate IQR by finding the so-called first quartiles, the point where 25% of the data are smaller, and the third quartile, the point where 25% of the data are larger. Then type

=QUARTILE(raw data,3) − QUARTILE (raw data,1),

which subtracts one from the other.

Another measure of variability is the **mean absolute deviation** (MAD), which is the average difference between each individual value and the mean—as described in the first paragraph of this section. For subtle mathematical reasons, this is not as popular or useful a measure of variability as SD. For the experience data it equals 4.8. It typically tends to be about 25% smaller than the SD. The Excel function for this is AVEDEV.

All the measures—SD, IQR, and MAD—measure typical variation on an absolute scale. Whether a given deviation is considered large or small depends on the context. But one natural benchmark is to compare the variability to the typical value of the business variable. For instance, a $32,000 variability in weekly revenue would be very large if a typical weekly revenue is $100,000, but would be practically negligible if the weekly revenue is $10,000,000.

Open the file ex3 (salries).xls with the data for the 273 employees. The column headed *salary* gives the annual salary of these employees in units of US$1000. The SD is $13.5K. This should be compared with the typical or mean salary, which is $74.3K.

DEFINITION

The ratio of the SD to the mean is called the **coefficient of variation** (CV). It measures typical variation as a proportion of a typical value of the variable. It is usual to quote it as a percentage.

CV = Sample SD/Sample Mean

= $13.5K/$74.3K = 0.182

or 18.2%. We say that there is an 18.2% variability in annual salaries of these 273 employees. This is quite a high level of variability and we expect that we will be able to explain much of these differences by looking at key variables such as age,

experience, and other variables. These other variables are given in the NTC Pay equity case in Chapter 18.

Check Your Understanding

Q2. You have 52 weekly percent growth figures for three different high-tech stocks, namely Cisco (CSCO), Microsoft (MSFT), and Hewlett Packard (HPQ) for the calendar years of 1999 and 2000. Below I have listed various summary statistics about the weekly growth of each stock. Each summary number is describing the 104 weekly growth figures over the 2 years. In the fourth column, I have given summaries of the growth performance of a hypothetical portfolio where you put 50% of your money in CSCO, 40% of your money in MSFT, and 10% of your money in HPQ.

	CSCO	MSFT	HPQ	Portfolio
Mean	1.81%	1.37%	0.55%	1.53%
Median	1.69%	0.76%	1.01%	1.59%
Std deviation	5.54%	5.41%	5.84%	4.63%
Minimum	–16.1%	–9.8%	–11.4%	–12.0%
Maximum	15.4%	22.8%	22.1%	11.4%
First quartile	–1.59%	–2.21%	–3.03%	–1.17%
Third quartile	5.65%	5.24%	3.95%	4.73%
IQR	7.24%	7.44%	6.98%	5.56%

(a) Comment on the figures for SD and IQR. What do these numbers indicate?

(b) Calculate the CV of the return on each investment. In simple words, what do these numbers indicate about the volatility of stock returns?

(c) Why might you want your investment to have a low CV? Which has the lowest CV?

1.6. Changing Units

The mean, median, and mode all measure typical or middle values of a distribution. They measure this in different ways but they are all so-called measures of *location*. Measures of location satisfy the following simple rules:

- If you add a constant K to all the data values then the location measure increases by K.
- If you multiply all the data values by a constant K then the location measure is multiplied by the same factor K.

The SD, IQR, and MAD all measure typical variability. They measure this in different ways but they are all so-called measures of *scale*.

- If you add a constant K to all the data values then the scale measure stays the same. The variable is exactly as variable as it was before.
- If you multiply all the data values by a constant K then the scale measure is multiplied by the same factor K.

The relevance of this to business analysis is illustrated by looking again at the data in ex2 (HKRE).xls, which describes 108 apartments sold in Hong Kong. If you look at the worksheet *final sale price* you will see the prices in the left column, which are expressed in Australian dollars, calculated at the December 1995 exchange rate. The mean is \$979.9 and the SD is \$264.3.

However, the estate agent charged a fixed fee of \$22K per sale. Column C gives the sale price net of these charges. The mean of these numbers is \$22K less than \$979.9, which is \$957.9. The SD is unchanged at \$264.3. The government charged a 12% value-added tax (VAT) on the sale price net of estate agent costs. The numbers in column D are equal to the numbers in column C times 0.88 and represent the money left after VAT is paid. The mean and SD are both multiplied by the factor 0.88. The mean becomes \$843.0 and the SD becomes \$232.6.

Finally, let us convert the figures back into Hong Kong dollars since that is where the sales took place. The exchange rate at the end of 1996 was \$5.16 Hong Kong dollars to the Australian dollar. Column E gives the received payment after commission and taxes in Hong Kong dollars, obtained from column D by multiplying by 5.16. The mean and SD are both multiplied by 5.16. The mean payment is HK\$4.35 million and the SD is HK\$1.2 million.

Price	Less k\$22 Commission	Less 12% VAT	in \$HK @ \$5.16
979.9	957.9	843.0	4349.8
264.3	264.3	232.6	1200.1
27.0%	27.6%	27.6%	27.6%

The CVs are also given in each case. Remember that this is the SD as a proportion of the mean, expressed as a percentage. Notice that when the data are multiplied by 88% and then multiplied by \$5.16, the CV does not change. This is because the mean and SD both change by the same factor, so the ratio is unaffected. When you add a number to the data, as when we subtracted the \$22K commission, the CV does change because the mean changes, while the SD does not.

The main upshot of all this is that when you subtract taxes from business outcomes or convert numbers into different currencies, the mean and SD also change in the same way—and the CV stays the same.

CHECK YOUR UNDERSTANDING

Q3 In 2007 there were 48 middle managers in a firm. Their mean salary was $100K pa. The SD of their base salary is $19.7K.

(a) In 2008, they received a salary increment of $10K each. What was the mean and SD of their salaries?

(b) In 2009, they received a salary increment of 10% each. What was the mean and SD of their salaries?

(c) In 2010, they received a variable percent salary increment, which averaged 10% but which depended on their base salary. What can you say about the mean and SD of their salaries?

(d) How is the CV affected in each case—higher, lower, the same, or not determined?

	2008	2009	2010
Mean			
Stdev			
CV			

Pushing a Bit Further

When you do something more complicated than just change units, it is not so clear what will happen to the mean and SD. For instance, suppose I have a portfolio of stocks and all of them rise. I check the percent growth of each of my stocks and find that the average growth was 10%. You would think then that the total value of my portfolio would increase by 10%. But it may not!

Open the spreadsheet ex4 (investing).xls in the Data and Examples folder. You will see 10 hypothetical stocks with different amounts invested. The total value of my portfolio was initially $1900, heavily weighted on stock 1. All stocks grew but, luckily for me, stock 1 grew the most. However, the average growth of the 10 stocks is still 10%. The final value of my portfolio is $2180, fully 14.7% higher than I started! This cannot happen if you have equal investment in each stock. But it is usually not a good idea to have equal exposure to all stocks, as we will see in a future topic.

You might like to play around with this template. What happens if I put more money in stock 10 and less money in stock 1? The moral of this story is to be careful about simplistic statements about means (and also about SD). For instance, if I give an average 10% bonus to my entire workforce, it does *not* follow that my wages bill will go up by 10%!!

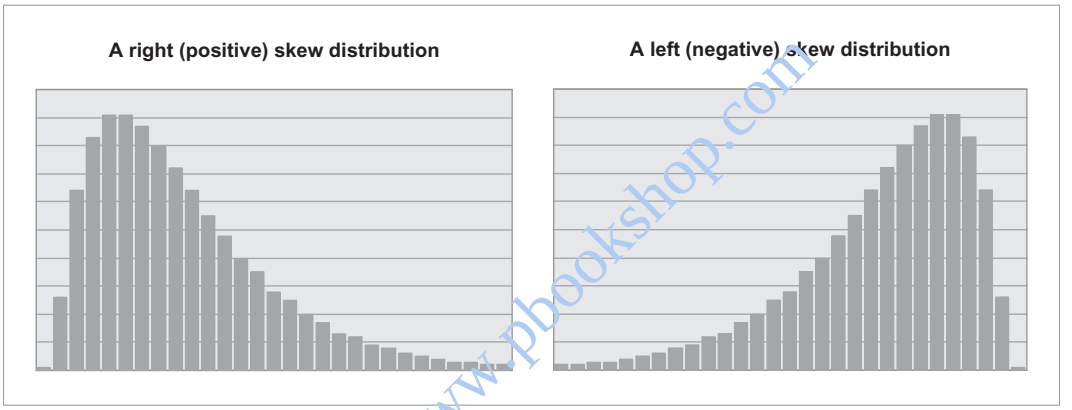
1.7. SHAPES OF DISTRIBUTIONS

Gross summary measures of location (such as mean or median) and scale (such as SD or IQR) tell us quite a lot about the data distribution. But there may be other fea-

tures of interest that are not revealed by these numbers. Thinking visually, the center and spread of a data distribution does not tell you everything about the *shape* of the distribution. Here are three issues to keep in mind when you are trying to understand, describe, and summarize a distribution.

Asymmetry and Skewness

Some data distributions are such that the largest observed values are way larger than the mean, whereas the smallest values are not all that small. Demand and price data are often like this, partly because the lower end of the distribution is constrained to be positive. Such asymmetry may be of some practical importance or interest. The SD would not tell us that variations on the upside are much larger than variations on the downside.



A generic chart displaying this kind of behavior is above left. Such a distribution is called *right or positively skew*. You will often know that a distribution is right skew if the mean is much larger than the median, though a better way to detect skewness is to look at the data distribution.

It is possible for a distribution to be left skew—where variation on the downside is larger than variation on the upside. This is common for human performance data where most people are performing close to some natural limit. The long left tail might be inexperienced employees who have not yet acquired the skills to perform at the highest levels.

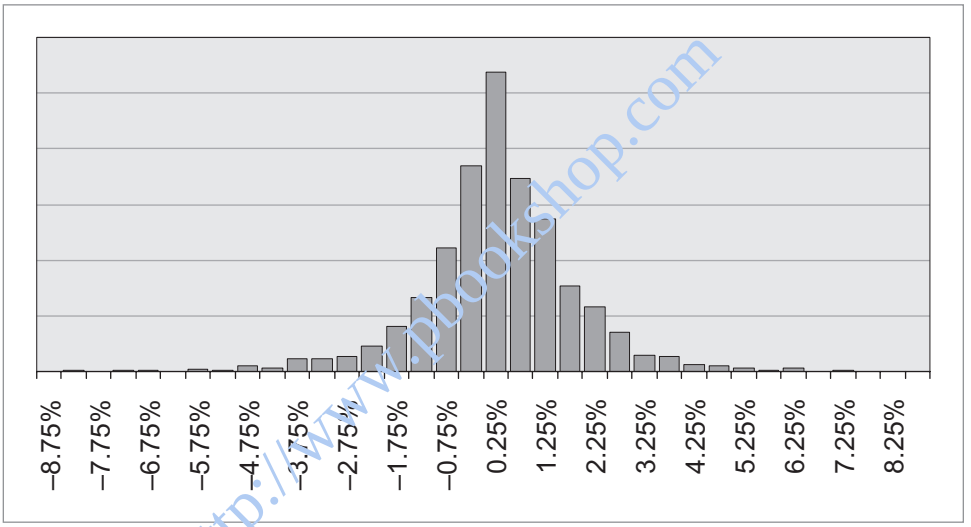
Skewness is worth keeping in mind when thinking about SD. It was earlier pointed out that roughly 95% of the data will be within two SDs of the mean. However, if the data is skewed then you would not find the extreme 5% evenly distributed at the top and bottom. With right skewed data you might find 4% in the upper tail and 1% in the lower tail.

There is a numeric measure of **skewness** but it is not easy to interpret and we do not cover it in this book. But the idea of skewness or asymmetry is worth keeping in the back of your mind when it comes to describing the more subtle aspects of the distribution. The business issue here is whether overperformance is more or less common than underperformance.

Very Heavy Tails

SD or IQR describe typical variation about the middle. These measures are not good at describing *extreme* variation. The *return* on a stock is its proportional change over a time period. Stock returns are typically subject to extreme outcomes—some trading periods see large proportional increases or decreases in value that are much larger than the typical changes.

The chart below is for weekly returns on the Hong Kong stock market index known as the Hang Seng Index. Notice the occurrence of some very high and also some very low returns. The SD of this data set is 1.7%, which does not tell us much about the extreme returns. There is a numeric measure of *heavy tails* called **kurtosis** but we do not cover it in this book. However, it is worth being aware of this further aspect of a variable's distribution.

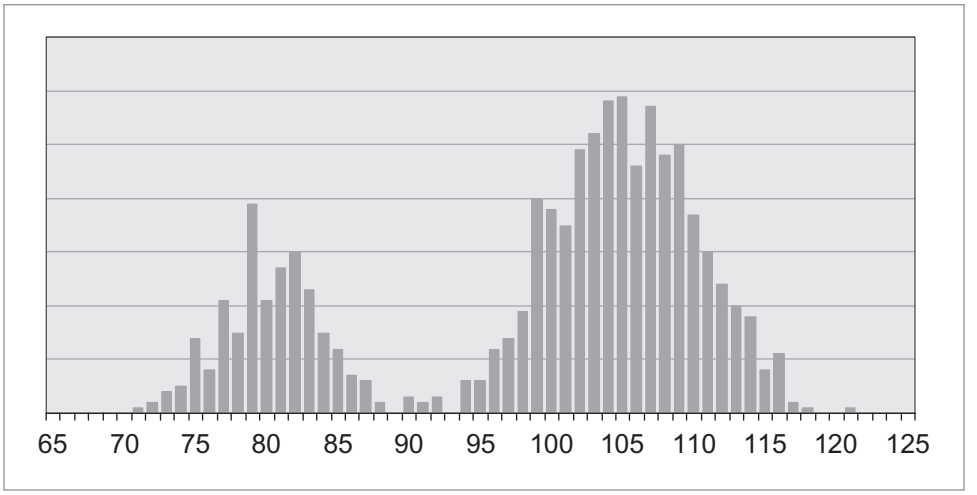


The business issue is whether or not extreme performance ever happens—where by extreme we mean more than 3 or 4 SDs from the mean. This is obviously always going to be rare, but for most regular performance outcomes it virtually never happens. For heavy-tailed distribution, however—such as occur in finance—you do find the occasional but truly exceptional (either good or bad) outcome.

Multimodality/Segmentation

A data distribution like the one below might describe the monthly mobile phone bills across a sample of customers. It is quite obvious that there are two distinct groups of customers—maybe on different call plans, or maybe the right-hand group consists of business customers and the left-hand group consists of domestic customers.

Measures of location and scale cannot alert you to this fact. Indeed, it would be rather meaningless here to quote the mean value of 98.4 and the SD of 11.8. When you have a data distribution like this you will not usefully summarize it with these two numbers. You would be much better off trying to discover what label defines the two market segments.



1.8. SUMMARY

Data come in various flavors—categorical, ordinal, and numeric/interval. You should also classify business variable into performance measures (the outcomes you really care about), input variables (the ones you think may affect the performance outcome), and segmenting variables (corresponding to different market segments).

The **sample mean** measures the center of a distribution. In practical terms this means the central or typical value of the variable. The **sample SD** measures the spread of a distribution. In practical terms this means the uncertainty or variability of the variable. There are alternative measures to mean and SD but these are less commonly used.

When you change currencies, mean and SD change in the same way. When you add or subtract a constant to a data set, the mean changes but the SD does not. The **CV** measures variability in proportional terms. It is the SD divided by the mean, expressed as a percentage.

The data distribution can be largely summarized by these three numbers. But there are more subtle aspects of the data distribution that the mean and SD suppress, such as asymmetry, heavy tails, and segmenting.

EXERCISES

E1. You are given the following 15 data values: 0, 3, 1, 0, 1, 1, 1, 3, 4, 3, 2, 0, 2, 0, and 0. Obtain the relative frequency distribution, the mean, and the SD. First use ordinary Excel and COUNTIF to calculate the frequencies and relative frequencies. Then use the Excel template data_distribution.xls to confirm your answers.

E2. Consider the frequency distribution below. What proportion of the data is between 2 and 4 inclusive? What proportion of the data values is odd? What is the mean and SD? Use the Excel template frequency_distribution.xls.

Value	1	2	3	4	5	6	7	8	9	10
Frequency	49	144	244	197	159	97	51	15	6	3

E3. As explained in the class notes, the symbol Σ means "sum" or "add." Using the samples below, calculate Σy_i, Σx_i^2, and $\Sigma\, x_i^2 y_i$. The point of this question is to make sure you understand this notation.

X-sample	1	2	3	4	5	6	7	8	9	10
Y-sample	9	14	24	17	59	7	5	5	2	1

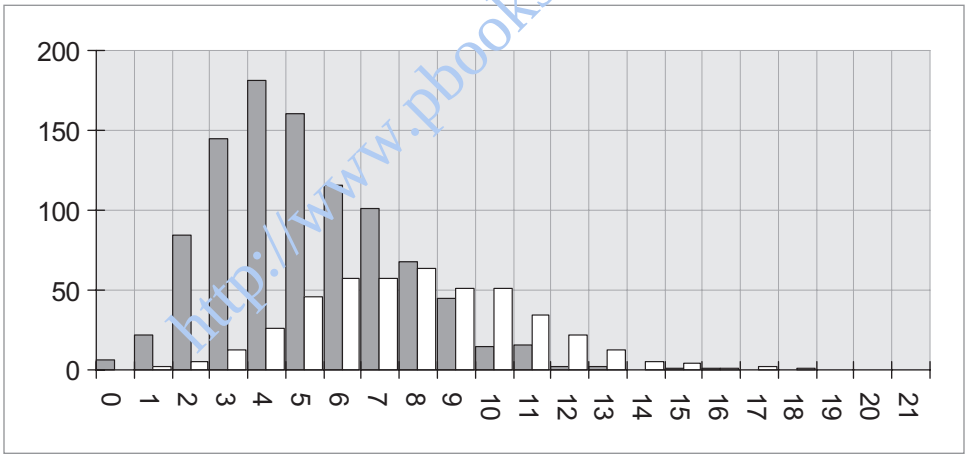
A quick way to do these calculations is to name the data ranges; for instance, call the x sample "x" and the y sample "y." Then you can calculate $\Sigma\, x_i^2 y_i$ by typing = sumproduct(x,x,y).

E4. Suppose that you have 1 year of data on weekly sales at a supermarket in units of $1000. The mean is 27.4 and the SD is 2.5. Describe the typical weekly sales *in* words that could be understood by a nonstatistician. What would happen to these figures if you subtracted the weekly costs of $12,400 from the 52 weekly sales figures?

E5. Below are the lengths of five waiting times in a queue measured in minutes. Calculate the mean and SD. Then convert the five times into seconds and calculate the mean and SD again. What happens to the mean, SD, and CV?

Data in minutes: 21.8, 26.1, 31.4, 15.7, and 20.0.

E6. In the histograms below the gray bars represent sample A and the white bars represent sample B. Estimate by eye the mean and SD of each sample. Why are the gray bars so much taller?



E7. For 1241 apartments sold in a certain area in 2002, the mean price is $459,000 and the SD is $76,000. Consider three apartments that sold for $545,000, $295,000, and $395,000. Which of these is most unusual? Are any of them really unusual statistically?

E8. A sample of 450 consumers was surveyed on their typical weekly consumption coffee in units of *cups*. They were asked how many cups of coffee they purchased over a typical week. There were 296 men in the survey and 154 women. For men, the mean consumption was 19 cups with an SD of 21.0. The median consumption was 13. For women, the mean consumption was 16.6 cups with an SD of 19.7. The median expenditure was 13. There were 73 male consumers and 65 female consumers who did not drink coffee at all.

Describe the distribution for males and females in ordinary words. How much difference is there between males and females?