



Richard N. Landers



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/1 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Kirsty Smy
Editorial assistant: Martha Cunneen
Assistant editor, digital: Chloe Statham
Production editor: Sarah Cooke
Copyeditor: Andy Baxter
Proofreader: Tom Hickman
Indexer: Martin Hargreaves
Marketing manager: Alison Borg
Cover design: Francis Kenney
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

© Richard N. Landers 2019

First published 2013. Reprinted 2017 (three times)
This second edition published 2019

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

SPSS Reprints Courtesy of International Business Machines Corporation, © International Business Machines Corporation (SPSS Inc. was acquired by IBM in October, 2009).

IBM, the IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "IBM Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Library of Congress Control Number: 2018954720

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-4739-4810-5
ISBN 978-1-4739-4811-2 (pbk)

At SAGE we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

CONTENTS

<i>Dedication</i>	vii
<i>Acknowledgements</i>	viii
<i>About the Author</i>	ix
<i>Your Guide to This Book</i>	x
<i>Online Resources</i>	xii
<i>Preface</i>	xiii
<i>Changes in the Second Edition</i>	xv
<i>Introduction</i>	xvii
PROLOGUE: REVIEW OF ESSENTIAL MATHEMATICS	XIX
PART 1 – DESCRIPTIVE STATISTICS	1
1 The Language of Statistics	3
2 Working with Numbers and Data Display	24
3 Central Tendency and Variability	75
4 Probability Distributions	109
5 Sampling Distributions	141
PART 2 – INFERENCE STATISTICS	163
6 Estimation and Confidence Intervals	165
7 Hypothesis Testing	192
8 z -Tests and One-Sample t -Tests	214
9 Paired- and Independent-Samples t -Tests	251
10 Analysis of Variance (ANOVA)	297
11 Chi-Squared (χ^2) Tests of Fit	335
12 Correlation and Regression	369

PART 3 – WRAP-UP	397
13 Matching Statistical Tests to Business Problems	399
<i>Bibliography</i>	423
<i>Appendices</i>	424
A1 <i>z</i> -Table	424
A2 <i>t</i> -Table	426
A3 <i>F</i> -Table ($\alpha = .05$)	428
A4 <i>F</i> -Table ($\alpha = .01$)	429
A5 χ^2 -Table (chi-squared)	430
B1 Inferential Test Reference Table	431
B2 Inferential Test Decision Tree	432
C Glossary	433
D Statistical Notation and Formulas	444
E Enabling Excel's Data Analysis Add-In	454
<i>Index</i>	457

4 PROBABILITY DISTRIBUTIONS

WHAT YOU WILL LEARN FROM THIS CHAPTER

- How to work with and interpret probabilities
- How to distinguish between normal, uniform and Poisson distributions
- How to predict the probability of cases based upon the normal distribution
- How to round appropriately so as to minimize errors
- How to convert between raw data, z-scores and proportions
- How to identify when it is appropriate to convert between raw data, z-scores and proportions

DATA SKILLS YOU WILL MASTER FROM THIS CHAPTER

- Computing z-scores
- Computing the mean of variables
- Computing percentiles

CASE STUDY WHO NEEDS EXTRA TRAINING?



Jill is the owner of Jill's Used Cars, a used vehicle business with a staff of 55 salespeople. With her new-found knowledge from previous chapters of this text, Jill decided it would be useful to analyse sales to see how well her employees were performing relative to each other. Because she wanted to examine employees' previous patterns of sales, and based upon reading Chapter 1, she decided to use a correlational design. Because part of her organization's mission

(Continued)

(Continued)


was to sell ‘the best car for the customer’, she decided to operationalize monthly sales as number of vehicles sold instead of the value of those vehicles.

This didn’t require much additional effort on Jill’s part, because she’d already been recording monthly sales for each employee and writing them on a whiteboard in the break room. This, she hoped, would motivate her salespeople to sell more. If the poorer-performing sellers could see how much the top sellers were bringing in, and they knew how much commission those sellers were earning for their increased sales, they might be motivated to sell a bit more. But in watching sales from month to month, Jill has noticed that her salespeople tend to stay around the same performance level relative to their peers.

Most of her salespeople have been working for her for three or more years. Jill believes the interrelationships between her team members are important, so she wants to keep as many

employees as she can. She doesn’t want to let anyone go unless there’s no other option. But some of her employees definitely need help to perform at a satisfactory level.

Last week, Jill received an e-mail from an online training provider offering an intensive online sales seminar. It looks very promising, but it’s expensive. Jill only wants to provide training to those who will benefit most from it. From reading Chapter 3, she has identified the mean and variability of the monthly sales numbers for the past six months. There is quite a bit of variability, as she suspected, and sales month-to-month are often quite different. But that doesn’t tell her anything about individual employees. Just how different are each of their sales numbers? What’s the best way to compare employees with each other across months?

 Take a look at Jill’s employee sales data for yourself in [chapter4.xls](#) (Excel) or [chapter4.sav](#) (SPSS).

Jill’s problem is a common one. Although employee success varies, it’s not immediately clear just how meaningful that variance is when considering the success of individuals. As we learned in Chapter 3, measures of central tendency and variability are excellent ways to precisely consider the quantitative characteristics of a group in aggregate. But those statistics alone don’t tell us much if we need to make judgements about individuals within those groups.

In this case study, to compare employees, Jill’s only option so far is to compare the raw numbers of sales between employees. But if one employee sells 10 vehicles and another sells 11 vehicles, is the one that sold 11 vehicles really a ‘better’ seller? Or is it, perhaps, that the seller was just lucky? In statistics, we refer to luck as **chance**, and one of the major goals of statistics is to explore and explain the effects of chance.

We began our exploration of chance in Chapter 3 by learning about the **normal distribution**. But why do data tend to be normally distributed? Are there other common distributions of data? In this chapter, we will begin our exploration of chance by considering several common shapes of data and why we tend to see these shapes. As you read, consider how knowledge of these distributions can help Jill understand the relative performance of her employees.

4.1 PROBABILITY



To identify why data typically appear in the various shapes they appear in, we first need to explore the concept of **probability**. Probability, broadly, refers to how likely a specific event is to occur. It is typically expressed as a proportion (see Chapter 2, p. 27) and ranges from .00 (a specific event will never occur) to 1.00 (a specific event will always occur). One of the major purposes of statistics is to accurately assess probabilities associated with data.

In some cases, probability is very easy to compute. For example, consider a coin – one side is heads; the other is tails. Any time you toss the coin in the air and let it fall, it will land either heads-up or tails-up. If there's nothing strange about the coin, it will fall heads-up about half of the time and tails-up the other half of the time. Thus, the probability of heads is .50. The probability of tails is also .50.

As we add more possible outcomes, the probabilities become more complex. Next consider a five-sided die with the numbers 0, 1, 2, 3 and 4 on its sides. The probability that any particular number will land facing up when rolling that die is $1/5$, which we can express as a probability as .20. Thus, the probability of a 0 is .20, 1 is .20, 2 is .20, 3 is .20 and 4 is .20. All together, these numbers add up to 1.00, since with any given roll of the die, one of these numbers will appear 100% of the time.

If we didn't know ahead of time the probabilities associated with this, we could use a process called the **classical method of assigning probability**. The classical method involves simply doing the thing you're curious about many, many times, and then calculating how many times each outcome occurred relative to the total. For example, let's roll our five-sided die 10000 times. In doing so, you might end up with 1942 rolls of 1, 2029 rolls of 2, 2013 rolls of 3, 2038 rolls of 4, and 1978 rolls of 5. By calculating relative frequencies (for example, $1942/10000 = .1942$), you can see that each is very close to .2. If I were really going to assign a probability to each size using this method, I'd want to collect many more than 10000 cases to make those numbers more stable and closer to .2.

An alternative way to assign probabilities is the **relative frequency of occurrence method of assigning probability** in which historical data are consulted. For example, if we want to know the probability that the person we just hired will quit within six months, we could consult our company's human resources records to determine the relative frequency of newly hired employees that quit within six months. If we found that 56 out of the 178 people we've ever hired had quit in the last six months, we could conclude that the probability of the new hire quitting is also $56/178 = .31$, or a 31% chance.

In organizational research, probability is much more complicated, because we typically don't know the total number of possibilities. In our case study, how likely is it that an

employee will sell ten cars in any particular month? What about six or 30? Should we consider 100? We could use the relatively frequency of occurrence method, but how do we know if last month's sales or this month last year's sales is a better choice? These are the problems that keep statisticians up at night. Without a list of every possibility, we cannot compute a specific, precise probability that any of these events will occur.

Fortunately, data typically take one of several common shapes, and we can compute the probability of data occurring within any of these shapes. The next sections will explore what these shapes look like and the relative probabilities of the data they contain.

COMBINING PROBABILITIES

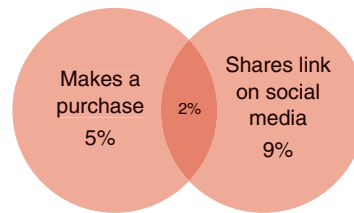


FIGURE 4.01 Unions and intersections

Sometimes you need to think and talk about probabilities in relation to one another. In these cases, probability often does not work the way that people intuitively think it should. In Figure 4.01, I've depicted three relative frequencies reported by a marketing department related to the performance of a product on their sales website. 5% of customers viewing the product purchase it. 9% of customers viewing it share a link to it on social media. 2% of customers do both. Implied by this diagram then is that $100\% - 5\% - 9\% - 2\% = 84\%$ of customers viewing this product neither purchase it nor share a link. We can utilize the relative frequency of occurrence method to convert these numbers into probabilities; for example, the probability of a customer viewing this page making a purchase is .05.

So what is the probability that a customer will make a purchase *or* share a link? Combining probabilities this way is called **union** and is signified by the word 'or'. To calculate this probability, we need to *add together* all of the component probabilities. In this case, $5\% + 2\% + 9\% = 16\%$. Thus, the probability of a customer seeing this webpage making a purchase or sharing a link is .16.

But what if we want to know the probability a customer will make a purchase *and* share a link? Combining probabilities is called **intersection** and is signified by the word 'and'. To calculate this probability, we need to simply look at the intersection point alone: 2%. Thus, the probability of a customer seeing this webpage making a purchase and sharing a link is .02.

If making a purchase and sharing a link were unrelated events, which is to say if we assumed the occurrence of one does not affect the other and that nothing externally affects them together, we'd expect the probability of intersection to be the value of the two component probabilities multiplied together. In this case, we'd expect a probability of $.05 * .09 = .0045$ for people to both make a purchase and share a link. Because the observed value, .02, is much higher than that, we could conclude that they probably are related. This is the basic logic of hypothesis testing, which we'll explore in much greater detail starting in Chapter 7.

4.2 DISTRIBUTIONS OF DATA

Data take different shapes, called **distributions**. There are several common distributions, three of which we'll cover in this chapter. It's important to realize that distributions are prototypes; 'real' data never conform perfectly to any prototype. You've already seen several examples of this in the previous chapter; although the normal distribution is a nice, predictable bell-shaped curve, the data you examined in the examples and case study created only a vaguely bell-shaped histogram.

Most statistical tests (especially the ones we cover in this book) are fairly robust (see Chapter 3, p. 80) to small variations from prototype distributions. For example, in the last chapter we learned that the mode and median reflect central tendency even if the data are skewed. So don't worry too much that the data you collect to answer your organization's research questions don't perfectly represent the prototypes. You just need to know which distribution to expect given your data and take a look at a histogram to verify that it is roughly the shape you're expecting.

4.2.1 UNIFORM DISTRIBUTION

The simplest data distribution is the **uniform distribution**. In a uniform distribution, every possibility is equally probable. We don't see this distribution very much among organizational data, but it is easy to interpret, so it's a good place to start.

Let's return briefly to die-rolling, this time with a ten-sided die with sides numbered from 1 to 10. If we were to roll our ten-sided die 50 times, we'd expect each number to appear five times. This is because each side is equally probable: each roll, the probability of any particular side appearing on the top of the die is $.10$. A bar graph of our expectations appears in Figure 4.02.

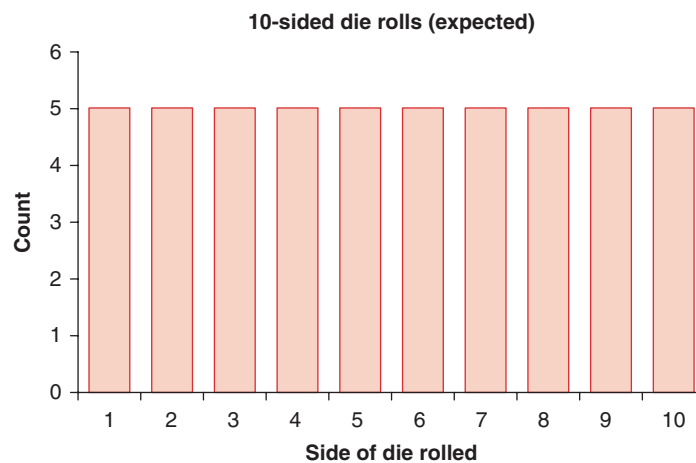


FIGURE 4.02 Expected sample of $n = 50$ die rolls drawn from a uniform distribution

If we were to actually roll a ten-sided die 50 times, it's very unlikely we'd see such a perfect distribution. However, you can easily see why the distribution is called 'uniform' – all options are equally probable. With real data, we'd be more likely to see something like the two samples in Figure 4.03.

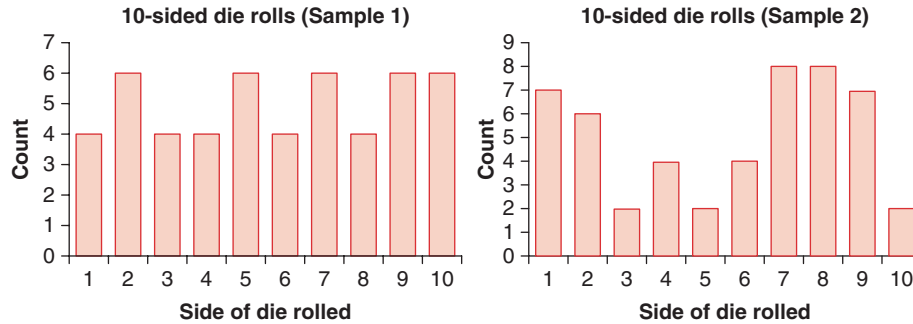


FIGURE 4.03 Examples of observed samples ($n = 50$) drawn from a uniform distribution

Both of these figures show data from 50 rolls of a 10-sided die. Yet, they look dramatically different from our expected distribution. This is because of chance. Although we expect a perfectly uniform distribution, real data are typically messy, as we see here, and luck alone can lead to substantial variation from the expected distribution. The more that data deviate from an expected distribution, the more sceptical we must be when interpreting statistics that expect that distribution. Fortunately, there's a fairly easy way to get closer-to-expected distributions of real data: large samples. We'll discuss this principle in much more detail in the next chapter.

4.2.2 POISSON DISTRIBUTION

The **Poisson** (pronounced: pwah-SAHN) **distribution** describes data containing small numbers of independent counts. It typically occurs when there are a small number of observations for a relatively large number of opportunities for those observations to occur. Like the normal distribution, it is a spread of random data.

Consider an example of how a Poisson distribution might appear in an organizational context. In a call centre, employees are rarely asked by the people they are calling to escalate the call to a supervisor. Every customer has the power to ask this, but relatively few actually do. On average, the call centre handles two elevated calls per day. This creates the Poisson distribution in Figure 4.04.

Thus, on roughly 13.5% of days, the call centre has no escalated calls. On 27% of days, the call centre has one escalated call. On 27% of days, the call centre has two. On 18% of days, the call centre has three. On 9% of days, the call centre has four (and so on).

This distribution thus appears similar to the skewed normal distribution we discussed in Chapter 3, with one key difference: it only applies to discrete, nominal data (counts; see Chapter 1, p. 8).

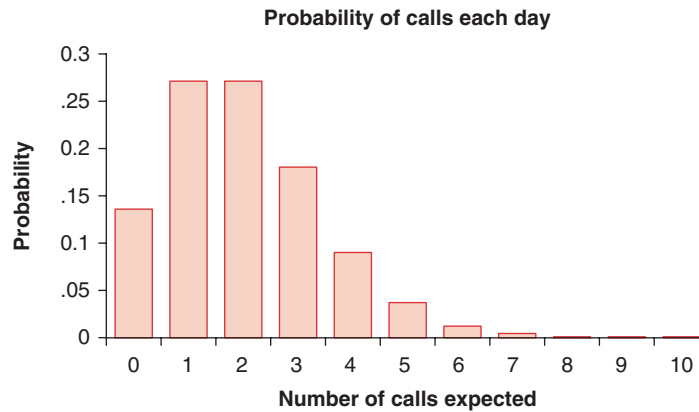


FIGURE 4.04 Example of a Poisson distribution from a call centre

4.2.3 NORMAL (AND STANDARD NORMAL) DISTRIBUTIONS

FOUNDATION CONCEPTS

A **normal distribution** is a common shape in which data are found resembling a bell or hill. See Chapter 3, p. 77.

The **standard deviation** is the average distance between all scores in a variable and their mean, represented in a population as σ (sigma) and in a sample as s , and is also the square root of the variance. See Chapter 3, p. 90.

Normal distributions are the most common shape of data you are likely to see, especially if you ask many organizational questions with surveys. We learned a bit about these distributions in Chapter 3. When data are normally distributed, cases are more likely to occur close to the mean and increasingly unlikely to appear the further we move away from the mean in either direction. This is what creates the bell shape.

We sometimes refer to the **standard normal distribution**, which is a perfectly normal distribution with a mean of zero and a standard deviation of one ($\mu = 0$; $\sigma = 1$). This is also called a **z-distribution**. You can see a z-distribution in Figure 4.05.

The standard normal distribution is useful because it is defined in terms of standard deviations. When looking at a standard normal distribution, -2 is two standard deviations below the mean, -1 is one standard deviation below the mean, 0 is the mean, $+1$ is one standard deviation above the mean, and $+2$ is two standard deviations above the mean. These numbers are called **z-scores**. They don't need to be whole numbers; for example, $z = -1.29$ describes a score 1.29 standard deviations below the mean. A z-score can be calculated from the scores in any dataset.

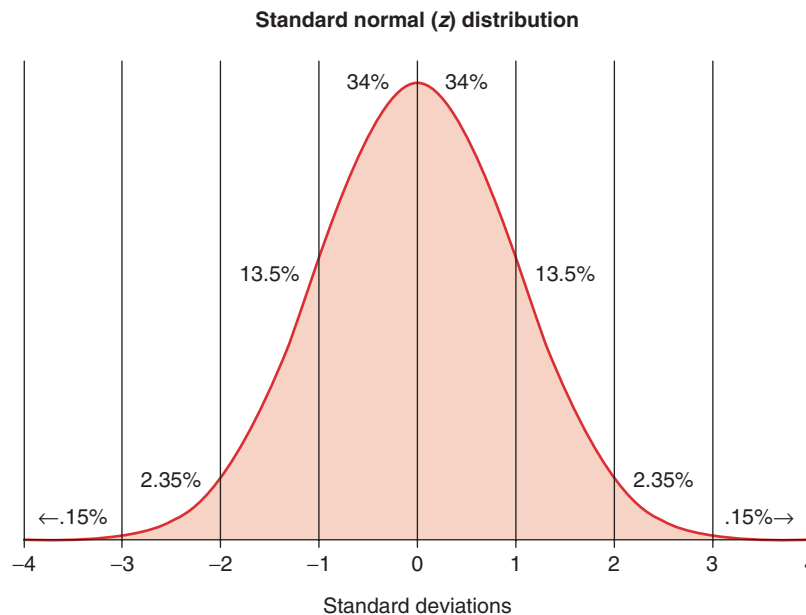


FIGURE 4.05 Normal (z) distribution and percentage of cases under the curve divided between standard deviations

WHY MIGHT DATA THAT WE EXPECT TO BE NORMAL BE SKEWED?

When there is some restriction that prevents normally distributed data from moving away from the mean in both directions, it bunches up wherever it is blocked, creating skew. This creates a problem in organizational research because skew distorts the results of many common statistical tests.

For example, the collection of supervisory ratings of job performance is often tricky for organizational researchers because many supervisors rate their direct reports (the employees that report to them) very highly. There are many reasons for this. Some supervisors don't want to expose their direct reports to the risk of layoffs. Others have personal relationships with their direct reports and don't want to rate them poorly regardless of their actual performance. Still other supervisors have a difficult time assessing the job performance of others so, to be safe, they rate every direct report identically.

Regardless of the reason, when responding to a five-point survey question about employee job performance where five is the strongest response, the mode will often be a five. Although we'd expect these data to be normally distributed, because there are no options higher than five, the data will be negatively skewed.

When analysing organizational data, z -scores are helpful because they can be **standardized** from **raw data**. Raw data are data as they are collected and thus cannot be compared across dissimilar variables. For example, it is impossible to directly compare 18 sales with a rating of '4' from a supervisor. However, each of these values can be standardized so that they are comparable. For example, 18 sales may be $z = 1.1$ (1.1 standard deviations above the mean) while '4' is $z = -.2$ (.2 standard deviations below the mean). Therefore, this employee is above average in sales but slightly below average in supervisory ratings.

By standardizing, we can compare these numbers directly with one another (below average in one regard; above average in another regard).

In our case study, standardization is critical because sales fluctuate radically by month, regardless of employee performance. Consider an employee whose raw sales in July and December are 10 and 10. Doesn't look bad, right? But if we convert these values to z -scores, we find that this employee's July sales were $z = -.10$, whereas his December sales were $z = -1.15$. From these standardized scores, we can conclude that this employee was slightly below average in July and very below average in December. This information was not obvious from the raw data alone. Car sales are simply higher in some months than others. By converting monthly sales to z -scores, we can therefore more accurately assess how well employees are doing relative to their co-workers.

4.3 PROPORTIONS OF CASES WITHIN NORMALLY DISTRIBUTED DATA

FOUNDATION CONCEPTS

A **proportion** is a portion of a whole represented as a decimal. For example, .35 is a proportion indicating 35 for every 100. See Chapter 2, p. 27.

A **percentage** is a portion of a whole represented as its share of 100. For example, 35% is a percentage indicating 35 for every 100. See Chapter 2, p. 28.

One of the advantages to collecting and analysing data that conform to a normal distribution is that the normal distribution is highly predictable. If we collected a large amount of normally distributed data, we would expect that:

- .15% (.0015) of cases fall below -3 standard deviations.
- 2.35% (.0235) of cases fall between -3 and -2 standard deviations.
- 13.5% (.1350) of cases fall between -2 and -1 standard deviations.
- 34.0% (.3400) of cases fall between -1 standard deviation and the mean (0).
- 34.0% (.3400) of cases fall between the mean (0) and $+1$ standard deviation.
- 13.5% (.1350) of cases fall between $+1$ and $+2$ standard deviations.
- 2.35% (.0235) of cases fall between $+2$ and $+3$ standard deviations.
- .15% (.0015) of cases fall above $+3$ standard deviations.

We also have some broader expectations:

- 50% (.50) of cases fall above the mean (0).
- 50% (.50) of cases fall below the mean (0).

You can actually *see* these percentages in Figure 4.05 above. Since the distribution represents 100% of the data, we can cut it up and look at the **area under the curve** to identify what percentage of cases fall between any two particular points of interest.

For example, if an employee's job performance was $z = +1$, that employee's performance would be higher than roughly 84% of her fellow employees. I got this number by adding the percentage of cases falling below the mean (50%) and the percentage of cases between 0 and +1 standard deviations (34%). That employee's performance would also be below approximately 16% of employees (13.5% + 2.35% + 0.15%).

By converting between raw scores, z -scores and the proportion of cases associated with particular z -scores, we can make many meaningful conclusions about the frequency of values relative to their populations. Exploring this concept mathematically will make up the remainder of this chapter.

4.3.1 CONVERTING FROM RAW DATA TO Z-SCORES

ACCOUNTING FOR ROUNDING ERROR

If you are using this textbook as part of a course, and you haven't talked about it already, this is an excellent point at which to talk to your instructor about measurement precision. Calculation in statistics often involves working with very small numbers, and the computations you'll be doing with z -scores are no exception. Every time you round a value, you lose a little bit of precision, which can lead to your final answer being incorrect.

To account for this, this text recommends carrying out all computations to six decimal places mid-computation, only then rounding the final answer to two decimal places. All computations displayed in this text from this point forward will adopt this strategy. However, if you are using this text for a course, your instructor may want your final answers to be more or less precise; if you don't know your instructor's expectation on this yet, you should ask as soon as possible.

To convert a raw score to a z -score, you must know the mean and standard deviation of the sample or population you want to compare it with. You might calculate this yourself from a dataset, or you might be given this information.

If you were analysing data from within your own organization, you would probably have access to the raw data and would therefore be able to compute the mean and standard deviation for yourself. But there are times when you might not have access to this information, or you might want to compare it with population data. Whichever mean and standard deviation you choose, this will be the comparison group for your z -scores.

For example, if you were worried about the attitudes toward customer service shown by some of your employees, you might hire a consultant to administer a customer service survey to them to assess this. After getting the results, you don't necessarily want to know how your employees compare with each other – instead, you want to know how they compare

with customer service employees *in general*. In this case, you might ask the consultant to provide the mean and standard deviation of their survey across all employees they've ever assessed. Any z -scores computed with this mean and standard deviation will be in reference to that group. A $z = -1$ in this context would indicate an employee with customer service attitudes one standard deviation below employees *in general*.

Regardless of the source of your comparison group, the formulas for converting a raw score to a z -score are similar whether you have samples or populations for comparison – see Figures 4.06 and 4.07.

$$\begin{array}{c}
 \text{Sample mean} \\
 \text{(see Ch. 3, p. 83)} \\
 \downarrow \\
 z = \frac{x - \bar{X}}{S} \\
 \uparrow \\
 \text{Sample standard deviation} \\
 \text{(see Ch. 3, p. 91)}
 \end{array}$$

FIGURE 4.06 Annotated z -score formula for a sample

$$\begin{array}{c}
 \text{Population mean} \\
 \text{(see Ch. 3, p. 88)} \\
 \downarrow \\
 z = \frac{x - \mu}{\sigma} \\
 \uparrow \\
 \text{Population standard deviation} \\
 \text{(see Ch. 3, p. 89)}
 \end{array}$$

FIGURE 4.07 Annotated z -score formula for a population

Take a moment to think about why the z -score formula is constructed this way. In each formula, we first subtract the sample or population mean from the value we're interested in. This results in a difference score representing the distance between the score and the mean. In the second step, we divide this difference by the standard deviation to convert the difference to standard deviation units.

As an example, consider the value of '15' sales from our case study dataset, coming from November sales. The mean of November sales is 12.327272, while the standard deviation of November sales is 4.830668. In the first step of the computation, we calculate $15 - 12.327272$, which gives us 2.672728. This number indicates the difference in sales between the case and the mean; in other words, 15 sales is 2.672728 sales higher than the mean. In the second step of the computation, we divide 2.672728 by the standard deviation, 4.830668, which gives us .553283. Therefore, the score is .55 standard deviations above the mean, and $z = .55$.

If we think about the relationship between these two numbers, this should seem obvious. 2.672728 is roughly half of 4.830668, so a z of .55 makes sense – the score (15) and the mean (12.327272) are roughly half of one standard deviation apart. We can see the steps of this computation in Figure 4.08, broken down step by step.

4.3.2 CONVERTING FROM Z-SCORES TO RAW DATA

To convert a raw score to a z -score, use the formula in Figure 4.08 in reverse. The formula solves for z , but we want to solve for x . Fortunately, as you can see from Figure 4.09, it's quite easy to convert.

$$\begin{aligned} z &= \frac{x - \bar{x}}{s} \\ &= \frac{15 - 12.327272}{4.830668} \\ &= \frac{2.672728}{4.830668} \\ &= .553283 = .55 \end{aligned}$$

FIGURE 4.08 Step by step calculation of a z -score from a sample

From this, you can see that converting a z -score back to a raw score requires you to know three pieces of information: the z -score, the standard deviation and the mean. This formula, like the one before, can be expressed in terms of either samples or populations (Figures 4.10 and 4.11).

Here's the starting formula:

$$z = \frac{x - \bar{x}}{s}$$

Next, we multiply both sides by s :

$$zs = x - \bar{x}$$

Finally, we add \bar{x} to both sides:

$$zs + \bar{x} = x$$

FIGURE 4.09 Algebraically modifying z -score formula to solve for a raw score

$$\begin{array}{c} \text{Sample mean} \\ \text{(see Ch. 3, p. 83)} \\ \downarrow \\ x = zs + \bar{x} \\ \uparrow \\ \text{Sample standard deviation} \\ \text{(see Ch. 3, p. 89)} \end{array}$$

FIGURE 4.10 Annotated formula to calculate a raw score from a sample z -score

$$x = z\sigma + \mu$$

Population mean
(see Ch. 3, p. 88)

↓

↑

Population standard deviation
(see Ch. 3, p. 89)

FIGURE 4.11 Annotated formula to calculate a raw score from a population z-score

In our case study dataset, imagine that Jill wanted to reward every employee selling more than +2 standard deviations of vehicles in November. What score would she look for? We'll use the same values from the example in the previous section: the mean of November sales is 12.327272, while the standard deviation of November sales is 4.830668 – see Figure 4.12.

$$\begin{aligned} x &= zs + \bar{x} \\ &= 2(4.830668) + 12.327272 \\ &= 9.661336 + 12.327272 \\ &= 21.988608 = 21.99 \end{aligned}$$

FIGURE 4.12 Step by step calculation of a raw score from a sample z-score

Thus, Jill would reward anyone selling 22 or more vehicles in November.

4.3.3 CONVERTING FROM Z-SCORES TO PROPORTIONS

Although conversions between raw scores and z-scores are interesting, the most useful information is when we then convert those z-scores to proportions. We've already done some simple conversions using the proportions associated with $z = -3, -2, -1, 0, 1, 2$ and 3 shown in the normal distribution in Figure 4.05 above. For example, 84% of cases fall above $z = -1$ (50% + 34%). But what if we want proportions for other z-scores? For that, we must reference the z-table, which appears in Appendix A1 (see p. 424).

In the z-table, you'll see columns and rows representing z-scores. To find the value in the table corresponding to any particular z-score, add the numbers you find at the top of the table with the numbers you find on the left side of the table. For example, if you want to find $z = 2.36$, look for the row labelled 2.3 and the column labelled .06. The value where these two numbers intersect is .0091 (Figure 4.13).

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014

FIGURE 4.13 Locating a z-score in the z-table (Appendix A1)

So what does this value mean? This is the proportion of cases falling *above* this positive z-score. Thus, .91% of cases fall above $z = 2.36$. To determine how many cases fall *below* this point, subtract it from 1. In this example, $1 - .0091 = .9909$. Therefore, 99.09% of cases fall below $z = 2.36$. You can see this graphically in Figure 4.14.

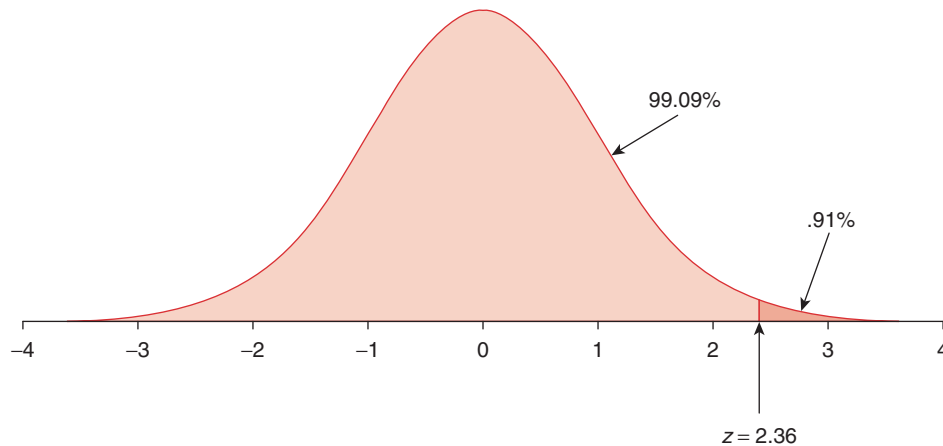


FIGURE 4.14 z-distribution and percentage of cases falling above $z = +2.36$, based upon the z-table

We sometimes refer to the smaller portion of this graph as the positive ‘tail’ of the distribution. So what do we do if our z is negative? The process is exactly the same – because the z -distribution is symmetrical, the proportions are the same on the negative tail. Simply look up the value associated with the positive z -score (Figure 4.15).

So given this, what’s the proportion of values *outside* $z \pm 2.36$? Just add them together: $.91\% + .91\% = 1.82\%$. We can then compute the remainder (the values *inside* $z \pm 2.36$) as $1 - .0182 = 98.18\%$ (Figure 4.16). The distribution always contains a total of 100% of all values. In this case, $.91\% + 98.18\% + .91\% = 100\%$.

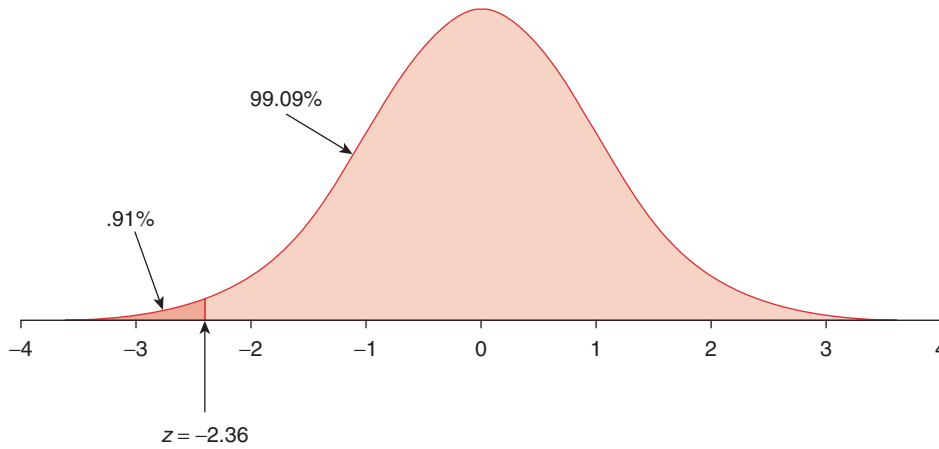


FIGURE 4.15 z-distribution and percentage of cases falling below $z = -2.36$, based upon the z-table

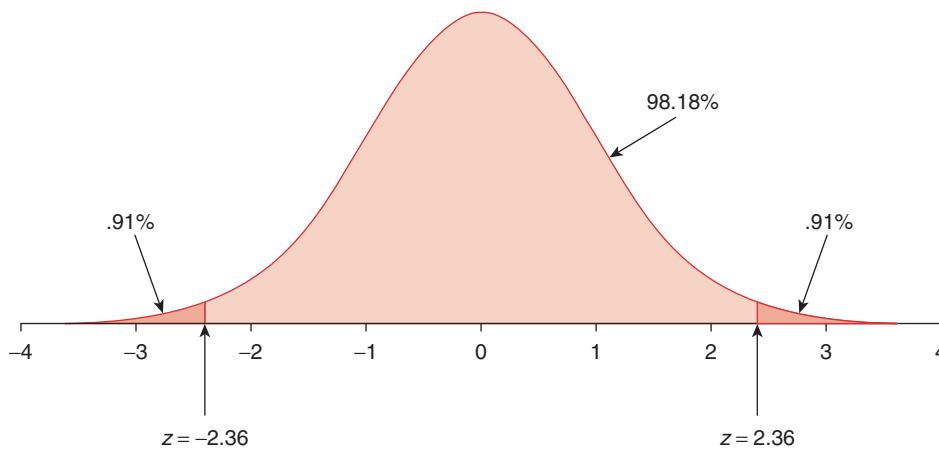


FIGURE 4.16 z-distribution and percentage of cases falling outside of $z = \pm 2.36$, based upon the z-table

WHY DOES THE Z-TABLE END AT 3.59?

Sometimes, when looking at the z-table, students conclude that z must always be between -3.59 and $+3.59$. This is not true! In fact, a z-score can be absolutely any value from negative infinity to positive infinity. Values below -3.59 or above 3.59 are simply extremely uncommon.

For example, if we look in Appendix A1, we see that the proportion associated with 3.59 is $.0002$. That means that only 0.02% of cases fall above 3.59 or below -3.59 (or $.04\%$ of cases beyond both). That is, we'd expect only four cases per 10000 to be this or more extreme. That doesn't mean these scores don't exist; in fact, you'll find these in your own data from time to time if you work with statistics long enough. It just means that you don't see them very often!

4.3.4 CONVERTING FROM PROPORTIONS TO Z-SCORES

Just as we might need to convert a z -score into a proportion, we might need to convert a proportion into a z -score. To do this, we simply follow the procedure above in reverse.

We need to do this most often when we are interested in **percentiles**. A percentile represents the point at which a particular percentage of a dataset falls below a specific point. For example, if the 90th percentile equals 10, 90% of data are smaller than 10 and 10% of data are bigger than 10. The 50th percentile is the median (see Chapter 3, p. 81) – the point at which 50% of data are bigger and 50% of data are smaller. You might see a percentile expressed as a proportion (e.g. the 30th percentile might be represented as .30).

So when converting between proportions and z -scores, you might see a question like this: ‘What z -score corresponds to the 63rd percentile?’ When trying to answer such questions, draw it to help yourself visualize what the question is asking. The 63rd percentile is greater than the median (the 50th percentile), so the dividing line will be somewhere on the right side of the distribution. 63% of scores fall below the point you’re interested in, which means that 37% (100% – 63%) fall above it. It doesn’t matter exactly where you draw the line, as long as it’s on the correct side of the median. What you draw should look something like Figure 4.17.

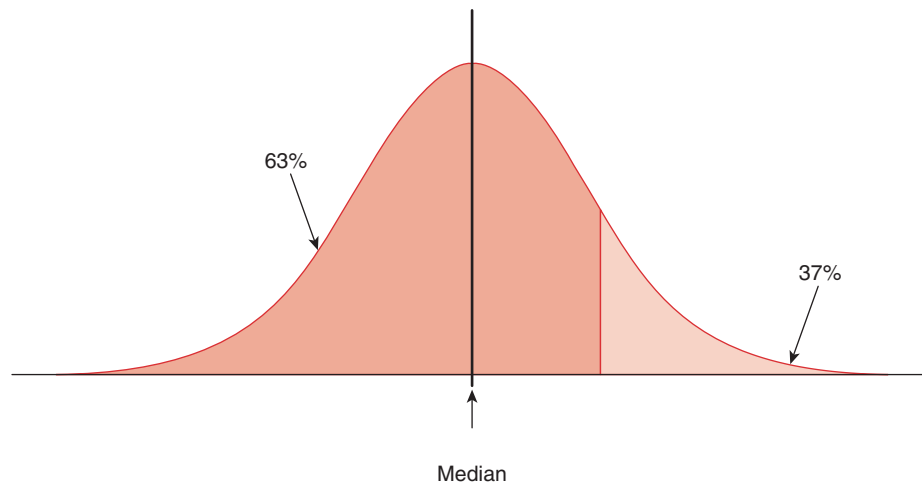


FIGURE 4.17 The 63rd percentile of a z -distribution

Although the question asks about the ‘63rd percentile’, the z -table only contains information about the tails. So instead of 63%, you’ll be looking up 37%.

Next, scan the z -table in Appendix A1 until you find the two proportions that surround .37 – one bigger and one smaller. In this case, the two closest proportions are .3707 and .3669. The cell we will concentrate on will be .3707, as it is closer to .37. As before, add together the column and row headings to determine the associated z -score. In this case, .3707 is in row .3 and column .03 (Figure 4.18). Thus, the z -score we’re looking for is $.3 + .03 = .33$. The 63rd percentile corresponds to $z = .33$.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451

FIGURE 4.18 Identifying the z-score corresponding to the 63rd percentile

If the z-score we were looking for was on the left side of the median, we'd simply use a minus sign with that value to reflect the negative z-score. You can use this general procedure for any type of conversion between a proportion and a z-score.

4.3.5 CONVERTING BETWEEN PROPORTIONS AND RAW DATA

When converting between raw data, z-scores and proportions, you will always use some combination of the tools listed above. These relationships are illustrated in Figure 4.19.

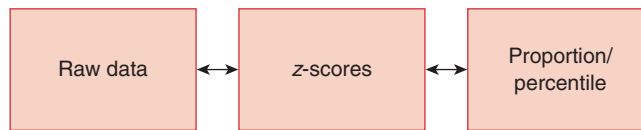


FIGURE 4.19 Illustration of necessary conversions between raw data and proportions/percentiles

You can consult Figure 4.19 whenever asked to make a conversion. For example, imagine you were provided a proportion and asked to identify the raw score at which that proportion would be found. Since you can never convert directly from a proportion, you must convert to a z-score first.

For example, in our case study, what would Jill do if she wanted to identify how many cars an employee would have sold in August if that employee was performing at the 20th percentile. From the data, we know that the mean of August sales is 9.727273 and its standard deviation is 2.805118.

First, Jill needs to draw it out (Figure 4.20).

This time, 20% of values fall in the tail, so that's the number we'll look up. When looking in the z-table (Appendix A1), the two closest values we can find are .2005 and .1977. The closer is .2005, so we'll use that – its associated z-value is .84 (Figure 4.21). Because the tail is on the left side of the distribution, we know that the z-score is negative; thus, the z-score associated with the 20th percentile is $z = -.84$.

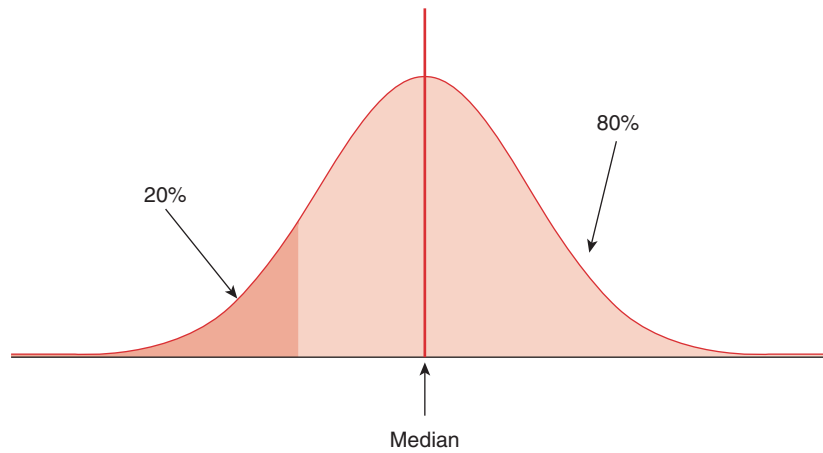


FIGURE 4.20 Drawing of 20th percentile with relevant percentages

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

FIGURE 4.21 Looking up the 20th percentile in the z-table (Appendix A1)

Next, we need to use the formula above to convert z -scores to raw data. The mean and standard deviation are provided in the question, so we'll use those values in the formula (Figure 4.22).

$$\begin{aligned}
 x &= zs + \bar{x} \\
 x &= -.84 * 2.805118 + 9.727273 \\
 x &= -2.356299 + 9.727273 \\
 x &= 7.370974 = 7.37
 \end{aligned}$$

FIGURE 4.22 Step by step calculation of raw score from a percentile

Thus, the 20th percentile is associated with 7.37 cars sold.

SUMMARY OF CONVERSIONS WITH Z-SCORES

Conversion from	Information needed	Procedure/formula used
Raw data to z-score	A score, the mean, and the standard deviation. You may need to calculate the mean and standard deviation or they may be provided for you.	<p>Sample mean (see Ch. 3, p. 83)</p> $z = \frac{x - \bar{x}}{s}$ <p>Sample standard deviation (see Ch. 3, p. 89)</p> <p>Population mean (see Ch.3, p. 88)</p> $z = \frac{x - \mu}{\sigma}$ <p>Sample standard deviation (see Ch. 3, p. 89)</p>
z-scores proportion/percentile	A z-score and a z-table (found in Appendix A1).	Find the z-score in the z-table by scanning across columns and down rows, adding the column head and row head together. For example, column .03 and row 1.0 intersect at $z = .03$. The number you find there is the proportion of values found in the tail. If you are interested in the proportion of values in the tail, use this value. If you are interested in the proportion of values in the rest of the distribution, subtract this value from 1.
Proportion/percentile to z-score	A proportion/percentile and a z-table (found in Appendix A1).	Find the proportion in the z-table by looking for the closest value to the proportion you are looking for. Next, add the numbers found at the left of the row and the top of the column to compute the z-score. For example, the closest proportion to .01 is .0099, which corresponds to $z = 2.33$. If you are identifying a proportion in the lower tail (left side of the distribution), your z-score will be negative.
z-score to raw data	A z-score, the mean, and the standard deviation. You may need to calculate the mean and standard deviation or they may be provided for you.	<p>Sample mean (see Ch. 3, p. 83)</p> $x = zS + \bar{x}$ <p>Sample standard deviation (see Ch. 3, p. 89)</p> <p>Population mean (see Ch.3, p. 88)</p> $x = z\sigma + \mu$ <p>Sample standard deviation (see Ch. 3, p. 89)</p>

FIGURE 4.23 Summary table of conversions involving z-scores

4.4 APPLYING PROBABILITY DISTRIBUTIONS

To apply what you've learned from this chapter, consider the following case study, questions posed about that case study, and discussion of those questions.

4.4.1 APPLICATION CASE STUDY

Raphael is head sales agent of Where the Heart Is, a residential estate agency. He is responsible for tracking sales for the entire firm. One of the owner's goals is for Where the Heart Is to sell across the full spectrum of the housing market, from the smallest flat to the largest mansion. This, the owner believes, will make the broadest impression on potential clients through word-of-mouth, improving the overall number of clients.

What statistical approach might Raphael use to ensure this goal is being met? He has access to the full list of houses put on the market over the past year, as well as the list of houses managed by Where the Heart Is. What statistics could he use to ensure that the owner's goal is being met?

4.4.2 APPLICATION DISCUSSION

Like most real business scenarios calling for statistics, there are multiple ways to approach Raphael's problem. However, an obvious approach here would be to use z -scores.

Raphael actually has a population – the current housing market. Since the owner's goal is for house values from Where the Heart Is to match those of the market, we can examine the z -score of each house represented by Where the Heart Is in relation to that market.

For example, if houses for Where the Heart Is were $z = +.5$, $z = +1.0$ and $z = +1.3$, we'd know from the mean of these z -scores ($z = +.93$) that the houses Where the Heart Is tends to sell are above average for the current housing market. We could further conclude from that average that houses sold were one standard deviation above the average house sold in the population, which is around the 84th percentile. If the owner's goal was being met, the average z -score should be close to zero, because the average house sold by Where the Heart Is should be approximately equal to the average house on the overall market.

EXPLORING PROBABILITY DISTRIBUTIONS IN EXCEL AND SPSS

EXCEL



Download the Excel dataset for the demonstration below as **chapter4.xls**. As you read this section, try to apply the terms you've learned in this chapter to the dataset and follow along with Excel on your own computer.



You can also get a video demonstration of the section below under **Excel Video: Chapter 4**.

In our case study, Jill is trying to identify which employees are consistently weak in sales. She has collected data on employee sales for the months July through December. Try using the techniques you learned in Chapter 3 to identify the mean and standard deviation of each month. You'll find that there is quite a bit of variation from month to month (if you don't find the values given in Figure 4.24, review Chapter 3's lesson on Excel).

	July	August	September	October	November	December
Mean	9.75	9.73	10.04	10.36	12.33	15.20
SD	2.65	2.81	2.98	3.42	4.83	4.53

FIGURE 4.24 Means and standard deviations from case study dataset in Excel

If we look at any one month, we may not get the whole story; perhaps an employee was simply having a bad month, but it doesn't reflect their overall performance. But if we add all sales across all months for each employee, we may miss patterns of poor performance. For example, what if an employee does well during the December rush, but is consistently poor during all other months? To address this problem, we will convert each month's sales for each employee to z-scores, so that we can identify how well each employee was doing each month relative to his or her co-workers.

First, create six new columns to represent where our new z-values will go (Figure 4.25).

	G	H	I	J	K	L
z	July	August	September	October	November	December

FIGURE 4.25 Six new column labels for z-scores in Excel

In G2, we want the number we find to represent the z-score for the first employee's July sales. We'll create this formula just like we would if we were doing it by hand: a z-score equals the score minus the mean, all divided by the standard deviation. We'll use the mean and standard deviation formulas we learned in Chapter 3.

- Excel 2010 or later:** To compute the z-score in G2, type: $= (A2 - AVERAGE(A:A)) / STDEV.S(A:A)$.
- Excel 2007 or earlier:** To compute the z-score in G2, type: $= (A2 - AVERAGE(A:A)) / STDEV(A:A)$.

You should end up with something like Figure 4.26.

		G2						
		fx						
		=(A2-AVERAGE(A:A))/STDEV(A:A)						
	A	B	C	D	E	F	G	
1	July	August	September	October	November	December	July	
2		12	9	13	15	16	18	0.851647

FIGURE 4.26 Computed z-scores for first case's data in Excel

(Continued)

(Continued)

This formula is a little different than the ones we've used before now. Instead of naming specific cells – like A2:A56 – we've named an entire column. Excel will ignore cells with text in them, like the word 'July'.

This is an important change because it changes how Excel fills formulas. Do you remember the fill from Chapter 3? We did a fill when we clicked on the little black box at the bottom-right corner of a cell and dragged it right to copy the content of the cells we'd highlighted. When you run a fill, Excel automatically updates the references of all cells copied to match wherever you copied them. For example, if you filled one cell right from a cell that referenced A2, the new cell's formula would reference B2 instead (because B2 is one cell to the right of A2).

When filling down, this becomes a problem if we've named specific cells. For example, if we'd named A2:A56 and filled down, the next cell would contain A3:A57, the next would contain A4:A58, and so on. We want our fill to always reference the entire column, so this is incorrect. By referencing A:A instead, Excel will simply copy A:A in every cell. When we fill right, Excel will still know to convert our A:A to B:B.

Do this now – fill down from G2 by highlighting G2 and click-dragging the little black box down to G56. Once at G56, click-drag again to the right to L56. You should now have a new table of values with z-scores for every employee for every month, as in Figure 4.27.

	A	B	C	D	E	F	G	H	I	J	K	L
1	July	August	September	October	November	December	zJuly	zAugust	zSeptember	zOctober	zNovember	zDecember
2	12	9	13	15	16	18	0.851647	-0.25927	0.994109	1.356605	0.760294	0.618472
3	13	13	12	12	12	18	1.229393	1.166699	0.658674	0.478802	-0.06775	0.618472
4	10	10	12	10	10	16	0.096154	0.097225	0.658674	-0.1064	-0.48177	0.176706
5	11	11	11	11	11	14	0.4739	0.453716	0.323238	0.186201	-0.27476	-0.26506
6	3	5	7	6	5	8	-2.54807	-1.68523	-1.0185	-1.27681	-1.51682	-1.59036
7	10	10	10	9	10	12	0.096154	0.097225	-0.0122	-0.399	-0.48177	-0.70683
8	8	9	11	11	15	16	-0.65934	-0.25927	0.323238	0.186201	0.553283	0.176706
9	9	8	9	9	9	13	-0.28159	-0.61576	-0.34763	-0.399	-0.68878	-0.48594
10	10	11	11	14	15	19	0.096154	0.453716	0.323238	1.064004	0.553283	0.839355
11	9	9	10	10	9	12	-0.28159	-0.25927	-0.0122	-0.1064	-0.68878	-0.70683
12	10	10	14	14	13	15	0.096154	0.097225	1.329545	1.064004	0.139262	-0.04418

FIGURE 4.27 Dataset with z-scores filled down in Excel

Now that we have z-scores for every month, we can more meaningfully compute a mean. Create a new column M with the mean of scores in columns G through L. If you aren't sure how to do this, review the Excel portion of Chapter 3 (see p. 96). You should end up with something like Figure 4.28.

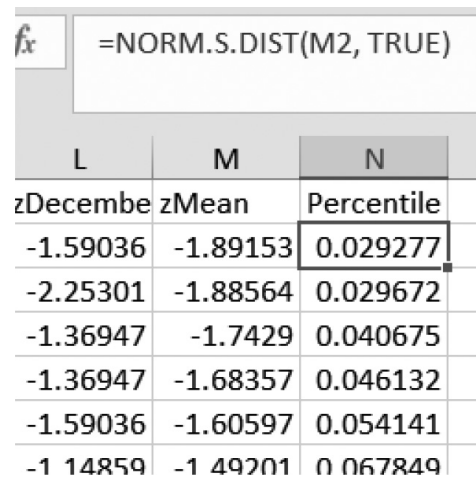
G	H	I	J	K	L	M
zJuly	zAugust	zSeptember	zOctober	zNovember	zDecember	zMean
0.851647	-0.25927	0.994109	1.356605	0.760294	0.618472	0.72031
1.229393	1.166699	0.658674	0.478802	-0.06775	0.618472	0.680715
0.096154	0.097225	0.658674	-0.1064	-0.48177	0.176706	0.073431
0.4739	0.453716	0.323238	0.186201	-0.27476	-0.26506	0.149539
-2.54807	-1.68523	-1.0185	-1.27681	-1.51682	-1.59036	-1.60597
0.096154	0.097225	-0.0122	-0.399	-0.48177	-0.70683	-0.2344
-0.65934	-0.25927	0.323238	0.186201	0.553283	0.176706	0.05347
-0.28159	-0.61576	-0.34763	-0.399	-0.68878	-0.48594	-0.46978
0.096154	0.453716	0.323238	1.064004	0.553283	0.839355	0.554958
-0.28159	-0.25927	-0.0122	-0.1064	-0.68878	-0.70683	-0.34251
0.096154	0.097225	1.329545	1.064004	0.139262	-0.04418	0.447002

FIGURE 4.28 Mean calculated from six z-scores in Excel

Now that we have mean z-scores for each employee, we should convert them to percentiles to make the z-scores more interpretable. Create a new column N labelled 'Percentile'.

- 1 **Excel 2010 or later:** To compute the percentile in N2, type: =NORM.S.DIST(M2, TRUE).
- 2 **Excel 2007 or earlier:** To compute the percentile in N2, type: =NORMSDIST(M2).

You should end up with percentiles as shown in Figure 4.29.




L	M	N
zDecember	zMean	Percentile
-1.59036	-1.89153	0.029277
-2.25301	-1.88564	0.029672
-1.36947	-1.7429	0.040675
-1.36947	-1.68357	0.046132
-1.59036	-1.60597	0.054141
-1.14859	-1.49201	0.067849

FIGURE 4.29 Percentiles computed from mean z-scores

Sort the data in column N in ascending order to see the lowest-ranking employees first. If you don't remember how to sort, see Chapter 2 (p. 47).

Now the situation is a bit clearer. Roughly a dozen employees stand out as particularly low sellers, month over month. Their z-scores are always negative (they are always below average), and there's a fairly steep decline in percentiles down to this lower group. While other employees may have good months and bad months, these employees are consistently poor. That means these employees are definitely the ones that Jill should target. But what to do with them is up to her.

SPSS

 Download the SPSS dataset for the demonstration below as **chapter4.sav**. As you read this section, try to apply the terms you've learned in this chapter to the dataset and follow along with SPSS on your own computer.

 You can also get a video demonstration of the section below under **SPSS Videos: Chapter 4**.

In our dataset, Jill is trying to identify which employees are consistently weak in sales. She has collected data on employee sales for the months July through December. Try using the techniques you learned in Chapter 3 to identify the mean and standard deviation of each month. You'll find that there is quite a bit of variation from month to month (if you don't find the values shown in Figure 4.30, review Chapter 3's lesson on SPSS).

(Continued)

(Continued)

Descriptive Statistics			
	N	Mean	Std. Deviation
July	55	9.75	2.647
August	55	9.73	2.805
September	55	10.04	2.981
October	55	10.36	3.418
November	55	12.33	4.831
December	55	15.20	4.527
Valid N (listwise)	55		

FIGURE 4.30 Descriptive statistics from case study dataset in SPSS

If we look at any one month, we may not get the whole story; perhaps an employee was simply having a bad month, but it doesn't reflect their overall performance. But if we add all sales across all months for each employee, we may miss patterns of poor performance. For example, what if an employee does well during the December rush, but is consistently poor during all other months? To address this problem, we will convert each month's sales for each employee to z-scores, so that we can identify how well each employee was doing each month relative to his or her co-workers.

In SPSS, we do this with a tool we first tried in Chapter 3 – Descriptives. Open **Analyze**, then **Descriptive Statistics**, then **Descriptives** (Figure 4.31).

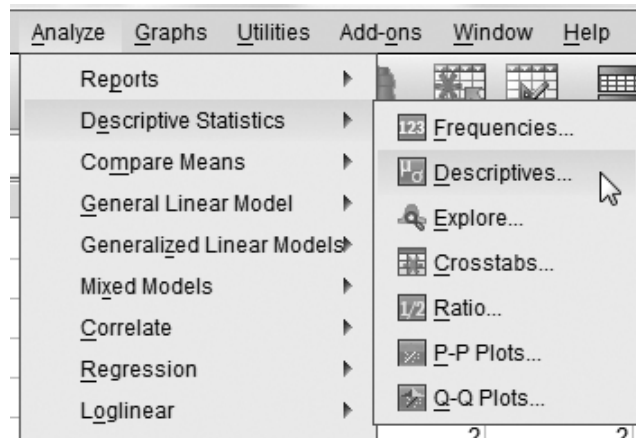


FIGURE 4.31 Menu option to run Descriptive statistics in SPSS

Move all of your variables to the right (again, see Chapter 3 if you've forgotten how).

This time, we're going to change one little thing – the checkbox that says 'Save standardized values as variables'. By checking this box, we tell SPSS not only to compute descriptives, but also to create new variables containing z-scores for each variable we are looking at. You should end up with something like Figure 4.32.

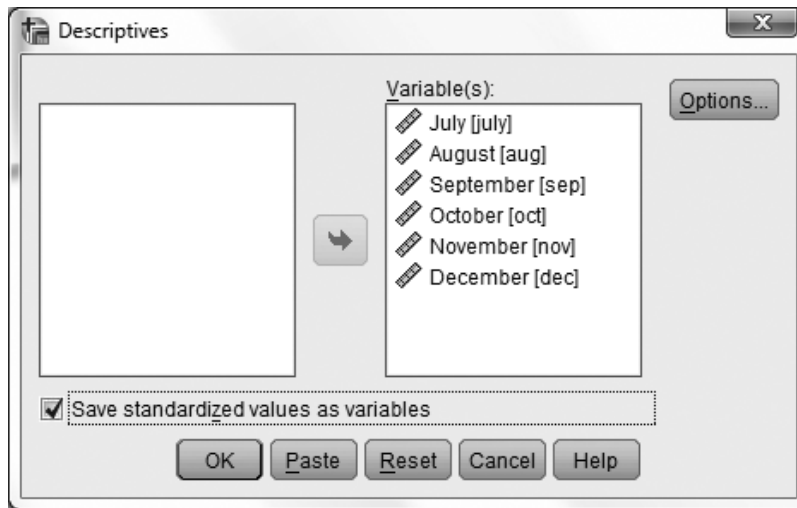


FIGURE 4.32 Descriptive statistics dialogue with option to save z-scores selected in SPSS

Click **OK**. The output pane will pop up as usual, but we're not worried about that this time. Instead, turn back to your raw data. You'll see six new variables containing z-scores (Figure 4.33).

	july	aug	sep	oct	nov	dec	Zjuly	Zaug	Zsep	Zoct	Znov	Zdec
1	12	9	13	15	16	18	85165	-25927	.99411	1.35651	76029	.61847
2	13	13	12	12	12	18	1.22939	1.16670	.62667	.47880	-.06775	.61847
3	10	10	12	10	10	16	.09615	.09722	.65867	-.10640	-.48177	-.17671
4	11	11	11	11	11	14	47390	45372	.32324	.18620	-.27476	-.26506
5	3	5	7	6	5	8	-2.54807	-1.68523	-1.01850	-1.27681	-1.51682	-1.59036
6	10	10	10	9	10	12	.09615	.09722	-.01220	-.39900	-.48177	-.70683
7	8	9	11	11	15	16	-.65934	-.25927	.32324	.18620	55328	-.17671
8	9	8	9	9	9	13	-.28159	-.61576	-.34763	-.39900	-.68878	-.48594

FIGURE 4.33 Dataset with new z-score variables in SPSS

Each set of z-scores has been calculated exactly as we would calculate it by hand.

Now that we have z-scores for every month, we can more meaningfully compute mean performance across months. To do this, we'll use a new tool called Compute. Open the **Transform** menu and select **Compute Variable** at the top (see Figure 4.34).

(Continued)

(Continued)

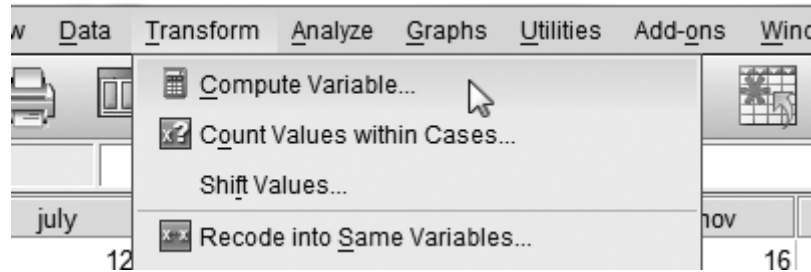


FIGURE 4.34 Menu option to compute values for new variables from current variables

You might be wondering why we're using a different tool to compute a mean. The difference is the type of mean. Before, we were interested in computing the mean score within a variable, so we used Descriptives. This time, we're interested in *creating a new variable* that is the mean of other variables. Any time you need to create a new variable, you'll generally use Compute.

If you've done this successfully, the Compute dialogue box will pop up. This looks fairly complicated, but we don't need to worry about anything here for now except the Target Variable and Numeric Expression sections. The Target Variable section lets you specify the name of your new variable. This can be whatever you want, as long as it doesn't start with a number or contain spaces and inappropriate symbols (SPSS will tell you if you try to name it something inappropriate, and it's easy to change). In our case, we want to compute the mean z-score, so we'll just call it zMean.

In the Numeric Expression section, we provide code to SPSS to tell it what to put into the new variable. In this case, we want the mean of the six new z-score variables we just created.

In the Numeric Expression section, type: `MEAN(Zjuly, Zaug, Zsep, Zoct, Znov, Zdec)`.

If you don't trust yourself to spell the variable names correctly, you can also click-drag the variables or use the arrow button, just like we've been doing in other SPSS dialogue boxes. You should end up with something like Figure 4.35.

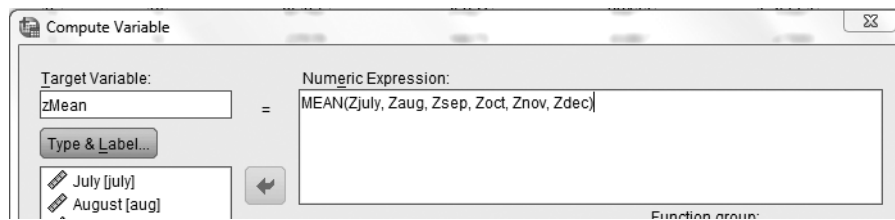


FIGURE 4.35 Compute variable dialogue in SPSS, with expression to compute new variable with mean of the six z-score variables

Click **OK**. Once again, the output pane will pop up, but we're interested in the dataset itself. Click back over to the main dataset, and you should now have a new column of mean z-scores at the far right (you may need to scroll to see them) (see Figure 4.36).

	Znov	Zdec	zMean
661	.76029	.61847	.72
880	-.06775	.61847	.68
640	-.48177	.17671	.07
620	-.27476	-.26506	.15
681	-1.51682	-1.59036	-1.61
900	-.48177	-.70683	-.23
620	.55328	.17671	.05
900	-.68878	-.48594	-.47
400	.55328	.83936	.55

FIGURE 4.36 Newly computed mean variable in SPSS

Now that we have mean z-scores for each employee, we should convert them to percentiles to make the z-scores more interpretable. Once again open the Compute dialogue, but this time, change two things:

In the Target Variable section, type: Percentile.

In the Numeric Expression section, type: CDF.NORMAL(zMean,0,1).

Remember to change the Target Variable name, or you could overwrite your other variables instead of creating a new one!

Once ready, the dialogue box should look like Figure 4.37.

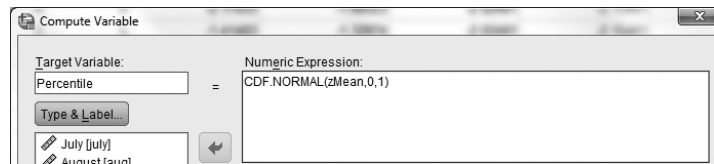


FIGURE 4.37 Compute variable dialogue in SPSS, with expression to compute new variable with percentile conversions of dataset z-scores

This formula is a little denser than the others we've covered, so let's break it down – see Figure 4.38.

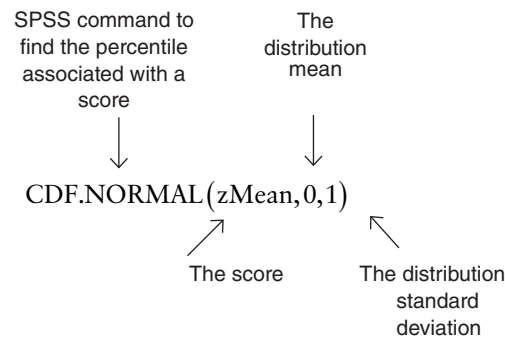


FIGURE 4.38 Annotated SPSS computation formula to convert z-scores to percentiles

(Continued)

(Continued)

In this case, we have a distribution of z-scores, so we want to know the percentile associated with a standard normal distribution – this distribution has a mean of 0 and standard deviation of 1. By including 'zMean' as the score, SPSS uses the 0 and 1 for every calculation but changes each zMean for each row of data.

You should end up with Figure 4.39 (you may need to scroll further right).

	Zdec	zMean	Percentile
129	.61847	.72	.76
75	.61847	.68	.75
177	.17671	.07	.53
176	-.26506	.15	.56
182	-1.59036	-1.61	.05
177	-.70683	-.23	.41
128	.17671	.05	.52
178	-.48594	-.47	.32
128	.83936	.55	.71

FIGURE 4.39 Final dataset in SPSS with percentile conversions from zMean variable

Sort the data in the Percentile column in ascending order to see the lowest-ranking employees first. To do this, click on **Data** and then **Sort Cases** (Figure 4.40).

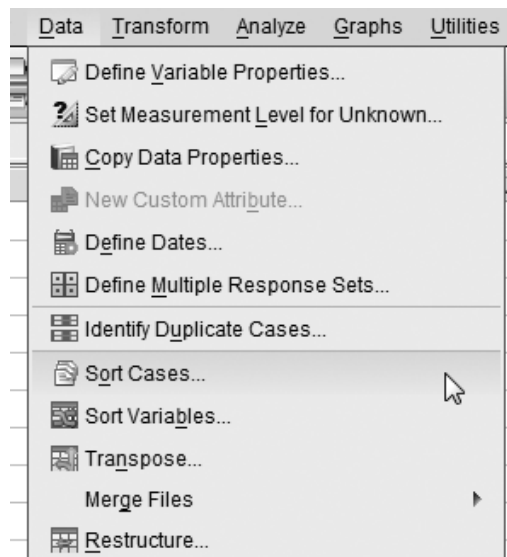


FIGURE 4.40 Menu option to sort data in SPSS

In the next dialogue, scroll down and drag Percentile to the right or use the arrow button. Ensure that 'Ascending' is selected so that the scores are ordered from smallest to largest (Figure 4.41).

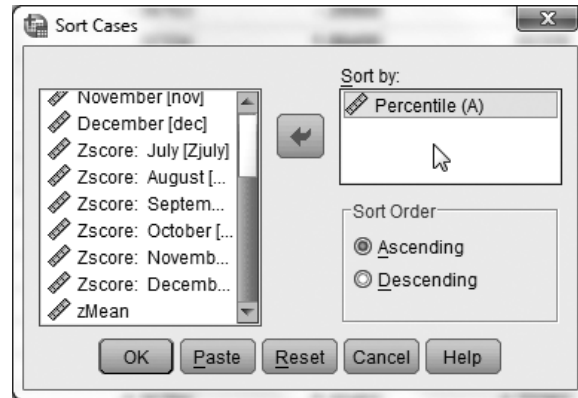


FIGURE 4.41 Dialogue to sort cases in SPSS

Click **OK**. At the top of your dataset, you'll see the lowest-performing sellers. Now the situation is a bit clearer. Roughly a dozen employees stand out as particularly low sellers, month over month. Their z-scores are always negative (they are always below average), and there's a fairly steep decline in percentiles down to this lower group. While other employees may have good months and bad months, these employees are consistently poor. That means these employees are definitely the ones that Jill should target. But what to do with them is up to her.

STATISTICS IN THE REAL WORLD

 These web links can be accessed directly from the book's website.

YouTube channel Numberphile gives an example of how probability doesn't always work the way you intuitively think it should, which could lose you money on a game show: <https://ed.ted.com/featured/PWb09pny>.

A demonstration from the National HE STEM Programme shows how normal distributions occur naturally given randomness alone using a real-life Galton Board: www.youtube.com/watch?v=6YDHBVlvs.

How do you see normal distributions of data in your everyday life?.

TEST YOURSELF



After you've completed the questions below, check your answers online.

- 1 What distribution would you expect and why for each of the following situations?
 - a Tossing a coin.
 - b Counting employee absences.
 - c Collecting survey data.

(Continued)

(Continued)

- 2 Using appropriate rounding as described in this chapter, what is the final answer for each of these computations?
- a $4.12 * 2.64$.
 - b $(7.1/2.31)/4.1$.
 - c $2.5^2 + 3.33^2$.
- 3 If $\bar{x} = 5$ and $s = 2$, calculate x for the following z-scores:
- a $z = 1.5$.
 - b $z = -3$.
 - c $z = 2.25$.
- 4 Without referencing a z-table, determine what proportion of cases we would expect to fall
- a Between $z = -3$ and $z = 1$.
 - b Below $z = -1$.
 - c Below $z = 2$.

DATA SKILL CHALLENGES



After you've completed the questions below, check your answers online.

Remember to try these calculations by hand and in the statistical program of your choice; the answers should agree.

- 1 Given this dataset: 1, 3, 2, 5, 4, 3, 2
- a Convert each of these values to a z-score.
 - b What percentage of cases would you expect to fall above 2.5?
 - c What score would be at the 20th percentile?
- 2 Given this dataset: 3, 6, 2, 1, 2, 3, 4
- a Convert each of these values to a z-score.
 - b What percentage of cases would you expect to fall below 2?
 - c What score would be at the 90th percentile?
- 3 Given this dataset: 5, 5, 7, 2, 3, 4, 4
- a Convert each of these values to a z-score.
 - b What percentage of cases would you expect to fall above 4?
 - c What score would be at the 75th percentile?
- 4 Given this dataset: 2, 3, 3, 1, 4, 2, 3
- a Convert each of these values to a z-score.
 - b What percentage of cases would you expect to fall below 3?
 - c What score would be at the 40th percentile?

NEW TERMS

area under the curve: chance: classical method of assigning probability: distribution: intersection: percentile: Poisson distribution: probability: raw data: relative frequency of occurrence method of assigning probability: standard normal distribution: standardization (or standardized): uniform distribution: union: z-distribution: z-score:

NEW STATISTICAL NOTATION AND FORMULAS

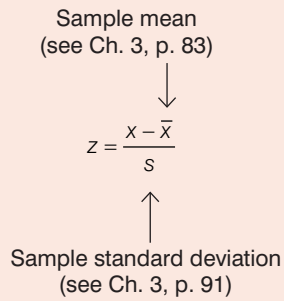


FIGURE 4.42 Annotated formula to compute a z-score from sample data

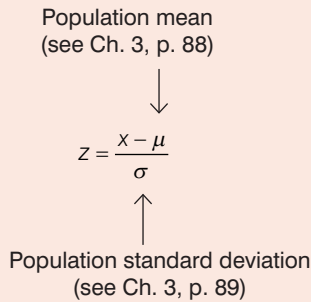


FIGURE 4.43 Annotated formula to compute a z-score from population data

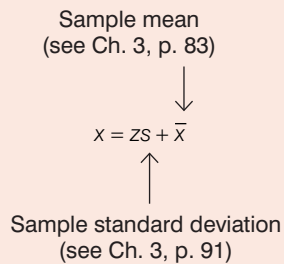


FIGURE 4.44 Annotated formula to compute a raw score from a sample z-score

(Continued)

(Continued)

Population mean
(see Ch. 3, p. 88)

↓

$$x = z\sigma + \mu$$

↑

Population standard deviation
(see Ch. 3, p. 89)

FIGURE 4.45 Annotated formula to compute a raw score from a population z-score

Visit <https://study.sagepub.com/landers2e> for free additional online resources related to this chapter.