Statistical Modelling and the World of Business Statistics

CHAPTER 16

CONNECTIONS CHAPTER

In Chapters 9 through 15 we developed a large set of methods of inference, *z*-tests, *t*-tests, chi-square tests, and linear regression. In the current chapter we develop a framework into which all these tests can fit, and a way to choose the appropriate procedure for a given data analysis situation. We also provide a little taste of the wide world of business statistics methods beyond where we end our study.

LEARNING OBJECTIVES

1. Form a statistical model to address a research question
2. Choose an appropriate statistical technique for a specific research question
3. Recognize the need for other techniques beyond this book

Rogers Communications Inc.

This book is about communication. What do data communicate to us and how do we communicate to others what we learn from the data? So our story to begin this chapter is about Rogers, a diversified public telecommunications and mass media company.

Rogers began in 1924 as the Standard Radio Manufacturing Corporation when Edward Rogers Sr. invented the world's first alternating current (AC) radio tube, which enabled radios to be powered by ordinary household electric currents. The invention was a technological breakthrough that helped popularize radios around the world. Rogers died very young at 38; his son E.S. "Ted" Rogers Jr. later carried on his father's business, acquiring CHFI-FM in 1960, then creating Rogers Broadcasting Limited from Rogers Radio Broadcasting in 1962. The business continued to evolve, to Rogers Cablesystems, Cantel, and then Rogers Communications Inc. in 1986. Following the death of Ted Rogers in 2008, control of Rogers Communications passed to the Rogers Control Trust, which operates, in part, for the benefit of current and future generations of the Rogers family.

With headquarters in Toronto, Rogers has three primary business lines. Rogers Cable is the largest cable television provider in Canada; it offers analog and digital television, high-speed internet access, residential telephony services, and home service monitoring. One division is Rogers Business Solutions, which provides voice communications services, data networking, and broadband internet connectivity to small, medium, and large businesses.

Rogers Media is Canada's premier collection of category-leading media assets with businesses in radio and television broadcasting, televised shopping, publishing, and sports entertainment. Rogers Wireless entered the mobile phone market in 1985 and is Canada's largest voice and data communications services provider. They launched the Apple iPhone in Canada in 2008. In 2011, Rogers launched a home monitoring service using both its wireless network and cable network. The service lets one manage multiple utilities at home, including security sensors and cameras, the thermostat, appliances, and lighting.

Rogers is also well-known for its leading role in Canadian sports. They acquired the Toronto Blue Jays Baseball Club in 2000 and the Skydome in 2004, renaming it the Rogers Centre. Rogers also purchased the naming rights to Rogers Arena, home of the Vancouver Canucks. They sponsor the Rogers Cup of Tennis Canada. Along with their chief competitor, Bell Canada, Rogers bought a majority stake in Maple Leaf Sports & Entertainment, the parent company of the Toronto Maple Leafs. In 2013, Rogers signed a landmark 12-year $5.2 billion agreement with the NHL for complete broadcast and multimedia rights on all platforms in all languages.

Rogers is listed on the TSX and NYSE, and with revenues exceeding $12 billion in 2011 and nearly 30 000 employees, Rogers is a leading name on the Canadian communications landscape.

Such a diversified company as Rogers collects massive amounts of data, and not just what their subscribers upload and download through their cellular phone data plans! Here are a few examples:

- Financial performance data, of interest to shareholders
- Surveys of customer satisfaction with wireless service, outlet stores, and technical assistance through their Canadian-based call centres
- Employee productivity: sales records, response times, absenteeism
- Wireless contract renewal rates
- Internet connectivity and usage statistics
- Magazine subscriptions and readership
- Television viewer and radio listener numbers.

All the statistical techniques presented in this textbook, and many more, can be applied to help Rogers draw conclusions from data, to understand and grow the activities of the company.

Even Rogers' current logo has a mathematical aspect. The symbol is a Mobius strip that has amazing mathematical properties. Look them up!

Now that we've reached the final chapter, we will look at how the wide range of techniques fits together. Is there a framework that makes it possible to see the common features of the techniques and, more importantly, how to decide which ones to use in which situations? Let's review the main headlines.

We began with a discussion of data—types, quality, sources, and sample surveys. Next we learned how to display and describe data, one variable at a time, and then in relationship with one another. We built some foundations on random variables and probability models and, in particular, the normal model. That was the first half of the book and it set the stage for statistical inference. That's been the second half of the book: confidence intervals and hypothesis tests for proportions and means with one sample, two independent samples and two related samples. And finally, we developed inference for simple and multiple regression to investigate more fully the relationships among two or more variables.

One word that showed up repeatedly was "model." We defined a **model** as a mathematical description of a real-world phenomenon. For example, the normal model is an equation that describes a common distribution displayed as a bell-shaped histogram. The *z*-model and *t*-model describe the distributions of the sample proportion and sample mean, respectively. The simple linear regression model describes a straight-line relationship between two quantitative variables. And so on. Let's look more deeply at what makes a statistical model and how statistical modelling is the key to understanding data.

16.1 Statistical Models

Science and scientists have developed countless models to explain physical situations. The models are mathematical representations that used data to estimate parameters. For example, Hooke's Law explains the relationship between the length of an extended spring and the mass hanging from the end of it. Newton's laws include the famous "force = mass × acceleration." Boyle's law in physics says that, at constant temperature, "pressure × volume = constant" for a given quantity of gas. In each of these examples there is a systematic relationship between the outcome and the predictors. Possibly the most famous model of all (no, not Gisele Bundchen, Channing Tatum, or Naomi Campbell) is Einstein's $E= mc^2$.

These are all examples of deterministic models, because the left-hand side of the equation is completely determined or explained by the right-hand side. The relationship is exact.

Statistical models are a little different. A **statistical model** has an added component. In addition to the systematic component there is a random component (also called "error," or a "stochastic" component if you want to impress people). The random component happens for a variety of reasons: measurement error, unaccounted-for factors, and natural variability between experimental units. For example, consider the height of people. Different people have different heights, because people are different! That's natural variability. But even if you measured the same person twice you would get slightly different results. That's measurement error.

Let's visualize a statistical model as a mathematical equation, as follows: write an "equal" sign. The variable(s) to the left are the outcome(s) or response(s); variables to the right are predictors or explanatory factors. But the right-hand side has one more term, representing the random component.

$$\begin{aligned}\text{Outcome} &= \text{Math Function of (Predictors)} + \text{Error}\\ &= \text{(Systematic component)} + \text{(Random component)}\end{aligned}$$

Imagine how much less impressive Einstein's model would look if it were:

$$E = mc^2 + \{\text{some other things I haven't figured out}\}.$$

We mentioned previously George Box's famous observation that, "All models are wrong, but some are useful." It is impossible to represent a real-world system exactly by a simple mathematical model. But a carefully constructed model can provide a good approximation to both the systematic and random components. That is, it can explain how the predictors affect the outcome and how big the uncertainty is. That's what we did in our study of regression.

What are the objectives of model building? Christopher Chatfield summarized them as follows:

- To provide a simple but adequate description of data
- To compare different sets of data
- To test a theory about a relationship
- To make predictions
- To give margins of error around estimates, predictions, and conclusions
- To understand the process that generated the data.

Note that this list *does not* include getting the best fit to the observed data. Recall our Chapter 15 discussion of over-fitting a model. Chatfield cautioned that the procedure of trying lots of different models until a good-looking fit is obtained is a dubious one. The purpose of model-building is not just to get the "best" fit, but rather to construct a model that is consistent, not only with the data, but also with background knowledge and with any earlier data sets. Remember, the model must apply not only to the data you have already collected but any other data that might be collected using the same procedures. In our text we have addressed the three stages in model building: formulation, estimation, and validation (checking assumptions and conditions).

Let's look at one famous and tragic illustration of how a good graph and a properly-built model could have saved lives.

The NASA Space Shuttle program operated from 1981 to 2011 and ran 135 missions to launch satellites, interplanetary probes, and the Hubble Space Telescope, to conduct science experiments in orbit, and to construct and service the International Space Station. The first orbiter, Enterprise, was built for testing and not for orbit. The original four fully functional orbiters were: Columbia, Challenger, Discovery, and Atlantis; Endeavour was added in 1991 to replace Challenger. Challenger and Columbia were lost in mission accidents in 1986 and 2003. We will investigate the Challenger disaster here.

The shuttle had a number of components: the orbiter vehicle, a pair of recoverable solid rocket boosters, and an expendable external tank. The shuttle was launched vertically, using the boosters and the orbiter vehicle's main engines fuelled by the external tank. The boosters were jettisoned before the vehicle reached orbit, and the external tank just before orbit insertion. Each booster rocket consisted of several pieces whose joints were sealed with rubber O-rings (think of them as giant rubber washers), designed to prevent the release of hot gases during combustion. Each booster contained three primary O-rings (for a total of six for the craft).

On January 28, 1986, Challenger took off, as the twenty-fifth flight in the space shuttle program. Two minutes into the flight, the spacecraft exploded, killing all on board. A presidential commission that included the late Nobel-prize-winning physicist Richard Feynman determined the cause of the accident and wrote a two-volume report.

The key issue was the forecasted temperature on launch day, a chilly 31°F (just below 0°C). The coldest previous launch temperature was 53°F. After each of the 23 previous flights for which there were data, the O-rings were examined for damage. The sensitivity of O-rings to temperature was well-known; a warm O-ring had greater elasticity so it would quickly recover its shape after being compressed, but a cold one would not. The inability of the O-ring to recover its shape would lead to joints not being sealed and might result in a gas leak. The commission determined that this was the cause of the Challenger explosion.

Could this have been foreseen? Engineers discussed whether the flight should go on as planned (no statisticians were involved). Here is a simplified version of one of the arguments.

The following table gives the ambient temperature at launch and number of primary O-rings damaged during the flight.

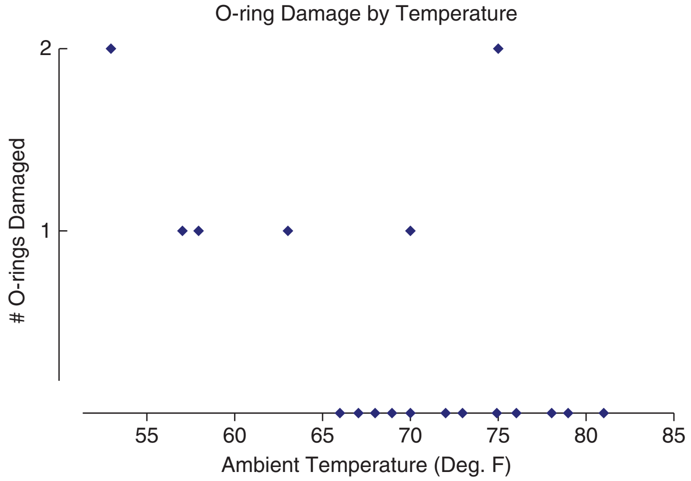
Ambient temp. at launch (°F):	53°	57°	58°	63°	70°	70°	75°
Number of O-rings damaged:	2	1	1	1	1	1	2

The table and a scatterplot (graph it yourself) shows no apparent relationship between temperature and the number of O-rings damaged; higher damage occurred at both lower and higher temperatures. Hence, just because it was cold the day of the flight doesn't imply that the flight should have been postponed or cancelled.

This is an inappropriate analysis! It ignores the 16 flights when zero O-rings were damaged. When those are included the scatterplot looks quite different, and in fact, shows a strong relationship between the number of O-rings damaged and temperature. Here is the complete data set:

Temp. (°F):	53	57	58	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	81
# damaged:	2	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	0

Here is a scatterplot of the data.



Except for the single observation in the upper right, there is a clear inverse relationship between the probability of O-ring damage and ambient temperature. Unfortunately, this plot was never made! The flaw in the analysis was not to include flights in which there was no O-ring damage.

On January 28, 1986, the ambient (outside) temperature was 31°F. Since this is off the scale of available data, extrapolation is needed. An appropriate statistical model (using a technique called logistic regression, discussed at the end of this chapter) estimates the probability of an O-ring failure at 31°F to be 96%! If you were in charge and knew that the probability of an O-ring failure was 96% would you have given go-ahead to launch?

There is a postscript to this illustration. One of the Commission's recommendations was that a statistician must be part of the ground control team for all flights.

Not only does this story emphasize the importance of graphing and using all available data, it also points out the vital role of a good statistical model. How was that model chosen? By first examining the data types. The outcome variable was binary—was there an O-ring failure, yes or no? The predictor variable was quantitative—temperature. Logistic regression (see the following section) was a suitable model for a situation where the outcome is binary and the predictor is measurement.

Perhaps this approach of classifying the role and type of variable can be applied to our other techniques and models.

16.2 A Modelling Framework

Let's look at a gender equity study that compares salaries of men and women, using a two-sample *t*-test. The following "data" table has five males and five females, with letters standing in for dollar values. A *t*-test compares the mean of the male salaries, *Mean(M)* to the mean of the female salaries, *Mean(F)*.

Male Salary	Female Salary
A	F
B	G
C	H
D	I
E	J
Mean(M)	*Mean(F)*

But we can rearrange the data as follows:

Salary ($)	Gender (0=M,1=F)
A	0
B	0
C	0
D	0
E	0
F	1
G	1
H	1
I	1
J	1

We can model the relationship by treating salary as the outcome variable and gender as the predictor variable. Salary is quantitative and gender is binary, so a two-sample *t*-test can be thought of as a model with a quantitative outcome variable and a binary predictor variable.

Let's try another situation: are male and female drivers equally likely to use a cellphone while driving? We can use a two-sample *z*-test of two proportions. Here is a data table, again for five males and five females. A *z*-test compares the proportion of male Yes responses, *Proportion(M)*, to the proportion of female Yes responses, *Proportion(F)*.

Male Driver Cell User	Female Driver Cell User
Yes	Yes
No	No
No	Yes
Yes	Yes
No	Yes
Proportion(M)	*Proportion(F)*

Again, we can rearrange the data as follows:

Cell Phone While Driving (Y/N)	Gender (0=M,1=F)
Yes	0
No	0
No	0
Yes	0
No	0
Yes	1
No	1
Yes	1
Yes	1
Yes	1

We can model the relationship by treating cell phone use as the outcome variable and gender as the predictor variable. Cell phone use is binary and gender is binary, so a two-sample *z*-test can be thought of as a model with a binary outcome variable and a binary predictor variable.

It is easy to extend this to categorical variables with more than two categories. For example, is ethnicity a predictor of smoking status (never, former, current)? We can use a chi-square test of independence to model this relationship, where smoking status is a categorical outcome variable and ethnicity is a categorical predictor variable.

Let's summarize these three situations in the following table.

Outcome Variable	Predictor Variable	Model or Technique
Quantitative	Binary	Two-sample *t*-test of means
Binary	Binary	Two-sample *z*-test of proportions
Categorical (2+ categories)	Categorical (2+ categories)	Chi-square test of independence

Now you can see how simple linear regression fits into this framework. The outcome variable and the predictor variable are each quantitative. If we allow more than one predictor variable, that's multiple regression.

Here is the previous table, expanded to include these and three possibilities we haven't yet considered.

Outcome Variable	Predictor Variable(s)	Model or Technique
Quantitative	Binary	Two-sample *t*-test of means
Binary	Binary	Two-sample *z*-test of proportions
Categorical (2+ categories)	Categorical (2+ categories)	Chi-square test of independence
Quantitative	Quantitative	Simple linear regression
Quantitative	Any combination of quantitative or categorical	Multiple linear regression
Binary	Quantitative	Simple logistic regression
Binary	Any combination of quantitative or categorical	Multiple logistic regression
Quantitative	Categorical (2+ categories)	One-way analysis of variance

One of the authors calls this his Grand Unified Theory of Statistics, which has the acronym G.U.T.S.! As the saying goes, "No guts, no glory."

The techniques called simple and multiple logistic regression and one-way analysis of variance have not been discussed. We will address them briefly in the next section. (And a fuller explanation of logistic regression is available on MyStatLab.)

However, the framework is not perfect. One-sample tests don't quite fit into this framework. Neither do situations like the paired *t*-test that has linkage between observations, or repeated measuring of subjects. Oh well, every theory has its limitations.

Outcome Variable	Predictor Variable	Model or Technique
Quantitative	None—compare to an external target instead	One-sample *t*-test of a single mean
Binary	None—compare to an external target instead	One-sample *z*-test of a single proportion

So we're right back where we started in Chapter 2. We wrote, "Section 2.2 is the most important section in the whole book. Why? … A statistical analysis cannot be done without knowing the type of variables or type of data to be analyzed." We also wrote, "Variables play different roles, and knowing the variable's *type* is crucial to knowing what to do with it and what it can tell us. The simplest and most important way to classify variables (and data) is either as *categorical* or *quantitative*."

There you have it. If you can figure out each variable's role, that is, which variable is the outcome and which is/are the predictor(s), and then decide whether they are categorical (including the simple two-category version called binary), you can pick a technique. It's that easy! The framework shows what all the important statistical modelling techniques have in common and, more importantly, how to choose an appropriate one. Let's try it out with a Mini Case Study Project from Chapter 2.

An anonymous online survey of a large undergraduate business statistics course gathered information for research about student life. Questions were asked about demographic characteristics, grades, study habits, and leisure activities. Here are some of the variables for which data were collected (they are labeled V1 through V9, for convenience—V for variable). Assume that the quantitative variables are normally distributed.

- V1 Gender (0=Male, 1=Female)
- V2 First-year overall grade (Percent)
- V3 Second-year overall grade (Percent)
- V4 Opinion of campus support services (1=Poor, 2=Fair, 3=Good, 4=Excellent)
- V5 Any paid part-time work (1=Yes, 0=No)
- V6 Total study time per week (Hours)
- V7 Mode of travel to campus (1=Car, 2=Transit, 3=Bicycle, 4=Walk/live on campus)
- V8 Regular Facebook user (1=Yes, 0=No)
- V9 Monthly amount spent on recreational activities (Dollars)

For each of the following research questions, suggest which of the techniques we have studied could be used. Here is a list of techniques to choose from. Consult the framework.

1. Is the average study time (V6) the same for males and females (V1)?

   Answer: Outcome variable is quantitative (study time); explanatory variable is binary (gender); = two-sample t-test of independent means

2. Is gender (V1) related to opinion about campus support services (V4)?

   Answer: Outcome variable is categorical (support services); explanatory variable is binary (gender); = chi-square test of independence

3. Can second-year overall grade (V3) be explained by study hours (V6), amount spent on recreational activities (V9), and paid part-time work (V5)?

   Answer: Outcome variable is quantitative (overall grade); explanatory variables are quantitative (study hours, recreational spending) and binary (part-time work); = multiple regression

4. Are there equal percentages of males and females (V1) who do paid part-time work (V5)?

   Answer: There are two possibilities here. Outcome variable is binary (paid part-time work—Yes/No); explanatory variable is binary (gender); = two-sample z-test of proportions OR chi-square test of independence

5. Do males and females (V1) achieve different first-year overall grades (V2)?

   Answer: Outcome variable is quantitative (overall grade); explanatory variable is binary (gender); = two-sample t-test of independent means

6. Is the rate of Facebook users (V8) different from the Canadian percentage?

   Answer: There is only one sample here and the variable is binary (Facebook user—Yes/No). Compare the estimate from the single sample with the external target; = one-sample z-test of a single proportion

7. Does monthly expenditure on recreational activities (V9) exceed $250?

   Answer: There is only one sample here and the variable is quantitative (recreational expenditure). Compare the estimate from the single sample with the external target; =one-sample t-test of a single mean

8. Are first-year overall grades (V2) and second-year overall grades (V3) different on average?

   Answer: Each respondent's first-year data value is matched with his/her second-year data value. Hence the outcome variable is the difference between first-year and second-year overall grade, so this is a one-sample test of the differences; = matched pairs t-test

9. Is second-year overall grade (V3) related to study hours (V6)?

   Answer: Outcome variable is quantitative (overall grade); explanatory variable is quantitative (study hours); = linear regression

   Bonus. Is mode of travel (V7) related to study hours (V6)?

   Answer: Outcome variable is quantitative (study hours); explanatory variable is categorical (mode of travel, 4 categories); = one-way analysis of variance (see section 16.3)

16.3 A Short Tour of Other Statistical Methods in Business

Our book is titled *Business Statistics: A First Course*. That suggests the possibility of a second course, and perhaps more beyond that. What topics and techniques might be found in such a book? Let's look briefly at a few of them, especially the ones identified in our modelling framework, and illustrate them with plausible research questions that might be of interest to Rogers Communications Inc.

One-Way Analysis of Variance (ANOVA) (a.k.a. comparison of multiple means)

Rogers' Human Resources Division wonders whether the average level of employee satisfaction is the same for full-time, part-time, and casual employees. A two-sample *t*-test compares the means of two independent populations. How could you compare the means of three independent populations? A crude solution would be to compare each pair of means with a series of three two-sample *t*-tests—that is, compare full-time to part-time, full-time to casual, and part-time to casual (i.e., 1 to 2, 1 to 3, and 2 to 3). But that's inefficient and prone to misinterpretation. A better solution would be to use multiple regression with two dummy predictor variables. There is a third solution: **one-way analysis of variance**, a method for comparing more than two means.

The hypotheses are: H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ vs H_a: at least one μ is different from the others. The test statistic is based on an *F*-ratio (which is why it is called ANOVA) that compares the variance *across* the means of the samples with the variance *within* the samples.

Why is one-way ANOVA preferred over a series of two-sample *t*-tests?

1. You get a single test that answers the question, "Is there ANY difference among group means?" If the answer is no, you are done. Only if you reject the null hypothesis would you go further to locate the source of the difference. So there are gains in efficiency.
2. When multiple tests are performed, each at a 5% significance level (and therefore a 0.05 chance of a Type I error), the overall chance of a Type I error increases dramatically. For example, if you had 10 means to compare, you would need 45 two-sample *t*-tests! The more tests you run, the greater the chance of finding a false positive. So you could be quite likely to get spurious significance; that is, conclude that a difference is real, when, in fact, it isn't.
3. When two-sample *t*-tests are used, the pooled variance estimate of the error variance uses only the two samples being compared, and this estimate will change in the next *t*-test. One-way ANOVA uses a pooled estimate of the error variance from *all* of the samples. So the standard error, P-values, and confidence intervals are more trustworthy because they are based on more data.

Warning #1: A common error is to confuse One-Way ANOVA with ANOVA in Regression. The common part of the name—ANOVA—represents the analysis of variance table that summarizes the computations. The "one-way" refers to the fact that we have classified the observations in one "way," according to one categorical variable.

Warning #2: The technique is really designed for experimental situations, not observational data. We'll discuss the difference below.

One-way ANOVA can also be thought of as a statistical model with one quantitative outcome variable and one categorical predictor variable (with two or more categories). We can extend this further to have two categorical predictor variables. Not surprisingly, it is called Two-Way ANOVA. And it leads to the idea of Design of Experiments.

Design of Experiments

A full discussion of design of experiments could fill its own textbook. Except for our work on sample survey design in Chapter 3, we have focused on the analysis of data.

Data come from two main types of situations. They can come from observational studies, so-called because the studies simply observe and record the behaviour or response of subjects. Surveys are the prime example of observational studies. Data can also come from experiments: studies in which the experimenter manipulates or controls attributes of what is being studied and sees the consequences. Each controlled attribute is called a factor. But if the experimenter is to manipulate things, he or she must make decisions about how many groups are to be compared, how the subjects should be assigned to each group, and how many factors are of interest; this is the **design of the experiment**.

Experiments are often harder to set up and carry out, but the conclusions that can be drawn by assessing cause and effect are stronger than observational studies, which can only look for correlations. Observational studies usually go backward in time (and so are called retrospective), for example, by asking respondents for opinions or recollections of things that have happened. Experiments usually go forward in time (and so are called prospective).

Here is an example of an experimental design. Rogers selects 30 stores to display a particular product using different types of packaging: 10 stores will get Packaging A, 10 stores will get Packaging B, and 10 stores will get Packaging C. The sales for each type of packaging are recorded. In this design, Packaging is the factor, with three levels. Sales volume is the response. If different packaging and different pricing levels are used, there would be two factors: *Packaging* and *Pricing*, and various combinations of the two would be analyzed.

Here's a more complicated design. Suppose a study is undertaken to examine the separate and combined effects of length of break during a class and the class start time on the attentiveness of university students. Three lengths of break (5, 10, and 15 minutes) are tested with two start times (8:30 a.m. and 10:30 a.m.). For nine lectures with each start time, three are taught with each length of break (i.e., three lectures use 5-minute breaks, three use 10-minute breaks, and three use 15-minute breaks). Random samples of 20 students in each of the lectures are given a test of "attentiveness." Two-way analysis of variance will compare the mean attentiveness for each of the break durations, compare mean attentiveness for early and late start times, and then assess whether the means for each break duration are dependent on the start time. A two-way analysis of variance is better than two one-way analyses of variance precisely because of the ability to test for the connection between factors. It may be that a 15-minute break is best only for the early class, while a 5-minute break is best for the later class, perhaps because the caffeine is already in effect!

Logistic Regression

How could Rogers model the likelihood of a wireless subscriber renewing a contract once it expires? That's a typical situation for **logistic regression**, an advanced statistical procedure that looks a great deal like regular linear regression, except that the response variable is binary. It takes only two values, such as success or failure, accept or reject, yes or no, or in Rogers' case, renew or not renew. If the values are denoted by 1 and 0, then the mean value of y is actually just the proportion of 1s, which we denoted by p, just as before.

There are three main reasons why the usual least squares regression is not suitable here. First, the outcome variable is obviously not normally distributed. How can it be? There are only two values! Second, the variance of the outcome variable depends on the value(s) of the predictor variables(s), thus violating the equal spread condition. And third, there is no guarantee that the least squares estimated value of p will fall between 0 and 1; but it has to, because it represents a probability, namely, the probability of getting a "yes" or "success" outcome!

Instead of using proportions, logistic regression works with odds. The odds equal the proportion of one outcome divided by the proportion of the other outcome. Gamblers are all too familiar with the concept of odds. $Odds = \frac{\hat{p}}{1 - \hat{p}}$

The odds gives us a convenient way to model the relationship between p and the predictor variables: x_1 to x_k. For Rogers, the predictor variables could include customer characteristics such as: age, gender, years as a Rogers' customer, number of calls/texts/instant messages, employment status, and so on. We take the logarithm of the odds and set it equal to the linear combination of the x-variables to get the logistic regression model: $log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$.

If there is only one x-variable it is called simple logistic regression; with more than one x-variable it is called multiple logistic regression.

Unlike linear regression, which uses least squares to estimate the parameters $\beta_0, \beta_1, \cdots, \beta_k$, logistic regression uses a method called maximum likelihood. The details are not necessary for you to use this technique. Simply rely on your software package. The software will also produce confidence intervals and hypothesis tests of the parameters, just as in linear regression. And the hypothesis tests are interpreted pretty much the same way. That is, rejecting a null hypothesis of $\beta_i = 0$ means that there is a statistically significant relationship between that x-variable and the binary outcome y-variable.

Once the model has been fit and tested it can be used to estimate the probability of a "success" for any given value of x, by solving for $\hat{p}$ as follows.

$$\hat{p} = \frac{1}{1 + \exp\left[-(b_0 + b_1x_1 + \cdots b_kx_k)\right]} = \frac{\exp\,(b_0 + b_1x_1 + \cdots b_kx_k)}{1 + \exp\,(b_0 + b_1x_1 + \cdots b_kx_k)}$$

A special case happens when x_i is also a binary (0/1) variable. In that case *exp*(b_i) can be interpreted as an odds ratio. For example, if the outcome is contract renewal (yes or no, coded as 1 or 0) and the predictor is "discount offered" (yes or no, coded as 1 or 0), then *exp*(b_i) can be interpreted as the number of times higher the chance of a renewal is when a discount is offered as when there is no discount.

Logistic regression is a powerful modelling technique for assessing the joint effect of many predictor variables (quantitative and categorical) on a binary outcome.

Factor Analysis

Rogers' Human Resources Division carries out a survey of employee attitudes about working at Rogers. The survey has 50 questions about different aspects of the workplace. How can you best group them into categories and compute summary scores?

That's a job for **factor analysis**, a technique that comes under the general heading of "methods of data reduction." It was developed as a method of uncovering a smaller number of concepts or "factors" from a larger set of measured variables. These concepts cannot easily be measured directly so they are measured indirectly instead.

For example, the Rogers' survey may ask employees to rate a series of items, using a five-point scale from "very dissatisfied" to "very satisfied." By summing up all 50 items you can get an overall measure of attitude. But there may also be subscales of interest. Some items may refer to physical environment, some to interpersonal interactions with staff, some to communications with management, some to a sense of shared decision-making and autonomy, and so on. Factor analysis is a technique to find those subscales. The basic idea is to collect variables that are most highly correlated with one another. The thinking is that if high scores on one variable go along with high scores on another, perhaps they are measuring the same thing. Factor analysis is different from our other techniques because it makes no distinction between response and predictor variables. And the computational procedures feel a bit like magic!

Cluster Analysis

Rogers' Marketing Division is interested in developing targeted advertising campaigns, and needs to know how to divide customers into identifiable subgroups or segments. This is known as market segmentation.

Cluster analysis is a set of methods designed to find similarities among cases (in this case, people) based on a set of variables, such as demographic characteristics and past customer behaviour information. It is a little like factor analysis because it, too, is a data reduction technique that makes no distinction between response and predictor variables, and because the computational procedures are complex.

If anyone ever asks you for a one-sentence description of what these two techniques can do, just tell them: Factor analysis reduces the number of variables by grouping them into a smaller set of factors, while cluster analysis reduces the number of cases by grouping them into a small set of clusters.

Nonparametric Methods

The confidence intervals and hypothesis tests for measurement data discussed in our text are based on the assumption that the populations from which the data are drawn are normally distributed. The *t*-distribution procedures are very popular because they work well even with a moderate lack of normality, as long as the samples are reasonably large. What can be done with severe non-normality, especially with small samples?

There are procedures that make no assumptions at all about the distribution of the populations. These are called **nonparametric methods** (and also distribution-free methods), because there are no distributional parameters, such as the mean or standard deviation of the normal. The tests are also known as rank tests, because they are based on the relative position of each data value in a sample, rather than the actual magnitude of each data value. For example, we know the ranking of the three Olympic medals—gold, silver, bronze—but we don't worry about the difference in actual performance that won the gold and the silver. The gold medallist could have beaten the silver medallist by a fraction of a second or by a minute; all that matters is that the gold medallist was ahead and was therefore ranked number one.

While a two-sample t-test can only compare means of two populations, there are non-parametric tests that can compare medians or even the entire distributions of two populations.

Just to introduce them to you by name, here is a short list of nonparametric procedures that correspond to the parametric procedures we developed.

Parametric (normal distribution test)	Nonparametric (rank) Test
One-sample *t*-test	Wilcoxon signed-rank test
Two-sample *t*-test for independent samples	Wilcoxon rank-sum (a.k.a. Mann-Whitney) test; and Tukey's quick test
Paired *t*-test for dependent samples	Wilcoxon signed-rank test
One-way ANOVA	Kruskal-Wallis test
… and many others!	… and many others!

Nonparametric procedures are a very handy addition to the statistical toolkit.

16.4 The Future of Business Statistics

Statistics has always been a crossover field. Its foundations are in mathematics. In economics, it is called econometrics; in psychology, it's psychometrics. Graphical displays of data include aspects of visual perception, art, and perspective, not to mention computing. We call it biostatistics if people are the cases and health care is the subject. Another separate field is epidemiology—statistical methods that deal with the incidence, distribution, and control of disease in a population. Courses in introductory statistics are taught as part of a wide range of academic programs, from accounting to zoology, from A to Z. Statistics is everywhere.

Where is Statistics headed and what new crossover fields is it encountering? New terms now in vogue are: data mining, big data, business analytics, business intelligence, data visualization, machine learning, and artificial intelligence. Very often they concern the collection of data sets so large and complex that traditional data processing applications are overwhelmed.

What is **data mining**? It is the name for a process that uses a variety of data analysis tools to discover patterns and relationships in data to help build useful models and make predictions. In particular, its purpose is to extract useful information hidden in very large databases. Many of the modelling techniques that we've covered in this book—especially multiple regression and logistic regression—are used in data mining. But because data mining has benefited from work in machine learning, computer science, and artificial intelligence, as well as statistics, it has a much richer set of tools than those we've discussed in this book. Data mining is similar to traditional statistical analysis in that it involves exploratory data analysis and modelling. But it has some different aspects too. The most important ones include: the size of the databases, the exploratory nature, the lack of a designed experiment or survey, and the automatic nature of modelling. Surprisingly, perhaps, there is no consensus on exactly what constitutes data mining.

What is business analytics? An excellent definition appears in Wikipedia. **Business analytics** "refers to the skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning." It is based on data and statistical methods, explanatory and predictive modelling. In other words, it is really just a new term for the techniques we have discussed in this textbook, and more.

The underlying theme is that business decision making (financial, managerial, policy, etc.) is improved considerably by an understanding of statistical concepts and statistical methods.

The traditional challenge facing statisticians has been *not enough data*. Today the challenge is a new one, *too much data*. Imagine the volume of data collected by Google, Facebook, Twitter, Wikipedia, Amazon, Netflix, and the GPS application Waze. In online commerce, meteorology, genomics, biological and environmental research, internet searching, finance, business informatics, remote sensing, software logs, etc. investigators regularly encounter limitations due to large data sets. According to IBM, as of 2012, 2.5 quintillion (2.5×10^{18}) bytes (10^{18} bytes is a million million MB) were created every day!

The size of a typical data warehouse makes any analysis challenging. The ability to store data is growing faster than the ability to use it effectively. Commercial data warehouses often contain terabytes (TB)—more than 1 000 000 000 000 (1 trillion) bytes or a million MB—of data (one TB is equivalent to about 260 000 digitized songs), and warehouses containing petabytes (PB—one PB = 1000 TB) are now common. The digital size of all 33 000 000 books in the U.S. Library of Congress is about 15 TB. According to *Wired* magazine, about 20 TB of photos are uploaded to Facebook every month. It's estimated that the servers at *Google* process a petabyte of data every 72 minutes. One key challenge of statistics and all its related fields is how to uncover important strategic information lying hidden within these massive collections of data.

As the statistician William G. Hunter observed more than 30 years ago, "We live not in a time of information explosion but in a time of data inundation." Statistics is how we turn all the data into information!

(Famous) Last Words

The year 2013 was designated The International Year of Statistics, a worldwide campaign promoting the power and impact of statistics on everyday life.

A special resolution in the U.S. Senate, introduced by Senator Kay Hagan, recognizes that "the science of statistics is vital to the improvement of human life because of the power of statistics to improve, enlighten, and understand" and that "statisticians contribute to the vital and excellent of myriad aspects of United States society, including the economy, health care, security, commerce, education, and research." The same is true in Canada and around the world.

We live in a data-centric world, and the future depends on data, not just their collection, but analysis, interpretation, and communication. We hope this textbook has succeeded in showing you the power and possibility that comes with knowledge of statistical thinking, and that we added quality to quantity!

We'll end by repeating the words of the great Canadian humourist, Stephen Leacock:

"I've been reading some very interesting statistics," he was saying to the other thinker.

"Ah, statistics!" said the other, "Wonderful things, sir, statistics; very fond of them myself."

WHAT HAVE WE LEARNED?

In this chapter we developed a framework that ties together the common features of the inference methods we have studied. We extended the idea of a model to a statistical model. We developed a method for choosing an appropriate inference technique by identifying the roles played by each variable and the type of data represented by each variable. We also provided brief introductions to many other statistical methods used in business applications.

- We learned that a statistical model connects an outcome variable to one or more predictor variables (the systematic component), but recognizes that there are also factors that we can't identify or measure (the random component).
- We learned that the main objective of statistical modelling is to build a model that is consistent with background knowledge, existing data, and future data. It is not simply an attempt to get the best fit to the observed data.
- We learned that statistical modelling has three stages: formulation, estimation, and validation (checking assumptions and conditions).
- We learned how each of the methods—two-sample t-test of means, two-sample z-test of proportions, chi-square test of independence, simple linear regression, and multiple regression—are examples of statistical models.
- We learned how to identify the appropriate inference method to use by determining which variable is the outcome, which variable(s) is/are the predictor variables, and whether each variable is binary, categorical, or quantitative.

Terms

Business analytics A new term for the techniques discussed in this book. It has also been defined as "the skills, technologies, applications, and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning."

Cluster Analysis A set of methods for finding similarities among cases based on a set of variables. It is used extensively in market segmentation.

Data Mining A process that uses a variety of data analysis tools to discover patterns and relationships in data to help build useful models and make predictions. In particular, its purpose is to extract useful information hidden in very large databases.

Design of Experiments Experiments are studies where the experimenter manipulates attributes of what is being studied. In that way they are different from observational studies. Designing experiments is a complex field in itself.

Factor Analysis A technique for discovering a smaller number of concepts or “factors” from a larger set of measured variables. It is used frequently in developing measurement tools.

Logistic Regression A method similar to regular linear regression, but where the model has a binary outcome variable. Simple logistic regression has one predictor; multiple logistic regression has multiple predictors, and they can be any combination of quantitative and categorical variables.

Model A mathematical description of a real-world phenomenon.

Nonparametric Methods These are procedures that make no assumptions about the population distributions and so don’t involve testing parameters such as the mean or proportion. They are based on comparing ranks, rather than the actual data values.

One-Way Analysis of Variance A method for comparing more than two means. The model has one quantitative outcome variable and one categorical predictor variable.

Statistical Model A model with an added component. It identifies a mathematical relationship between an outcome variable and one or more predictor variables plus a random error term.