3

INTRODUCTORY STATISTICS

> In spite of the immense amount of fruitful labour which has been expended in its practical applications, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved.
>
> (Fisher, 1922, p. 310)

Our statistics review includes topics that would customarily be seen in a first course in statistics at the undergraduate level, but depending on the given course and what was emphasized by the instructor, our treatment here may be at a slightly deeper level. We review these principles with demonstrations in R and SPSS where appropriate. As was true for the mathematics review of the previous chapter, should any of the following material come across as entirely "new," then a review of any introductory statistics text is recommended. Kirk (2008), Moore, McCabe, and Craig (2014), and Box, Hunter, and Hunter (1978) are relatively nontechnical sources, whereas Degroot and Schervish (2002), Wackerly, Mendenhall, and Scheaffer (2002) and Evans and Rosenthal (2010) are much deeper and technically dense in coverage. Casella and Berger (2002), Hogg and Craig (1995), and Shao (2003) are much higher level theoretically oriented texts intended mainly for mathematical and theoretical statisticians. Other sources include Panik (2005), Berry and Lindgren (1996), and Rice (1995). For a lighter narrative on the role of statistics in social science, refer to Abelson (1995).

Applied Univariate, Bivariate, and Multivariate Statistics, First Edition. Daniel J. Denis.

Companion Website: www.wiley.com/go/denis/appliedmultivariatestatistics

Because of its importance in the interpretation of evidence, we close the chapter with an easy but powerful demonstration of what makes a *p*-value small or large in the context of statistical significance testing and the testing of null hypotheses. It is imperative that as a research scientist, you are knowledgeable of this material before you attempt to evaluate *any* research findings that employ statistical inference.

3.1 DENSITIES AND DISTRIBUTIONS

When we speak of *density* as it relates to distributions in statistics, we are referring generally to theoretical distributions having *area* under their curves. There are numerous probability distributions or density *functions*. Empirical distributions, on the other hand, rarely go by the name of densities. They are in contrast "real" distributions of real empirical data. In some contexts, the identifier *normal distribution* may be given without indicating whether one is referring to a density or to an empirical distribution. It is usually evident by the context of the situation which we are referring to. We survey only a few of the more popular densities and distributions in our discussion that follows.

The univariate normal density is given by

$$f(x_i, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

where

- μ is the population mean for the given density,
- σ^2 is the population variance,
- π is a constant of approximately 3.14,
- e is a constant of approximately 2.71,
- x_i is a given value of the independent variable, assumed to be a real number.

When μ is 0 and σ^2 is 1, which implies that the standard deviation (SD) σ is also equal to 1 (i.e., $\sqrt{\sigma^2} = \sigma = 1$), the normal distribution is given a special name. It is called the *standard normal distribution*, and can be written more compactly as

$$\begin{aligned} f(x_i, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi(1)}} e^{-(x_i-0)^2/2(1)} = \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} \\ &= \frac{e^{-x_i^2/2}}{\sqrt{2\pi}} = \frac{e^{-(1/2)x_i^2}}{\sqrt{2\pi}} \end{aligned} \tag{3.1}$$

Notice that in (3.1), $e^{-(x_i-\mu)^2/2\sigma^2} = e^{-(x_i-0)^2/2(1)}$, where μ is now 0 and σ^2 is now 1. Note as well that the density depends only on the *absolute* value of x_i, because both x_i and $-x_i$ give the same value x_i^2; the greater is x_i in absolute value, the smaller the density at that point, because the constant e is raised to the *negative* power $-x_i^2/2$.

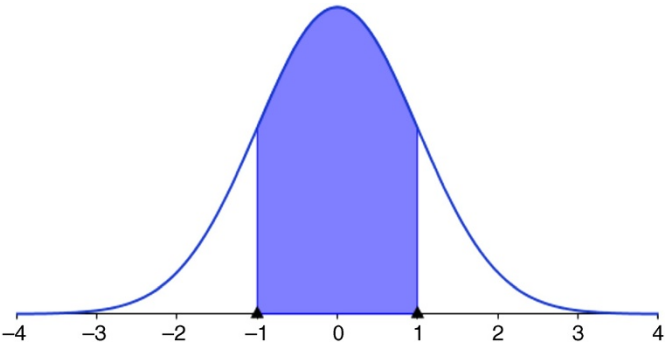


FIGURE 3.1 Standard normal distribution with shaded area from −1 to +1 standard deviations from the mean.

The *standard normal distribution* is the classic z-distribution whose areas under the curve are given in the appendices of most statistics texts, and are more conveniently computed by software. An example of the standard normal is featured in Figure 3.1.

Scores in research often come in their own units, with distributions having means and variances different from 0 and 1. We can transform a score coming from a given distribution with mean μ and standard deviation σ by the familiar z-score:

$$z = \frac{x_i - \mu}{\sigma}$$

A z-score is expressed in units of the standard normal distribution. For example, a z-score of +1 denotes that the given raw score lay one standard deviation above the mean. A z-score of −1 means that the given raw score lay one standard deviation below the mean. In some settings (such as school psychology), t-scores are also useful, having a mean of 50 and standard deviation of 10. In most contexts, however, z-scores dominate.

A classic example of the utility of z-scores typically goes like this. Suppose two sections of a statistics course are being taught. John is a student in section A and Mary is a student in section B. On the final exam for the course, John receives a raw score of 80 out of 100 (i.e., 80%). Mary, on the other hand, earns a score of 70 out of 100 (i.e., 70%). At first glance, it may appear that John was more successful on his final exam. However, scores, considered *absolutely*, do not allow a comparison of each student's score relative to their *class distributions*. For instance, if the mean in John's class was equal to 85% with a standard deviation of 2, this means that John's z-score is

$$z = \frac{x_i - \mu}{\sigma} = \frac{80 - 85}{2} = -2.5$$

Suppose that in Mary's class, the mean was equal to 65% also with a standard deviation of 2. Mary's z-score is thus

$$z = \frac{x_i - \mu}{\sigma} = \frac{70 - 65}{2} = 2.5$$

As we can see, *relative to their particular distributions*, Mary greatly outperformed John. Assuming each distribution is approximately normal, the density under the curve for a normal distribution with mean 0 and standard deviation of 1 at a score of 2.5 is

```
> dnorm(0, 1, 2.5)
[1] 0.1473081
```

where `dnorm` is the density under the curve at 2.5. This is the value of $f(x)$ at the score of 2.5. What then is the probability of scoring 2.5 or greater? To get the cumulative density up to 2.5, we compute

```
> pnorm(2.5)
[1] 0.9937903
```

The given area is represented in Figure 3.2. The area we are interested in is that at or above 2.5 (the area where the arrow is pointing). Since we know the area under the normal density is equal to 1, we can subtract `pnorm(2.5)` from 1:

```
> 1-pnorm(2.5)
[1] 0.006209665
```

We can see then the proportion of students scoring higher than Mary is in the margin of approximately 0.6% (multiply the proportion by 100).

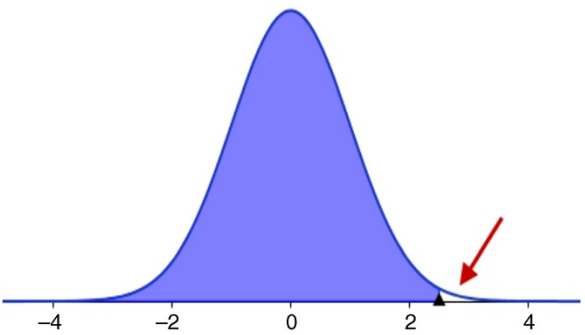


FIGURE 3.2 Shaded area under the standard normal distribution at a z-score of up to 2.5 standard deviations.

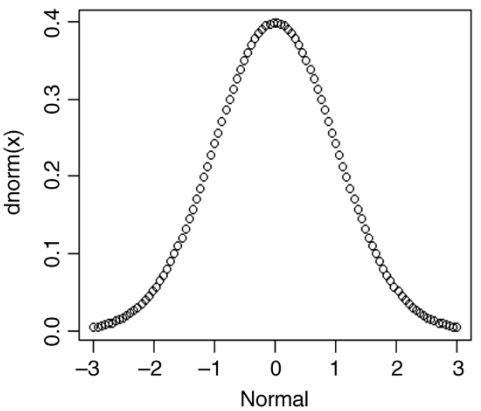
What proportion of students scored better than John in his class? Recall that his z-score was equal to −2.5. Because we know the normal distribution is symmetric, we already know the area lying below −2.5 is the same as that lying above 2.5. This means that approximately 99.38% of students scored higher than John. Hence, we see that Mary drastically outperformed her colleague when we consider their scores *relative* to their classes. Be careful to note that in drawing these conclusions, we had to assume each score (that of John's and Mary's) came from a normal distribution. The mere fact that we transformed their raw scores to z-scores in no way *normalizes* their raw distributions. Standardization *standardizes*, but it does not *normalize*.

One can also easily verify that approximately 68% of cases in a normal distribution lie between −1 and +1 standard deviations, while approximately 95% of cases lie between −2 and +2 standard deviations.

3.1.1 Plotting Normal Distributions

We can plot normal densities in R by simply requesting the lower and upper limits on the abscissa:

```
> x <- seq(from = -3, to = +3, length.out = 100)
> plot(x, dnorm(x))
```



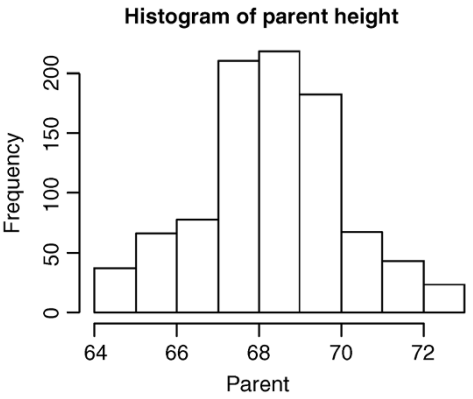
Distributions (and densities) of a single variable often go by the name of *univariate* distributions to distinguish them from distributions of two (*bivariate*) or more variables (*multivariate*).

For example, we consider some of Galton's data on parent and child heights (the height of the children were measured when they were adults, not actual toddlers). Some of Galton's data appear below, retrieved from the `HistData` package (Friendly, 2014) in R:

```
> library(HistData)
> attach(Galton)

> Galton
   parent child
1    70.5  61.7
2    68.5  61.7
3    65.5  61.7
4    64.5  61.7
5    64.0  61.7
6    67.5  62.2
7    67.5  62.2
8    67.5  62.2
9    66.5  62.2
10   66.5  62.2
```

The `library` statement in R loads the package `HistData`. From there, we `attach` the Galton data to include the object in R's workspace. We generate a histogram of parent height:

```
> hist(parent, main = "Histogram of Parent Height")
```



One can overlay a normal density over an empirical plot to show how closely observed data match that of a theoretical normal distribution, as was done by Fisher in 1925 displaying a distribution of heights of 1375 women (see Figure 3.3, taken from *Classics in the History of Psychology*[1]). R.A. Fisher is usually regarded as the father of modern statistics and among his greatest contributions was the publication of

[1]*Classics in the History of Psychology* is an online educational resource hosted by Christopher D. Green of York University in Toronto, Canada. It contains a huge selection of milestone papers and articles in the history of psychology. It can be accessed at http://psychclassics.yorku.ca/

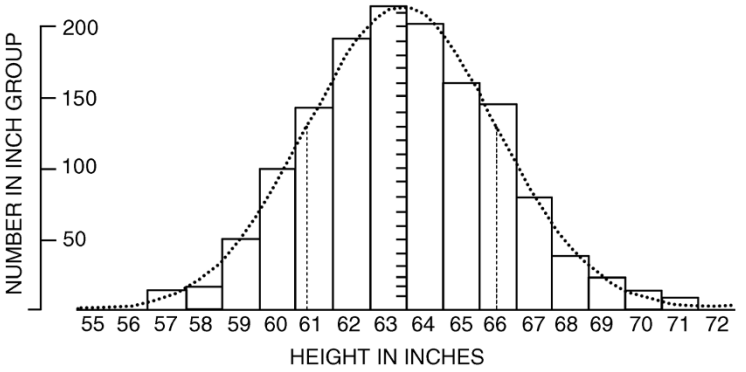


FIGURE 3.3 Fisher's overlay of normal density on empirical observations (Fisher, 1925).

Statistical Methods for Research Workers in 1925 in which he discussed such topics as tests of significance, correlation coefficients, and the analysis of variance.

We can see that the normal density serves as a close, and very convenient, *approximation* to empirical data. Indeed, the normal density has figured prominent in the history of statistics largely because it serves as a useful *model* for many phenomena, and also because it provides a very convenient starting point for much work in theoretical statistics. Oftentimes the assumption of normality will be invoked in a derivation because it makes the problem simpler and easier to solve.

3.1.2 Binomial Distributions

The binomial distribution is given by

$$\begin{aligned} p(r) &= \binom{n}{r} p^r (1-p)^{n-r} \\ &= \left(\frac{n!}{r!(n-r)!}\right) p^r (1-p)^{n-r} \end{aligned}$$

where

- $p(r)$ is the probability of observing r occurrences out of n possible occurrences,[2]
- p is the probability of a "success" on any given trial, and
- $1-p$ is the probability of a failure on any given trial, often simply referred to by q (i.e., $q = 1-p$).

[2]We can also extend the binomial distribution to a multinomial one in which instead of n trials giving rise to r occurrences, we have n trials giving rise to outcomes in k categories:

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where $\mathbf{x}$ is now a vector of random variables $\mathbf{x} = [x_1, x_2, \ldots, x_k]'$.

The binomial setting provides an ideal context to demonstrate the essentials of hypothesis testing logic. In a binomial setting, the following conditions must hold:

- The variable under study must be *binary* in nature. That is, the outcome of the experiment can result in only one category or another. For instance, the flipping of a coin has this characteristic, because the coin can either come up "head" or "tail" and nothing else (yes, we are ruling out the possibility that it lands on its side, and I think it is safe to do so).
- The probability of a "success" on each trial remains constant (or *stationary*) from trial to trial. For example, if the probability of head is equal to 0.5 on our first flip, we assume it is also equal to 0.5 on the second, third, fourth flips, and so on.
- Each trial is independent of each other trial. That is, the fact that we get a head on our first flip of the coin in no way changes the probability of getting a head or tail on the next flip, and so on for the other flips.

We can easily demonstrate hypothesis testing in a binomial setting using R. For instance, let us return to the coin-flipping experiment. Suppose you would like to know the probability of obtaining two heads on five flips of a fair coin, where each flip is assumed to have a probability of heads equal to 0.5. In R, we can compute this as follows:

```
> dbinom(2, size = 5, prob = 0.5)
[1] 0.3125
```

where `dbinom` calls the "density for the binomial," "2" is the number of successes we are specifying, "size = 5" represents the number of trials we are taking, and "prob = 0.5" is the probability of success on any given trial, which recall is assumed constant from trial to trial.

Suppose instead of two heads, we were interested in the probability of obtaining five heads:

```
> dbinom(5, size = 5, prob = 0.5)
[1] 0.03125
```

Notice that the probability of obtaining five heads out of five flips on a fair coin is quite a bit less than that of obtaining two heads. We can continue to obtain the remaining probabilities and get the complete binomial distribution for this experiment:

Heads	**0**	**1**	**2**	**3**	**4**	**5**	
Prob	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125	$\sum 1.0$

A plot of this binomial distribution is given in Figure 3.4.

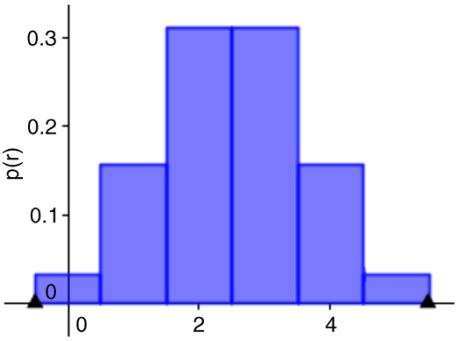


FIGURE 3.4 Binomial distribution for the probability of the number of heads on a fair coin.

Suppose that instead of wanting to know the probability of getting two heads out of five flips, we wanted to know the probability of getting two *or more* heads out of five flips. Because the events 2 heads, 3 heads, 4 heads, and 5 heads are *mutually exclusive events*, we can add their probabilities by the probability rule featured in Chapter 2 that said $p(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$: $0.3125 + 0.3125 + 0.15625 + 0.03125 = 0.8125$. Hence, the probability of obtaining two or more heads on a fair coin on five flips is equal to 0.8125.

Binomial distributions are useful in a great variety of contexts in modeling a wide number of phenomena. But again, remember that the outcome of the variable must be *binary*, meaning it must have only *two* possibilities. If it has more than two possibilities or is continuous in nature (see Section 2.9.8), then the binomial distribution is not suitable. Binomial data will be featured further in our discussion of logistic regression in Chapter 11.

One can also appreciate the general logic of hypothesis testing through the binomial. If our null hypothesis is that the coin is fair, and we obtain five heads out of five flips, this result has only a 0.03125 probability of occurring. Hence, because the probability of these data is so low under the model that the coin is fair, we decide to reject the null hypothesis and infer the *statistical alternative* hypothesis that $p(H) \neq .5$. Substantively, we might infer that the coin is not fair, though this *substantive alternative* also assumes it is the *coin* that is to "blame" for it coming up five times heads. If the flipper was responsible for biasing the coin, for instance, or a breeze suddenly came along that helped the result occur in this particular fashion, then inferring the *substantive* alternative hypothesis of "unfairness" may not be correct. Perhaps the *nature* of the coin is such that it *is* fair. Maybe the flipper or other factors (e.g., breeze) are what are ultimately responsible for the rejection of the null. This is one reason why rejecting null hypotheses is quite easy, but inferring the correct *substantive* alternative hypothesis (i.e., the hypothesis that explains *why* the null was rejected) is much more challenging (see Denis, 2001). As concluded by Denis, "Anyone can reject a null, to be sure. The real skill of the scientist is arriving at the true alternative."

The binomial distribution is also well-suited for comparing proportions. For details on how to run this simple test in R, see Crawley (2013, p. 365). One can also use `binom.test` in R to test simple binomial hypotheses, or the `prop.test` for testing null hypotheses about proportions. A useful test that employs binomial distributions is the *sign test* (see Siegel and Castellan, 1988, pp. 80–87 for details).

3.1.3 Normal Approximation

Many distributions in statistics can be regarded as *limiting forms* of other distributions. What this statement means can be best demonstrated through an example of how the binomial and normal distributions are related. When the number of discrete numbers along the x-axis grows larger and larger, the areas under the binomial distribution more and more resemble the probabilities computed under the normal curve. It is in this sense that for a large number of trials on the binomial, it begins to more closely *approximate* the normal distribution.

As an example, consider once again the binomial distribution for $n = 5, p = 0.5$, but this time with a normal density overlaying the binomial (Figure 3.5).

We can see that the normal curve "approximates" the binomial distribution, though perhaps not tremendously well for only five trials. If we increase the number of trials however, to say, 20, the approximation is much improved. And when we increase the number of trials to 100, the binomial distribution looks virtually like a normal density. That is, we say that *the normal distribution is the limiting form of the binomial distribution*.

We can express this idea more formally. If the number of trials n in a binomial experiment is made large, the distribution of the number of successes x will tend to resemble a normal distribution. That is, the normal distribution is the limiting form of a binomial distribution as $n \to \infty$ for a fixed p (and where $q = 1 - p$), where $E(x_i)$ is the expectation of the random variable x_i (the meaning of which will be discussed shortly):

$$z = \frac{x_i - \mu}{\sigma} \leftrightarrow z_m = \frac{x_i - E(x_i)}{\sigma} \leftrightarrow z_m = \frac{x_i - np}{\sqrt{npq}}$$

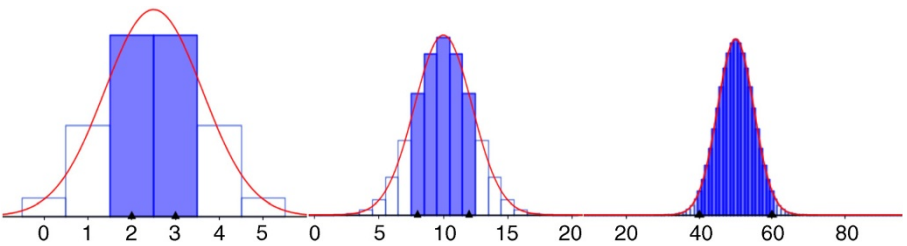


FIGURE 3.5 Binomial distributions approximated by normal densities for 5 (far left), 20 (middle), and 100 trials (far right).

Notice that in a z-score calculation using the population mean μ, in the numerator, we are actually calculating the difference between the obtained score and the *expectation*, $E(x_i)$. We can change this to a binomial function by replacing the expectation μ with the expectation from a binomial distribution, that is, np, where np is the mean of a binomial distribution. Similarly, we replace the standard deviation from a normal distribution with the standard deviation from the binomial distribution, $\sqrt{npq}$. As n grows infinitely large, the normal and the binomial probabilities become identical for any standardized interval.[3]

3.1.4 Joint Probability Densities: Bivariate and Multivariate Distributions

A univariate density expresses the probability of a single random variable within a specified interval of values along the abscissa. A joint probability density, analogous to a joint probability, expresses the probability of simultaneously observing *two* random variables over a given interval of values. The bivariate normal density is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

where ρ^2 is the squared Pearson correlation coefficient between x_1 and x_2.

When plotted, the bivariate density resembles a pile of raked leaves in the Autumn. A plot generated in R is given in Figure 3.6.

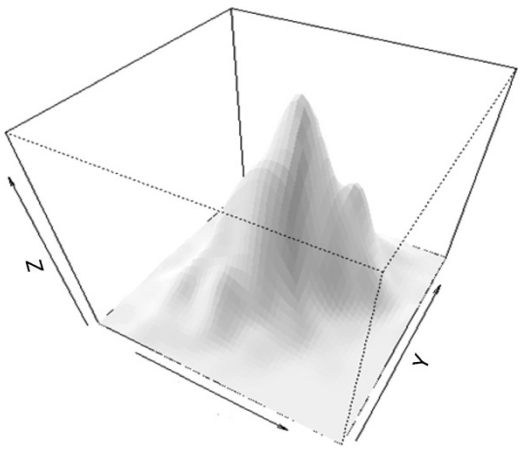
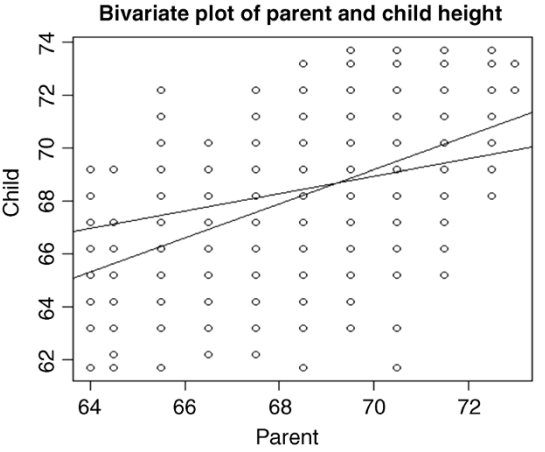


FIGURE 3.6 Bivariate density.[4]

[3]For a more technical demonstration of how and why this convergence occurs, see Proschan (2008).
[4]Code for this plot was retrieved from https://stat.ethz.ch/pipermail/r-help/2003-September/038314.html

Empirical bivariate distributions (as opposed to bivariate *densities*) are those showing the joint occurrence on two variables. For instance, again using Galton's data, we plot parent height by child height, in which we also fit both regression lines (see Chapter 8) using `lm`:

```
> plot(parent, child, main = "Bivariate Plot of Parent and Child
Height")
> abline(lm(parent~child))
> abline(lm(child~parent))
```



Note the relation between parent height and child height. Recall that a mathematical relation is a subset of the Cartesian product. The Cartesian product in the plot consists of *all* theoretically possible parent–child pairings. The fact that shorter than average parents tend to have shorter than average children and taller than average parents tend to have taller than average children reveals the linear form of the mathematical relation. In the plot are regression lines for child height as a function of parent height and parent height as a function of child height. Computing both the mean of child and of parent, we get

```
> mean(child)
[1] 68.08847
> mean(parent)
[1] 68.30819
```

Notice that both regression lines, as they are required to do whatever the empirical data, pass through the means of each variable. The reason for this will become clearer in Chapter 8.

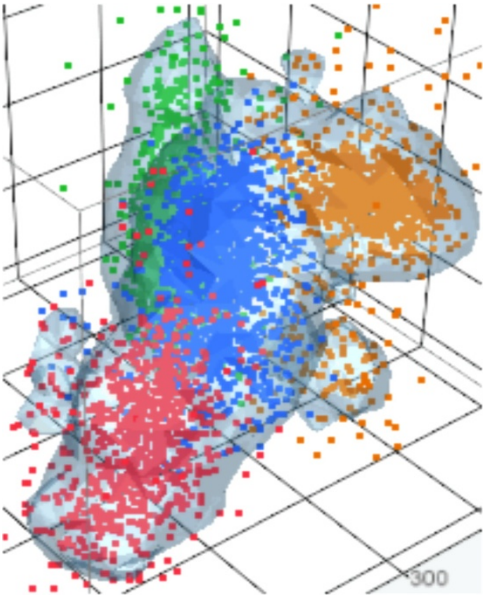


FIGURE 3.7 A 3D scatterplot with density contour and points.[5]

Turning now to multivariate distributions, the multivariate density is given by

$$g(x_i) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$$

where p is the number of variables and $|\Sigma|$ is the determinant of the population covariance matrix, which can be taken as a measure of *generalized variance* since it incorporates both variances *and* covariances. Multivariate distributions represent the joint occurrence of three or more variables, and thus are quite difficult to visualize. One way, however, of representing a density in three-dimensions is attempted in Figure 3.7.

Most multivariate procedures make some assumption regarding the multivariate normality of sampling distributions. Evaluating such an assumption is intrinsically difficult due to high dimensionality of the data. The best researchers can usually do is attempt to verify univariate and bivariate normality through such devices as histograms and scatterplots. Fortunately, as is the case for methods assuming univariate normality, multivariate procedures are relatively robust to violations. Though Mardia's test (Mardia, 1970) is favored by some (Romeu and Ozturk, 1993), no single method for evaluating multivariate normality appears to be fully adequate. Visual inspections of Q–Q plots (to be discussed) are usually sufficient for applied purposes.

[5]Image taken from http://www.jmp.com/support/help/Scatterplot_3D_Platform_Options.shtml

In cases where rather severe departures of normality exist, one may also choose to perform data transformations on the "offending" variables to better approximate normal distributions. However, it should be kept in mind that sometimes a severely nonnormal distribution can be evidence more of a *scientific* problem than symptomatic of a statistical issue. For example, if we asked individuals in a sample how many car accidents they got into this month, the vast majority of our responses would indicate a count of "0." Is the distribution skewed? Yes, but this is not a statistical problem, it is a *substantive* one. We would likely not even have sufficient variability in our measurement responses to conduct any meaningful analyses since probably close to 100% of our sample will respond with "0." If virtually everyone in your sample responds with a constant, then one might say the very process of *measurement* may have been a failure. The difficulties presented in subjecting that data to statistical analyses should be an afterthought, second in priority to the more pressing scientific issue.

3.2 CHI-SQUARE DISTRIBUTIONS AND GOODNESS-OF-FIT TEST

The chi-square distribution is given by

$$f(x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{[(v/2)-1]} e^{-x/2}$$

for $x > 0$, where v are degrees of freedom and Γ is the gamma function.[6] The chi-square distribution of a random variable is also equal to that of the sum of squares of n independent and normally distributed z-scores. That is,

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}$$

[6]For details on the gamma function, see Degroot and Schervish (2002, p. 295). A plot of the gamma function appears as follows (see Crawley, 2013, p. 264, for the R code):

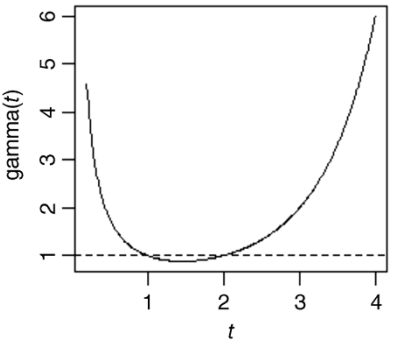


TABLE 3.1 Contingency Table for 2 × 2 Design

	Condition Present (1)	Condition Absent (0)	Total
Exposure yes (1)	20	10	30
Exposure no (2)	5	15	20
Total	25	25	50

The chi-square distribution plays an important role in mathematical statistics and is associated with a number of tests on model coefficients in a variety of statistical methods. The multivariate analog to the chi-square distribution is that of the *Wishart distribution*, not discussed here (see Rencher, 1998, p. 53).

The chi-square goodness-of-fit test is one such statistical method that uses the chi-square test statistic to evaluate the tenability of a null hypothesis. Recall that such a test is suitable for categorical data in which counts (i.e., instead of means, medians, etc.) are computed within each cell of the design. The goodness-of-fit test is given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_i - E_i)^2 / E_i$$

where O_i and E_i represent observed and expected frequencies, respectively, summed across r rows and c columns.

As a simple example, consider hypothetical data (Table 3.1), where the frequencies of those exposed to something adverse are related to whether a condition is absent or present. If you are a clinical psychologist, then you might define *exposure* as, perhaps, a variable such as combat exposure, and *condition* as posttraumatic stress disorder.

The null hypothesis is that the 50 counts making up the entire table are randomly distributed across each of the cells. We can easily test this hypothesis in SPSS by weighting the relevant frequencies by cell total:

Exposure	Condition	Frequency
1.00	0.00	10.00
1.00	1.00	20.00
2.00	0.00	15.00
2.00	1.00	5.00

```
WEIGHT BY Frequency.
CROSSTABS
  /TABLES=Condition BY Exposure
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

The output follows in which it is first confirmed that we set up our data file correctly:

Exposure * Condition Crosstabulation

Count

		Condition		Total
		1.00	0.00	
Exposure	1.00	20	10	30
	2.00	5	15	20
Total		25	25	50

We focus on the Pearson chi-square test value of 8.3 on a single degree of freedom. It is statistically significant ($p = 0.004$), and hence we can reject the null hypothesis of no association between condition and exposure group.

Chi-square Tests

	Value	df	Asymp. Sig. (two-sided)	Exact Sig. (two-sided)	Exact Sig. (one-sided)
Pearson chi-square	8.333[a]	1	0.004		
Continuity correction[b]	6.750	1	0.009		
Likelihood ratio	8.630	1	0.003		
Fisher's exact test				0.009	0.004
Linear-by-linear association	8.167	1	0.004		
No. of valid cases	50				

[a]0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.00.
[b]Computed only for a 2×2 table.

In addition to concluding an association between group and condition, we can state a lot more about these data. For instance, the odds are defined as

$$\frac{p}{1-p}$$

where p is the probability of a "success" and $1 - p$ is the probability of "failure." For our data, we can conclude the following:

- Given that you have been exposed (1), the odds of the condition being present are 20/30 to $10/30 = 2.00$.
- Given that you have not been exposed (2), the odds of the condition being present are 5/20 to $15/20 = 0.33$.
- Being exposed makes it $2/0.33 = 6$ times more likely for the condition to be present vs. absent than not being exposed. This is the *odds ratio* comparing odds for those

exposed to odds for those not exposed. We discuss odds ratios more extensively in Chapter 11.

The *relative risk* is computed as the ratio of p for *exposure yes* to p for *exposure no*. It is a measure of association to aid in assessing exposure to a condition and risk associated with a particular outcome. A relative risk of 1.0 indicates independence. For our data, it is equal to $0.66/0.25 = 2.64$. The number of 2.64 indicates that a person who has been exposed is 2.64 times as likely to have the condition present as someone who has not been exposed. The odds ratio can also be related to relative risk (see Agresti, 2002, p. 47).

In R, we can easily perform the chi-square test on these data. We first build the matrix of cell counts, calling it `diag.table`:

```
> diag.table <- matrix(c(20, 5, 10, 15), nrow = 2)
> diag.table
     [,1] [,2]
[1,]   20   10
[2,]    5   15

> chisq.test(diag.table, correct = F)

        Pearson's Chi-squared test

data:  diag.table
X-squared = 8.3333, df = 1, p-value = 0.003892
```

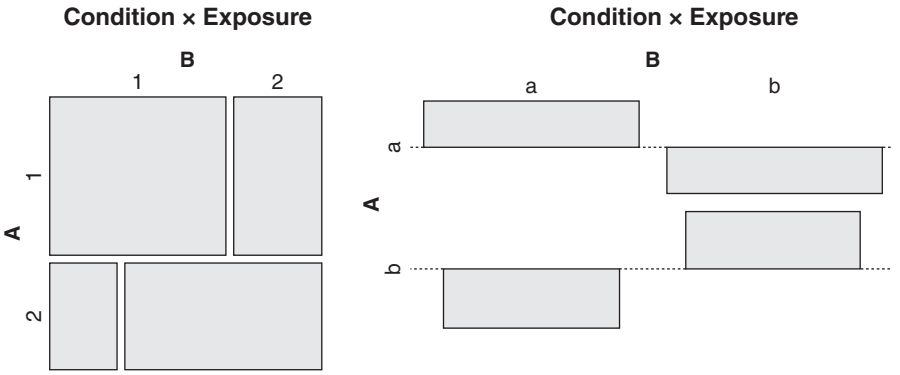
We see that the result in R agrees with what we obtained in SPSS. Note that specifying `correct = F` (correction = false) negated what is known as *Yates' correction for continuity*, which involves subtracting 0.5 from positive differences in $O - E$ and adding 0.5 to negative differences in $O - E$ in an attempt to better make the multinomial distribution approximate that of a chi-square distribution (i.e., in a loose sense, to help make discrete probabilities more continuous). To adjust for Yates, we can either specify `correct = T` or simply `chisq.test(diag.table)`, which will incorporate the correction. With the correction implemented, our p-value increases from 0.003 to 0.009 (not shown). We notice that this adjustment parallels that made in SPSS by adjusting for continuity. When expected counts per cell are relatively small (a working rule is that they should be at least five in each cell), one can also request *Fisher's exact test* (see Fisher, 1922a), which we note also mirrors the output generated by SPSS:

```
> fisher.test(diag.table)

        Fisher's Exact Test for Count Data

data: diag.table
```

```
p-value = 0.008579
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.466377 26.597383
sample estimates:
odds ratio
 5.764989
```

To visualize findings, we can produce what is known as a *mosaic plot* (a) and an *association plot* (b) for our analyzed frequency data:

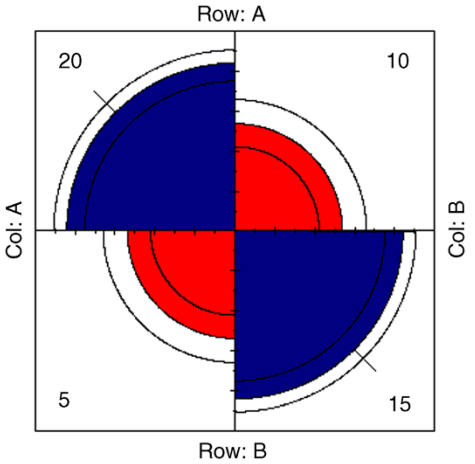
```
> library (vcd)
> mosaic(diag.table, main = "Condition x Exposure")
> assoc(diag.table, main = "Condition x Exposure")
```



The mosaic plot displays the relevant frequencies by associated areas in each rectangle proportional to their cell totals. For instance, the cell in row 1, column 1 is represented with much area since it comprises a substantial frequency of the column and row totals. The frequency of 15 is represented in the lower right of the mosaic (a), also denoting a substantial proportion. The association plot (b) communicates deviation from independence. In the case of complete independence, the plot would consist simply of two horizontal dotted lines. For the current data in which there is a lack of independence, cells with observed frequencies greater than expected are indicated by areas rising above the line (e.g., cells in row 1, column 1, and row 2, column 2), while cells with observed frequencies less than expected are indicated by areas falling below the line (e.g., cells in row 2, column 1 and row 1, column 2). For further details on these plots, see Friendly (1991, 2000).

One can also generate what is known as a *fourfold* plot, available in the `vcd` package in R (Meyer, Zeileis, and Hornik, 2014). Frequencies (and confidence rings, see Friendly, 2000, p. 75) are given in each quadrant, also reflected by their respective areas, with odds ratio equal to $20(15)/10(5) = 6$.

```
> library(vcd)
> fourfold(diag.table)
```



Other useful statistics for contingency tables include the *phi coefficient* and *Cramer's V*. Phi, ϕ is a measure of association for 2×2 contingency tables, computed as

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

where χ^2 is the chi-square statistic calculated on the 2×2 table and n is the total sample size. The maximum ϕ can attain is 1.0, indicating maximal association. ϕ can be computed in SPSS by/`statistics = phi` and is available in R in the `psych` package (Revelle, 2015). Cramer's ϕ_c extends on ϕ in that it allows for contingency tables of greater than 2×2. It is included in the /`statistics = phi` command and is also available in R's `psych` package. It is given by

$$\phi_c = \sqrt{\frac{\chi^2}{n(k-1)}}$$

where k is the minimum of the number of rows or columns. The relationship between ϕ_c and ϕ is easily shown for $k = 2$:

$$\phi_c = \sqrt{\frac{\chi^2}{n(2-1)}} = \sqrt{\frac{\chi^2}{n}} = \phi$$

3.2.1 Power for Chi-Square Test of Independence

We can estimate power[7] and required sample size for the chi-square test of independence using the package `pwr` (Champely, 2014) in R:

```
> library(pwr)
> pwr.chisq.test (w =, N =, df =, sig.level =, power =)
```

[7]Power will be discussed later in this chapter.

where w is the anticipated or required effect size, estimated as

$$w = \sqrt{\sum_{i=1}^{m} \frac{(p0_i - p1_i)^2}{p0_i}}$$

and $p0_i$ and $p1_i$ are the probabilities in a given cell i under the null and alternative hypotheses, respectively. We demonstrate by estimating power for $w = 0.2$:

```
> pwr.chisq.test(w = 0.2, N =, df = 5, sig.level = .05, power = 0.90)

     Chi squared power calculation

              w = 0.2
              N = 411.7366
             df = 5
      sig.level = 0.05
          power = 0.9
NOTE: N is the number of observations
```

R estimates that a total of approximately 411 subjects are required to achieve power set at 0.90. Such a large sample is required because $w = 0.2$ constitutes a relatively small effect size (see Cohen (1988) for details).

The reader may ask at this point how one might go about analyzing data for higher dimensional frequency tables. The example for the chi-square test of the data in Table 3.1 is only for that of a 2×2 layout. Suppose we added a third factor to our analysis, such as *gender*, making our contingency table appear as in Table 3.2.

For data such as that in Table 3.2 featuring higher-dimensional frequency data, *log-linear models* are appropriate (Agresti, 2002). Log-linear models are an option in the wider class of *generalized linear models*, to be discussed further in Chapter 11, where we discuss in some detail a special case of the generalized linear model called the *logistic regression model*.

TABLE 3.2 Contingency Table for $2 \times 2 \times 2$ Design

	Exposure	Condition Absent (0)	Condition Present (1)	Total
Males	Yes	10	20	30
	No	15	5	20
Females	Yes	13	17	30
	No	12	8	20
	Total	50	50	100

TABLE 3.3 Contingency Table for 2 × 2 Diagnostic Design

	Diagnosis Yes	Diagnosis No	Total
Disease Yes	20	10	30
Disease No	5	15	20
Total	25	25	50

3.3 SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are measures historically used in diagnostic situations but can be applied in other contexts as well. We can easily adapt the data in Table 3.1 to suit a brief discussion of these measures. We keep the same cell frequencies, but modify variable names so that the data become a bit more applicable to a discussion of sensitivity and specificity (see Table 3.3).

The *sensitivity* of the diagnostic instrument is the probability that the test is positive given that the individual has the disease. In the margins we see that 30 people have the disease, of which 20 were diagnosed with it. Thus, the sensitivity of the test is $20/30 = 0.66$. The *specificity* of the diagnostic instrument is the probability that the test is negative, given that the individual does not have the disease. In the margins we see that 20 people do not have the disease, of which 15 were diagnosed with not having the disease. Hence, the specificity of the test is $15/20 = 0.75$. The overall *prevalence* of the disease is equal to 30/50 (i.e., 30 people have the disease out of 50). One can also compute what are known as positive and negative *predictive values* from such tables. For these and other measures useful for diagnostic situations, see Dawson and Trapp (2004).

3.4 SCALES OF MEASUREMENT: NOMINAL, ORDINAL, AND INTERVAL, RATIO

Recall that in our discussion of the so-called "soft" versus "hard" sciences in Chapter 1, we concluded that a key principal difference between the two does not necessarily lie in different statistical or analytical methods used in drawing conclusions, but rather in the actual *material* that is subjected to measurement. Though this book is not about measurement *per se*, we nonetheless wish to review the scales of measurement as first proposed by Stevens in 1946 (Stevens, 1946).

Before we discuss these scales, it would do well to remind ourselves just what is measurement in the first place. We propose the following workable definition:

> Measurement is the systematic assignment of numbers to observations according to a well-defined set of rules.

The job of the "rules" is to make good sense of the measurement process. For instance, if we simply assigned numbers to observations without a set of rules to govern the assignment, then even if I weigh more than you, I could be assigned 150 lb and you 180 lb. The requirement of having rules of measurement avoids such meaningless and contradictory assignments. If I weigh more than you, rules of measurement imply that my weight measurement will be larger than yours within the margin of measurement error.

3.4.1 Nominal Scale

Measurement at the *nominal* level is hardly considered *real* measurement, because it is simply the process of grouping objects or subjects into *classes*. Each class is usually represented by a number, letter, name, etc. Other than naming these categories, no other properties are assumed or inferred, such as distance between objects or magnitude.

A classic example of measurement at the nominal level is that of hockey jersey numbers. That the number "99" is greater than the number "22" on the shirts of two hockey players does not imply anything about magnitude (though Wayne Gretzky did in this case wear "99" and was perhaps the best hockey player ever). The numbers 99 and 22 are simply "classes," they are symbols used to identify (or *name*) one class as different or distinct from the other. The fact that we use a rational system such as the real numbers to identify these different classes of "99" versus "22" does not imply anything about order or magnitude *at the level of substantive measurement*. Yes, to the mathematician, 99 is indeed numerically greater than 22. That is, an order property is implied in the *numbers*. However, to the scientist, nothing of order or magnitude needs to be implied when working with a nominal scale.

To briefly elaborate on this point, the concept of using numbers to represent classes makes for an ideal example of the distinction between mathematical measurement versus scientific measurement. In the mathematical measurement of the distance on the real line (e.g., the "length" between two real numbers), order is a necessary implication and differentiates any two numbers on the line. In scientific measurement, though we may still use the "objects" (i.e., the numbers) of pure mathematics, whether there exist order or magnitude in our *empirical* objects of study is for us to decide as scientists with the aid of our measurement tools. It is not solely a mathematical or "abstract" consideration.

As an example, consider the following objects:

* $ # %

Though we can say, at minimum, that nominal level measurement has been achieved (the objects have different symbols, that is, different *names*), we cannot say anything more about either the distance or magnitude between the objects, *unless we decide to impose an order relation* on the above objects. For instance, if we decide, based on our rules of measurement, that $ is greater than *, then not only have we measurement at the nominal level, we also have measurement at the *ordinal* level.

3.4.2 Ordinal Scale

In addition to categorizing objects into classes, measurement at the *ordinal* level imposes an *order relation* between objects. For instance, if \$ is greater than * for some characteristic that these symbols represent, then we have measurement at the ordinal level. The imposition of an order relation is fundamental to any sort of true measurement. Consider that if your measurement system does not even allow you to say that one thing has *more* of a characteristic than another, what could be the purpose of even measuring?

Ordinal measurement, however, does not say anything about the *precise* amount of magnitude between objects. For example, first place, second place, and third place in a race constitute measurement at the ordinal level, but that you finished second does not immediately tell us the *distance* between first and second, or the *distance* between second and third. To speak of distances between objects, we require measurement at the *interval* level.

3.4.3 Interval Scale

Measurement at the *interval* level possesses all the features of measurement of both nominal and ordinal scales, but with the extra requirement that *distances between measured objects are quantifiable, and that distances between successive measuring points on the scale are equal in magnitude*. For instance, consider the measurement of temperature in degrees Fahrenheit. The change in temperature from 10 to 20 degrees contains the same "amount" of temperature change as that from 20 to 30 degrees. That is, the *intervals* between measurement points are meaningful and represent an equal distance in the "thing" (i.e., temperature, in this case) we are measuring.

Is intelligence measurable on an interval scale? What would it mean for it to be measurable at the interval level? Well, supposing we base our measurements on a reputable standardized test, for IQ to be measurable at the interval scale would imply that the *distance* in the thing called "IQ" is equivalent from, say, 90 to 100 as it is from 100 to 110. At first glance, this might appear an easy condition to satisfy, after all, the real number distance in each interval is equal to 10. However, recall that *that is a distance of real numbers, not necessarily of IQ*. As William James put it, *we must not confuse the phenomena we study with the abstractions we use to study them*. The real numbers are the abstraction. The IQ is the phenomenon. That we used a real line to measure these distances does not necessarily imply that the actual *true* distances in terms of "IQ substance" corresponds one-to-one (or even at all) to our measurement tool. It is entirely possible that 90 to 100 represents a greater increase in IQ than does 100 to 110, making the relation between our measurement of IQ versus "true IQ" *nonlinear*. Our measurement of IQ is simply not that precise to make such statements. *Numerical length* in this case may not translate to the *substantive length* of the difference under study.

3.4.4 Ratio Scale

The most sophisticated scale of measurement is that of the *ratio scale*. It is the most sophisticated because it is the only scale for which we can speak meaningfully about

ratios between competing measurement intervals. By "ratio," we simply mean we have the power to make such statements as "object *a* is twice as large as object *b*." Up to now, no other scale has allowed us to make such statements. For instance, in the interval scale, concluding that *a* is any factor greater than *b* made no sense. We did not have a starting point to base such conclusions. An IQ of zero did not necessarily mean the *absence* of intelligence. Rather, it was simply an arbitrary point on the IQ scale presumably denoting a particular quantity of IQ (even if, in all probability, very small).

What gives us license to make statements of ratios? The element of the ratio scale that permits us to make such statements is the fact that the ratio scale has at its origin a *true zero point*. When something is deemed measurable at the ratio level, a measurement of zero *actually means zero of the thing that is being measured*. Was this fact true of the interval scale? No, because zero degrees Fahrenheit did not equate to there being *zero* temperature. "Zero" was simply an arbitrary value on the scale. However, the fact that I have zero coins in my pocket actually means that I have *zero* coins. "Zero" is said to be, in this case, "absolute," meaning that there is truly nothing there.

Physical quantities such as weight, distance, velocity, and motion, are all measurable at the ratio level. Variables such as reaction time in sensation experiments are also measurable at the ratio level. Phenomena such as intelligence, anxiety, and attitude are generally not. More often we deem them measurable at the interval level or less, and when we *really* get critical, it is even a stretch at times to consider the ordinal level of measurement as being satisfied for such variables. Then again, if we decided to operationally define anxiety by *beats per minute* of one's heart, then theoretically at least one could conclude that an individual has *zero* anxiety if that individual has zero beats per minute (though of course this could make for an awkward definition for the absence of anxiety!).

3.5 MATHEMATICAL VARIABLES VERSUS RANDOM VARIABLES

When we speak of a *mathematical variable* (or simply, *variable*), we mean a symbol that at any point could be replaced by values contained in a specified set. For instance, consider the mathematical variable y_i. By the subscript i is indicated the fact that y_i stands for a *set* of values, not all equal to the same number (otherwise y would be a *constant*) such that at any point in time any of these values in the set could serve as a temporary "replacement" for the symbol.

Of course, social and natural sciences are all about variables. Here are some examples:

- Height of persons in the world is a variable because persons of the world have different heights. However, height would be considered a *constant* if 10 people in a room were of the exact same height (and those were the only people we were considering).
- Blood pressure is a variable because persons, animals, and other living creatures have different blood pressure measurements.

TABLE 3.4 Mathematical versus Discrete Random Variable

Mathematical Variable y_i	Random Variable y_i
$y_1 = 1$	$y_1 = 1$ $(p = 0.20)$
$y_2 = 3$	$y_2 = 3$ $(p = 0.50)$
$y_3 = 5$	$y_3 = 5$ $(p = 0.30)$

- Intelligence (IQ) of human beings (difficult to measure to be sure, although psychology has developed instruments in an attempt to assess such things) is a variable because people have differing intellectual capacities.
- Earned run average (ERA) of baseball players is a variable because all players do not have the same ERA.

A *random variable* is a mathematical variable that is associated with a probability distribution. That is, as soon as we assign probabilities to *values* of the variable, we have a random variable. More formally, we can say that a random variable is a *function from a sample space into the real numbers* (Casella and Berger, 2002), which essentially means that elements in the set (i.e., sample space) have probabilities associated with them (Dowdy, Wearden, and Chilko, 2004).

Consider a simple comparison between a mathematical variable and a discrete random variable in Table 3.4.

Notice that for the mathematical variable, probability does not enter the picture, it is not of any consideration. For the discrete random variable, each value of the variable has a probability associated with it. Note as well that the probabilities must sum to 1.0 for it to be a legitimate probability distribution (i.e., $0.20 + 0.50 + 0.30 = 1.0$). How the given probabilities are assigned is a matter to be governed by the specific context of the problem. Recall as well from Chapter 2 that variables can be classified as discrete or continuous. This same distinction can be applied to random variables as to ordinary mathematical variables. Table 3.4 features a discrete random variable. For continuous random variables, since the probability of any particular value in a continuous distribution is theoretically zero, instead of associating probabilities with particular values, probabilities are associated with areas under the curve computed by way of integration in calculus.

The distinction between mathematical and random variables is important when we discuss such things as means, variances, and covariances. A reader first learning about random variables, having already mastered the concept of sample or population variance (to be discussed shortly), can be somewhat taken aback when encountering the variance of a random variable, given as

$$\sigma^2 = E(y_i - \mu)^2$$

and then attempting to compare it with the more familiar variance of a population:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n}$$

Realize however that both expressions are essentially similar, they both account for squared deviations from the mean. However, the variance of a random variable is stated in terms of its *expectation*, E. Throughout this book, we will see the operator E at work. What is an expectation? The expectation E of a random variable is the *mean* of that random variable, which amounts to it being a probability-weighted average (Gill, 2006). The operator E occurs several times throughout this book because in theoretical statistics, *long-run averages* of a statistic are of especial interest. As noted by Feller (1968, p. 221), should an experiment be repeated n times under identical conditions, the average of such trials should be *close to* expectation. Perhaps less formally, the operator E then tells us what we might expect to see in the *long run* for large n. Theoretical statisticians love taking expectations, because the short run of a variable is seldom of interest at a theoretical level. It is the long (probability) run that is often of most theoretical interest. As a crude analogy, on a personal level, you may be "up" or "down" now, but if your expectation E pointed to a favorable long-run endpoint, then perhaps that is enough to convince you that though "on the way" is a rough tumbly road, in the end, as the spiritual would say, we "arrive" at our expectation (which perhaps some would denote as an afterlife of sorts).

The key point is that when we are working with expectations, we are working with *probabilities*. Thus, instead of summing squared deviations of the kind $\sum_{i=1}^{n}(y_i-\mu)^2$ as one does in the sample or population variance for which there is specified n, one must rather assign to these squared deviations *probabilities*, which is what is essentially being communicated by the notation "$E(y_i-\mu)^2$." We can "unpack" this expression to read

$$\sum p(y_i)(y_i-\mu)^2$$

where $p(y_i)$ is the probability of the given deviation, $(y_i-\mu)$, for in this case, a discrete random variable.

3.6 MOMENTS AND EXPECTATIONS

When we speak of *moments* of a distribution or of a random variable, we are referring to such things as the mean, variance, skewness, and kurtosis.

The first moment of a distribution is its mean. For a discrete random variable y_i, the expectation is given by

$$E(y_i)=\sum_{i=1}^{n} y_i p(y_i)$$

where y_i is the given value of the variable and $p(y_i)$ is its associated probability. When y_i is a continuous random variable, the expectation is given by

$$E(y_i)=\int_{-\infty}^{\infty} y_i p(y_i)\mathrm{d}y$$

FIGURE 3.8 Because the sum of deviations about the arithmetic mean is always zero, it can be conceptualized as a balance point on a scale.

Notice again that in both cases, whether the variable is discrete or continuous, we are simply summing products of values of the variable with its probability, or *density* if the variable is continuous. In the case of the discrete variable, the products are "explicit" in that our notation tells us to take each value of y (i.e., y_i) and multiply by the probability of that given value, $p(y_i)$. In the case of a continuous variable, the products are a bit more *implicit* one might say, since the "probability" of any *particular* value in a continuous density is equal to 0. Hence, the product $y_i p(y_i)$ is equal to the given value of y_i multiplied by its corresponding density.

The arithmetic mean is a point such that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$. That is, the sum of deviations around the mean is always equal to 0 for any data set we may consider. In this sense, we say that the arithmetic mean is the *center of gravity* of a distribution, it is the point that "balances" the distribution (see Figure 3.8).

3.6.1 Sample and Population Mean Vectors

We often wish to analyze data simultaneously on several response variables. For this, we require vector and matrix notation to express our responses. The matrix operations presented here are a direct extension of those introduced in Chapter 2 and surveyed more comprehensively in Appendix A and in any book on elementary matrix algebra.

Consider the following vector:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{pmatrix}$$

where y_1 is observation 1 up to observation y_n.

We can write the sample mean vector $\bar{\mathbf{y}}$ for several variables y_1 through y_p as

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \vdots \\ \vdots \\ \bar{y}_p \end{pmatrix}$$

where $\bar{y}_p$ is the mean of the pth variable.

The expectation of individual observations within each vector is equal to the population mean μ, of which the expectation of the sample vector $\mathbf{y}$ is equal to the population vector $\boldsymbol{\mu}$. This is simply an extension of scalar algebra to that of matrices:

$$E(\mathbf{y}) = E\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ \vdots \\ \vdots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}$$

Likewise, the expectations of individual sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p$ are equal to their population counterparts $\mu_1, \mu_2, \ldots, \mu_p$. The expectation of the sample mean vector $\bar{\mathbf{y}}$ is equal to the population mean vector $\boldsymbol{\mu}$:

$$E(\bar{\mathbf{y}}) = E\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \vdots \\ \vdots \\ \bar{y}_p \end{pmatrix} = \begin{pmatrix} E(\bar{y}_1) \\ E(\bar{y}_2) \\ \vdots \\ \vdots \\ \vdots \\ E(\bar{y}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

We also note that $\bar{\mathbf{y}}$ is an *unbiased* estimator of $\boldsymbol{\mu}$ since $E(\bar{\mathbf{y}}) = \boldsymbol{\mu}$.[8]

Recall that we said that the mean is the first moment of a distribution. We will discuss the second moment of a distribution, that of the *variance*, shortly. Before we do so, a brief discussion of estimation is required.

[8]This is a result analogous to the simpler case where the expectation of the sample mean is equal to the population mean. That is, $E(\bar{y}) = \mu$. Only now, we are generalizing this to vectors.

3.7 ESTIMATION AND ESTIMATORS

The goal of statistical inference is, in general, to estimate parameters of a population. We distinguish between point estimators and interval estimators. A *point estimator* is a function of a sample and is used to estimate a parameter in the population. Because estimates generated by estimators will vary from sample to sample, and will thus have a probability distribution associated with them, estimators are also often *random variables*. For example, the sample mean $\bar{y}$ is an estimator of the population mean μ. However, if we sample a bunch of $\bar{y}$ from a population for which μ is the actual population mean, we know, from both experience and statistical theory, that $\bar{y}$ will vary from sample to sample. This is why the estimator $\bar{y}$ is often a random variable, because each of its values will have associated with them a given probability (density) of occurrence. When we use the estimator to obtain a particular number, that number is known as an *estimate*. An *interval estimator* provides a range of values within which the true parameter is hypothesized to exist within some probability. A popular interval estimator is that of the *confidence interval*, a topic we discuss later in this chapter.

More generally, if T is some statistic, then we can use T as an estimator of a population parameter θ. Whether the estimator T is any *good* depends on several criteria, which we survey now.

On average, in the long run, the statistic T is considered to be an *unbiased estimator* of θ if

$$E(T) = \theta$$

That is, an estimator is considered unbiased if its expected value is equal to that of the parameter it is seeking to estimate. The *bias* of an estimator is measured by how much $E(T)$ deviates from θ. When an estimator is biased, then $E(T) \neq \theta$, or, we can say $E(T) - \theta \neq 0$. Since the bias will be a positive number, we can express this last statement as $E(T) - \theta > 0$.

Good estimators are, in general, unbiased. The most popular example of an unbiased estimator is that of the arithmetic sample mean since it can be shown that

$$E(\bar{y}) = \mu$$

An example of an estimator that is biased is the uncorrected sample variance, as we will soon discuss, since it can be shown that

$$E(S^2) \neq \sigma^2$$

However, S^2 is not *asymptotically* biased. As sample size increases without bound, $E(S^2)$ converges to σ^2. Once the sample variance is corrected, it leads to an unbiased estimator, even for smaller samples:

$$E(s^2) = \sigma^2$$

where now

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

An alternative, but perhaps slightly less immediately insightful definition of biasedness is that given by Anderson (2003) in which a test is considered unbiased if, in general, power achieves its *minimum* at the null hypothesis. In other words, an unbiased test is most powerful for detecting deviations from the null hypothesis (which is usually the purpose of our investigations) rather than detecting parameters under the null hypothesis.

Consistency[9] of an estimator means that as sample size increases indefinitely, the variance of the estimator approaches zero. That is, $\sigma_T^2 \to 0$ as $n \to \infty$. We could also write this using a limit concept:

$$\lim_{n \to \infty} \sigma_T^2 = 0$$

which reads "the variance of the estimator T as sample size n goes to infinity (grows without bound) is equal to 0." Fisher called this the *criterion of consistency*, informally defining it as "when applied to the whole population the derived statistic should be equal to the parameter" (Fisher, 1922, p. 316). The key to Fisher's definition is *whole population*, which means, theoretically at least, an infinitely large sample, or analogously, $n \to \infty$. More pragmatically, $\sigma_T^2 \to 0$ when we have the entire population.

An estimator is regarded as *efficient* the lower is its mean squared error. Estimators with lower variance are more efficient than estimators with higher variance. Fisher called this the *criterion of efficiency*, writing "when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error" (Fisher, 1922, p. 316). Efficient estimators are generally preferred over less efficient ones.

An estimator is regarded as *sufficient* for a given parameter if the statistic "captures" everything we need to know about the parameter and our knowledge of the parameter could not be improved if we considered additional information (such as a secondary statistic) over and above the sufficient estimator. As Fisher (1922, p. 316) described it, "the statistic chosen should summarize the whole of the relevant information supplied by the sample." More specifically, Fisher went on to say:

> If θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of distribution of pairs of values of θ_1 and θ_2, for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2 does not involve θ. In other words, when θ_1 is known, knowledge of the value of θ_2 throws no further light upon the value of θ. (Fisher, 1922, pp. 316–317)

[9]Though in this text we define consistency of an estimator quite simply, further distinctions exist between *weak* and *strong* consistency. See Shao (2003, pp. 132–133).

3.8 VARIANCE

Returning to our discussion of moments, the *variance* is the second moment of a distribution. For the discrete case, variance is defined as

$$\sigma^2 = \sum_{i=1}^{n} [(y_i - E(y_i)]^2 p(y_i)$$

while for the continuous case,

$$\sigma^2 = \int_{-\infty}^{\infty} [(y_i - E(y_i)]^2 p(y_i) \mathrm{d}y$$

Since $E(y_i) = \mu$, it stands that we may also write $E(y_i)$ as μ. We can also express σ^2 as $E(y_i^2) - \mu^2$ since, when we distribute expectations, we obtain

$$\begin{aligned}
\sigma^2 &= E(y_i - \mu)^2 \\
&= E(y_i - \mu)(y_i - \mu) \\
&= E(y_i^2 - y_i\mu - y_i\mu + \mu^2) \\
&= E(y_i^2) - E(y_i)\mu - E(y_i)\mu + \mu^2 \\
&= E(y_i^2) - \mu\mu - \mu\mu + \mu^2 \\
&= E(y_i^2) - \mu^2 - \mu^2 + \mu^2 \\
&= E(y_i^2) - \mu^2
\end{aligned}$$

Recall that the uncorrected and *biased* sample variance is given by

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}$$

As earlier noted, taking the expectation of S^2, we find that $E(S^2) \neq \sigma^2$. The actual expectation of S^2 is equal to

$$E(S^2) = [(n-1)/n]\sigma^2$$

which implies the degree to which S^2 is biased is equal to

$$\frac{-\sigma^2}{n}$$

We have said that S^2 is biased, but you may have noticed that as n increases, $n-1/n$ approaches 1, and so $E(S^2)$ will equal σ^2 as n increases without bound. This was our basis for earlier writing $\lim_{n \to \infty} E(S^2) = \sigma^2$. That is, we say that the estimator S^2, though biased for small samples, is *asymptotically unbiased* because its expectation is equal to σ^2 as $n \to \infty$.

When we lose a degree of freedom in the denominator and rename S^2 to s^2, we get

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

Recall that when we take the expectation of s^2, we find that $E(s^2) = \sigma^2$ (see Wackerly, Mendenhall, and Scheaffer (2002, pp. 372–373) for a proof).

The population *standard deviation* of σ^2 is given by the positive square root of σ^2, that is, $\sqrt{\sigma^2} = \sigma$. Analogously, the sample standard deviation is given by $\sqrt{s^2} = s$.

Recall the interpretation of a standard deviation. It tells us on average how much scores deviate from the mean. In computing a measure of dispersion, we initially squared deviations so as to avoid our measure of dispersion always equaling zero for any given set of observations, since the sum of deviations about the mean is always equal to 0. Taking the average of this sum of squares gave us the variance, but since this is in squared units, we wish to return them to "unsquared" units. This is how the standard deviation comes about. Studying the analysis of variance, the topic of the following chapter, will help in "cementing" some of these ideas of variance and the squaring of deviations, since ANOVA is all about generating different sums of squares and their averages, which go by the name of *mean squares*.

The variance and standard deviation are easily obtained in R. We compute for parent in Galton's data:

```
> var(parent)
[1] 3.194561

> sd(parent)
[1] 1.787333
```

One may also wish to compute what is known as the *coefficient of variation*, which is a ratio of the standard deviation to the mean. We can estimate this coefficient for `parent` and child, respectively, in Galton's data:

```
> cv.parent <- sd(parent)/mean(parent)
> cv.parent

[1] 0.02616573

> cv.child <- sd(child)/mean(child)
> cv.child

[1] 0.03698044
```

Computing the coefficient of variation is a way of comparing the variability of competing distributions relative to each distribution's mean. We can see that the

dispersion of child relative to its mean (0.037) is slightly larger than that of the dispersion of parent relative to its mean (0.026).

3.9 DEGREES OF FREEDOM

In our discussion of variance, we saw that if we wanted to use the sample variance as an estimator of the population variance, we needed to subtract 1 from the denominator. That is, S^2 was "corrected" into s^2:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

We say we *lost a degree of freedom* in the denominator of the statistic. But what are degrees of freedom? They are *the number of independent units of information in a sample that are relevant to the estimation of some parameter* (Everitt, 2002). In the case of the sample variance, s^2, one degree of freedom is lost since we are interested in using s^2 as an estimator of σ^2. We are losing the degree of freedom because the numerator, $\sum_{i=1}^{n} (y_i - \overline{y})^2$, is not based on n independent pieces of information since μ had to be estimated by $\overline{y}$. Hence, a degree of freedom is lost. Why? Because values of y_i are not independent of what $\overline{y}$ is, since $\overline{y}$ is fixed in terms of the given sample data.

A conceptual demonstration may prove useful in understanding the concept of degrees of freedom. Imagine you were asked to build a triangle such that there was to be no overlap of lines on either side of the triangle. In other words, the lengths of the sides had to join neatly at the vertices. We shall call this the "Beautiful Triangle" as depicted in Figure 3.9. You are now asked to draw the first side of the triangle. Why did you draw this first side the length that you did? You concede that the length of the first side is arbitrary, you were *free* to draw it whatever length you wished. In drawing the second length, you acknowledge you were also *free* to draw it whatever length you wished. Neither of the first two lengths in any way violated the construction of a beautiful triangle with perfectly adjoining vertices.

However, in drawing the third length, what length did you choose? Notice that to complete the triangle, you were not *free* to determine this length arbitrarily. Rather,

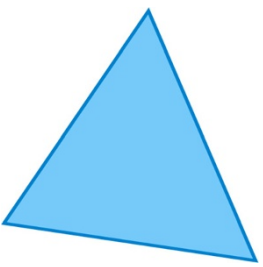


FIGURE 3.9 The "Beautiful Triangle" as a way to understanding degrees of freedom.

the length was *fixed* given the constraint that the triangle was to be a beautiful one. In summary then, in building the beautiful triangle, you lost 1 degree of freedom, in that two of the lengths were of your free choosing, but the third was fixed. Analogously, in using s^2 as an estimator of σ^2, a single degree of freedom is lost. If $\bar{y}$ is equal to 10, for instance, and the sample is based on five observations, then y_1, y_2, y_3, y_4 are freely chosen, but the fifth data point y_5 is *not* freely chosen so long as the mean must equal 10. The fifth data point is fixed. We lost a single degree of freedom.

Degrees of freedom occur throughout statistics in a variety of statistical tests. If you understand this basic example, then while working out degrees of freedom for more advanced designs and tests may still pose a challenge, you will nonetheless have a conceptual base from which to build your comprehension.

3.10 SKEWNESS AND KURTOSIS

The third moment of a distribution is its *skewness*. Skewness of a random variable generally refers to the extent to which a distribution lacks symmetry. Skewness is defined as

$$\gamma = E\left[\left(\frac{y_i - \mu}{\sigma}\right)^3\right] = \frac{E[(y_i - \mu)^3]}{(E[(y_i - \mu)^2])^{3/2}}$$

- Skewness for a normal distribution is equal to 0, just as skewness for a rectangular distribution is also equal to 0 (one does not require a *bell-shaped* curve for skewness to equal 0).
- Skewness for a *positively* skewed distribution is greater than 0; these distributions have tails that stretch out into values on the abscissa of greatest value.
- Skewness for a *negatively* skewed distribution is less than 0; these distributions have tails that stretch out to values on the abscissa of least value.

An example of a positively skewed distribution is that of the typical F density, given in Figure 3.10.

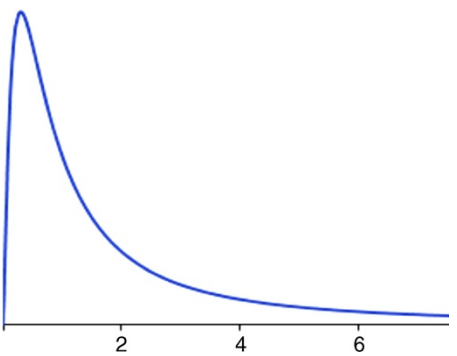


FIGURE 3.10 F distribution on 2 and 5 degrees of freedom. It is positively skewed since the tail stretches out to numbers of greater value.

The fourth moment of a distribution is its *kurtosis*, generally referring to the peakness of a distribution:

$$k = \frac{E[(y_i - \mu)^4]}{(E[(y_i - \mu)^2])^2}$$

With regard to kurtosis, distributions are defined:

- *mesokurtic* if the distribution exhibits kurtosis typical of a bell-shaped curve
- *platykurtic* if the distribution exhibits kurtosis more "plump" in the tails and flatter in the center than a normal distribution
- *leptokurtic* if the distribution exhibits kurtosis less "plump" in the tails and more narrow in the center than a normal distribution

We can easily compute moments of empirical distributions in R or SPSS. Several packages in R are available for this purpose. We could compute skewness for parent on Galton's data by

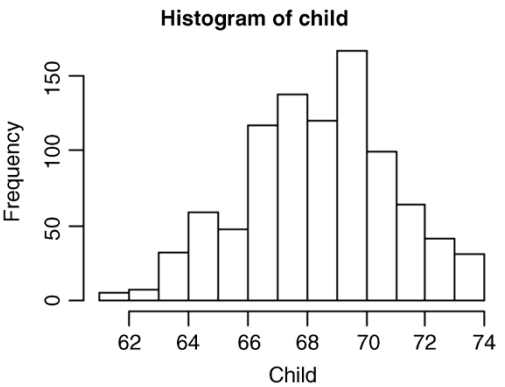
```
> library(psych)
> skew(parent)
[1] -0.03503614
```

The `psych` package (Revelle, 2015) also provides a range of descriptive statistics:

```
> library(psych)
> describe(Galton)
       vars   n  mean   sd median trimmed  mad  min  max range  skew kurtosis
parent    1 928 68.31 1.79   68.5   68.32 1.48 64.0 73.0     9 -0.04     0.05
child     2 928 68.09 2.52   68.2   68.12 2.97 61.7 73.7    12 -0.09    -0.35
         se
parent 0.06
child  0.08
```

The skew for child has a value of −0.09, indicating a *slight* negative skew. This is confirmed by visualizing the distribution (and by a relatively close inspection in order to spot the skewness):

```
> hist(child)
```



3.11 SAMPLING DISTRIBUTIONS

Sampling distributions are at the cornerstone of statistical inference. The sampling distribution of a statistic is a *theoretical probability distribution of that statistic*. As defined by Degroot and Schervish (2002), "the sampling distribution of a statistic tells us what values a statistic is likely to assume and how likely it is to assume those values prior to observing our data" (p. 391).

As an example, we will generate a theoretical sampling distribution of the mean for a given population with mean μ and variance σ^2. The distribution we will create is entirely *idealized* in that it does not exist in nature anywhere. It is simply a statistical *theory* of how the distribution of means might look if we were able to take an infinite number of samples from a given population, and on each of these samples, calculate the sample mean statistic.

When we derive sampling distributions for a statistic, we ask the following question:

> If we were to draw an infinite number of samples of size n from this population and calculate the sample mean on each sample, what would the distribution of sample means look like?

If we can specify this distribution, then we can evaluate obtained sample means *relative* to it. That is, we will be able to compare our obtained means (i.e., the ones we obtain in real empirical research) to the theoretical sampling distribution of means, and answer the question:

> If my obtained sample mean really did come from this population, what is the probability of obtaining a mean such as this?

If the probability is low, you might then decide to reject the assumption that the sample mean you obtained arose from the population in question. It could have, to be sure, but it *probably* did not. For continuous measures, our interpretation above is slightly informal, since the probability of any *particular* value of the sample mean in a continuous distribution is equal to 0. Hence, the question is usually posed such that we seek to know the probability of obtaining a mean such as the one we obtained *or more extreme*.

3.11.1 Sampling Distribution of the Mean

Since we regularly calculate and analyze sample means in our data, we are often interested in the sampling distribution of the mean. If we regularly computed medians, we would be equally as interested in the sampling distribution of the median.

Recall that when we consider any distribution, whether theoretical or empirical, we are usually especially interested in knowing two things about that distribution: a measure of central tendency and a measure of dispersion or variability. Why do we want to know such things? We want to know these two things because they help

summarize our observations, so that instead of looking at each individual data point to get an adequate description of the objects under study, we can simply request the mean and standard deviation as telling the story (albeit an incomplete one) of the obtained observations. Similarly, when we derive a sampling distribution, we are interested in the mean and standard deviation of that theoretical distribution of a statistic.

We already know how to calculate means and standard deviations for real empirical distributions. However, we do not know how to calculate means and standard deviations for sampling distributions. It seems reasonable that the mean and standard deviation of a sampling distribution should depend in some way on the given population from which we are sampling. For instance, if we are sampling from a population that has a mean $\mu = 20.0$ and population standard deviation $\sigma = 5$, it seems plausible that the sampling distribution of the mean should look different than if we were sampling from a population with $\mu = 10.0$ and $\sigma = 2$. It makes sense that different populations should give rise to different theoretical sampling distributions.

What we need then is a way to specify the sampling distribution of the mean for a given population. That is, if we draw sample means from this population, what does the sampling distribution of the mean look like for this population? To answer this question, we need both the expectation of the sampling distribution (i.e., its mean) as well as the standard deviation of the sampling distribution (i.e., its standard error (SE)). We know that the expectation of the sample mean $\bar{y}$ is equal to the population mean μ. That is, $E(\bar{y}) = \mu$. For example, for a sample mean $\bar{y} = 20.0$, the expected value of the sample mean is equal to the population mean, μ of 20.0.

To understand why $E(\bar{y}) = \mu$ should be true, consider first how the sample mean is defined:

$$\bar{y} = \frac{(y_1 + y_2 + \cdots + y_n)}{n}$$

Incorporating this into the expectation for $\bar{y}$, we have

$$E(\bar{y}) = E\left(\frac{(y_1 + y_2 + \cdots + y_n)}{n}\right)$$

There is a rule of expectations that says *the expectation of the sum of random variables is equal to the sum of individual expectations*. This being the case, we can write the expectation of the sample mean $\bar{y}$ as

$$\begin{aligned} E(\bar{y}) &= \frac{E(y_1 + y_2 + \cdots + y_n)}{n} \\ &= \frac{[E(y_1) + E(y_2) + \cdots + E(y_n)]}{n} \end{aligned}$$

Since the expectation of each y_1 through y_n is $E(y_1) = \mu$, $E(y_2) = \mu$, . . . , $E(y_n) = \mu$, we can write

$$E(\bar{y}) = \frac{[\mu + \mu + \cdots + \mu]}{n}$$
$$E(\bar{y}) = \frac{n\mu}{n}$$

We note that the n values in numerator and denominator cancel, and so we end up with

$$E(\bar{y}) = \mu$$

Using the fact that $E(y_i) = \mu$, we can also say that the expected value of a sampling distribution of the mean is equal to the mean of the population from which we did the theoretical sampling. That is, $\mu_{\bar{y}} = \mu$ is true, since given $E(\bar{y}) = \mu$, it stands that if we have, say, five sample means $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \bar{y}_5$, the expectation of each of these means should be equal to μ, from which we can easily deduce $\mu_{\bar{y}} = \mu$. That is, *the mean of all the samples we could draw is equal to the population mean*.

We now need a measure of the *dispersion* of a sampling distribution of the mean. At first glance, it may seem reasonable to assume that the variance of the sampling distribution of means should equal the variance of the population from which the sample means were drawn. However, this is not the case. What is true is that the variance of the sampling distribution of means will be equal to only a *fraction* of the population variance. It will be equal to $1/n$ of it, where n is equal to the size of samples we are collecting for each sample mean. Hence, the variance of means of the sampling distribution is equal to

$$\frac{1}{n}(\sigma^2)$$

or simply

$$\frac{\sigma^2}{n}$$

The mathematical proof of this statistical fact is in most mathematical statistics texts. A version of the proof can also be found in Hays (1994). The idea, however, can be easily and perhaps even more intuitively understood by recourse to what happens as n changes. Suppose that we calculate the sample mean from a sample size of $n = 1$, sampled from a population with $\mu = 10.0$ and $\sigma^2 = 2.0$. Suppose the sample mean we obtain is equal to 4.0. Therefore, the sampling variance of the corresponding sampling distribution is equal to

$$\frac{\sigma^2}{n} = \frac{2}{1} = 2$$

That is, the variance in means that you can expect to see if you sampled an infinite number of means based on samples of size $n = 1$ repeatedly from this population is equal to 2. Notice that 2 is exactly equal to the original population variance.

Consider now the case where $n > 1$. Suppose we now sampled a mean from the population based on sample size $n = 2$, yielding

$$\frac{\sigma^2}{n} = \frac{2}{2} = 1$$

What has happened? What has happened is that the variance in sample means has decreased by 1/2 of the original population variance. Why is this decrease reasonable? It makes sense, because we already know from the law of large numbers that as the sample size grows larger, one gets closer and closer to the true probability of a parameter. That is, for a consistent estimator, our estimate of the true population mean should get better and better as sample size increases. This is exactly what happens as we increase n, our *precision* of that which is being estimated increases. In other words, the sampling variance of the estimator *decreases*.

Analogous to how we defined the standard deviation as the square root of the variance, it is also useful to take the square root of the variance of means:

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

which we call the *standard error of the mean* σ_{M}. The standard error of the mean is the standard deviation of the sampling distribution of the mean. Finally, it is important to recognize that $\sigma/\sqrt{n}$ is not "the" standard error. It is merely the standard error *of the mean*. Other statistics will have different standard errors.

3.12 CENTRAL LIMIT THEOREM

It is not an exaggeration to say that the central limit theorem, in one form or another, is probably the most important and relevant theorem in theoretical statistics, which translates to it being quite relevant to applied statistics as well.

We borrow our definition of the central limit theorem from Everitt (2002):

> If a random variable y has a population mean μ and population variance σ^2, then the sample mean, $\bar{y}$, based on n observations, has an approximate normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, for sufficiently large n. (p. 64)

More generally, the theorem pertains to the limiting form of the cumulative distribution function (cdf) of a normal random variable (Casella and Berger, 2002, p. 236). Asymptotically, the distribution of a normal random variable *converges* to that of a normal distribution as $n \to \infty$.

A multivariate version of the theorem can also be given (Rencher, 1998, p. 53).[10] The relevance of the central limit theorem cannot be overstated: It allows one to know, at least on a theoretical level, what the distribution of a statistic (e.g., sample mean) will look like for increasing sample size. This is especially important if one is drawing samples from a population for which the shape is not known or is known *a priori* to be nonnormal. *Normality of the sampling distribution is still assured even if samples are drawn from nonnormal populations*. Why is this relevant? It is relevant because if we know what the distribution of means will look like for increasing sample size, then we know we can compare our obtained statistic to a normal distribution in order to estimate its probability of occurrence. Normality assumptions are also typically required for assuming independence between $\bar{y}$ and s^2 in univariate contexts, and between $\bar{\mathbf{y}}$ (mean vector) and $\mathbf{S}$ (covariance matrix) in multivariate ones. When such estimators can be assumed to arise from normal or multivariate normal distributions (i.e., in the case of $\bar{\mathbf{y}}$ and $\mathbf{S}$), we can generally be assured one is independent of the other.

3.13 CONFIDENCE INTERVALS

Recall that a goal of statistical inference is to estimate functions of parameters, whether a single parameter, a difference of parameters (e.g., in the case of population differences), or some other function of parameters. Though the sample mean $\bar{y}$ is an unbiased estimator of μ, the probability that $\bar{y}$ is equal to μ in any given sample is equal to zero. For this reason, and to build some flexibility in estimation overall, the idea of interval estimation in the form of *confidence intervals* was developed. Confidence intervals provide a range of values for which we can be relatively certain lay the true parameter we are seeking to estimate. In what follows, we provide a brief review of 95% and 99% confidence intervals.

We can say that over all samples of a given size n, the probability is 0.95 for the following event to occur:

$$-1.96\sigma_{\mathrm{M}} < \bar{y} - \mu < 1.96\sigma_{\mathrm{M}} \tag{3.2}$$

How was (3.2) obtained? Recall the calculation of a z-score for a mean (see Section 3.19.1):

$$z = \frac{\bar{y} - \mu}{\sigma_{\mathrm{M}}}$$

Suppose now that we want to have a 0.025 area on either side of the normal distribution. This value corresponds to a z-score of 1.96, since the probability of a z-score of ± 1.96 is $2(1 - 0.9750021) = 0.0499958$, which is approximately 5% of the

[10]We can also distinguish between weaker versus stronger forms of the theorem. For details, see Casella and Berger (2002, pp. 236–238).

total curve. So, from the z-score, we have

$$\begin{aligned} z &= \frac{\overline{y}-\mu}{\sigma_{\mathrm{M}}} \\ \pm 1.96 &= \frac{\overline{y}-\mu}{\sigma_{\mathrm{M}}} \\ (\sigma_{\mathrm{M}}) \pm 1.96 &= \overline{y}-\mu \end{aligned}$$

We can modify the equality slightly to get the following:

$$\overline{y} - 1.96\sigma_{\mathrm{M}} < \mu < \overline{y} + 1.96\sigma_{\mathrm{M}} \tag{3.3}$$

We interpret (3.3) as follows:

> Over all possible samples, the probability is 0.95 that the range between $\overline{y} - 1.96\sigma_{\mathrm{M}}$ and $\overline{y} + 1.96\sigma_{\mathrm{M}}$ will include the true mean, μ.

Very important to note regarding the above statement is that μ is *not* the random variable. The part that is random is the sample on which is computed the interval. That is, the probability statement is not about μ but rather about *samples*. The population mean μ is assumed to be *fixed*.

The 99% confidence interval for the mean is likewise given by

$$\overline{y} - 2.58\sigma_{\mathrm{M}} < \mu < \overline{y} + 2.58\sigma_{\mathrm{M}} \tag{3.4}$$

Notice that the only difference between (3.3) and (3.4) is the choice of different critical values on either side of μ (i.e., 1.96 for the 95% interval and 2.58 for the 99% interval).

Though not very useful, a 100% confidence interval, if constructed, would be defined as

$$\overline{y} - \infty\sigma_{\mathrm{M}} < \mu < \overline{y} + \infty\sigma_{\mathrm{M}}$$

If you think about it carefully, the 100% confidence interval should make perfect sense. If you would like to be 100% "sure" that the interval will cover the true population mean, then you have to extend your limits to negative and positive infinity, otherwise, you could not be *fully* confident. Likewise, on the other extreme, a 0% interval would simply have $\overline{y}$ as the upper and lower limits:

$$\overline{y} < \mu < \overline{y}$$

That is, if you want to have zero confidence in guessing the location of the population mean, μ, then guess the sample mean $\overline{y}$. Though the sample mean is an unbiased

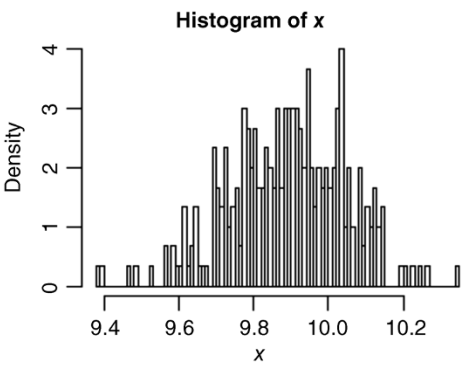
estimator of the population mean, the probability that the sample mean covers the population mean exactly, as mentioned, is equal to 0.

3.14 BOOTSTRAP AND RESAMPLING TECHNIQUES

There are times where estimating parameters through analytical methods proves futile or is otherwise very difficult. As a result of the advent of high-speed computing, techniques for what amounts to essentially simulating sampling distributions have come into vogue. Instead of deriving closed-form solutions (i.e., algebraically solvable equations or systems) for confidence intervals, for instance, one can run simulations on a given sample in order to approximate the behavior of the given sampling distribution and hence obtain an estimate of its sampling variability and stability. The so-called *bootstrap* technique (Efron and Tibshirani, 1993) is a special case of the wider *resampling techniques* available in estimating parameters. For a very user-friendly introduction to resampling procedures, see Howell (2002, pp. 691–707). Fox (1997) also provides an excellent discussion of bootstrapping in the context of regression analysis. The deeper technicalities of the bootstrap are beyond the scope of this book.

As a simple example of the bootstrap technique, suppose we wished to bootstrap a mean and standard deviation of the mean. In what follows, we first obtain a random sample of size $n = 100$ from a population with mean $\mu = 10$ (`rnorm(100, mean = 10)`). We then resample 300 times of size 50 (`300, sample(random.sample, 50)`), upon which the mean of each of these resamples is plotted with a histogram:

```
> random.sample <- rnorm(100, mean = 10)
> resample <- replicate(300, sample(random.sample, 50, TRUE),
simplify = FALSE)
> x <- sapply(resample, mean, simplify = TRUE)
> hist(x, breaks = 100, prob = TRUE)
```

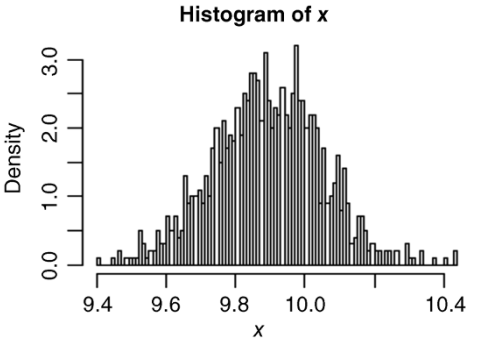


The mean (`mean(x)`) and standard deviation of the distribution (`sd(x)`) are equal to 9.89 and 0.16, respectively. You may obtain slightly

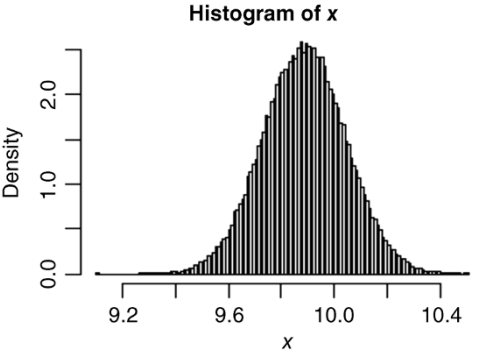
different values for these, since by the very nature of the resampling technique, they are not expected to be constant for each simulation.

Suppose we now increase the number of samplings from 300 to 1000:

```
> resample <- replicate(1000, sample(random.sample, 50, TRUE),
simplify = FALSE)
> x <- sapply(resample, mean, simplify = TRUE)
> hist(x, breaks = 100, prob = TRUE)
```



The mean and standard deviation of the sampling distribution based on 1000 samplings remains approximately the same (i.e., mean = 9.892, SD = 0.156); however, note that the *shape* of the distribution has converged closer to one of idealized normality. This is a direct consequence of the increase in samplings. To further demonstrate, let us sample 100,000 cases:



We note that for 100,000 samples, the empirical distribution closely resembles that of a smooth normal distribution. Indeed, the bootstrap technique, in addition to being useful in estimating parameters, can be used in demonstrating the convergence of the central limit theorem for increasingly large samples.

3.15 LIKELIHOOD RATIO TESTS AND PENALIZED LOG-LIKELIHOOD STATISTICS

When we speak of *likelihood*, we mean the probability of some sample data or set of observations dependent on some hypothesized parameter or set of parameters (Everitt, 2002). Probability statements such as $p(D/H_0)$ are simple examples of likelihoods, where typically the set of parameters in this case may be simply μ and σ^2.

A likelihood ratio test is a comparison (in the form of a ratio) of two likelihoods. Oftentimes in statistical procedures, we evaluate the *log* of the likelihood ratio test of the form:

$$\lambda = -2\ln\left(\frac{L_c}{L_s}\right) = -2[\log_e L_c - \log_e L_s] \tag{3.5}$$

where L_c (or H_0) is the likelihood of the observed data under the current model under investigation, and L_s (or H_1) is the likelihood of the observed data under what is often (but not always) a *saturated* model. Saturated models fit the data perfectly as a result of having as many parameters as there are values to be fit. Hence, we can see that (3.5) compares a model c based on some restrictions imposed by the researcher with that of one s that has no such restrictions. As we will discuss further in Chapter 16 when we survey structural equation models, saturated models also have zero degrees of freedom. Note that the reason we are able to write L_c/L_s as $\log_e L_c - \log_e L_s$ is due simply to the property of logarithms that says the quotient of logarithms $\log(a)/\log(b)$ is equal to $\log(a) - \log(b)$. If the likelihood under each model is the same, then L_c/L_s will be equal to 1, and we obtain

$$\lambda = -2\ln(L_c/L_s) = -2[\log_e L_c - \log_e L_s] = 0.$$

The extent to which $L_c \neq L_s$ is the extent to which λ will be unequal to 0.

When we speak of *maximum-likelihood* (ML) estimation, we mean the process of maximizing a likelihood subject to certain parameter conditions. As a simple example, suppose we obtain 8 heads on 10 flips of a presumably fair coin. Our null hypothesis was that the coin is fair, meaning that the probability of heads is $p(H) = .5$. However, our actual obtained result of 8 heads on 10 flips would suggest the true probability of heads to be closer to $p(H) = .8$. Thus, we ask the question:

> Which value of θ makes the observed result most likely?

If we only had two choices of θ to select from, 0.5 and 0.8, our answer would have to be 0.8, since this value of the parameter θ makes the sample result of 8 heads out of 10 flips most *likely*. This is the essence of how maximum-likelihood estimation works. ML is by far the most common method of estimating parameters in many

models, including factor analysis, path analysis, and structural equation models to be discussed later in the book. There are very good reasons why mathematical statisticians generally approve of maximum likelihood. We summarize some of their most favorable properties.

First, ML estimators are *asymptotically unbiased*, which means that bias essentially vanishes as sample size increases without bound (Bollen, 1989). Second, ML estimators are *consistent* and asymptotically *efficient*, meaning that the estimator has a small asymptotic variance relative to many other estimators. Third, ML estimators are asymptotically normally distributed, meaning that as sample size grows, the estimator takes on a normal distribution. Finally, ML estimators possess the *invariance* property (Casella and Berger, 2002). This property essentially states that if for a maximum-likelihood estimator θ' of θ, θ' remains as the ML estimator for any function of θ. As noted in Bollen (1989), a concept related to the invariance property is that of *scale freeness*, which essentially means that even if we linearly transform observed variables, the estimates of the parameters of the model remain unchanged.

3.16 AKAIKE'S INFORMATION CRITERIA

A measure of model fit commonly used in comparing nested models that uses the log-likelihood, $-2L_{\mathrm{m}}$, is *Akaike's information criteria* (*AIC*) (Sakamoto, Ishiguro, and Kitagawa, 1986). This is one statistic of the kind generally referred to as *penalized log-likelihood* statistics (another is the *Bayesian Information Criterion* (*BIC*)). It is defined as

$$-2L_{\mathrm{m}} + 2m$$

where L_{m} is the maximized log-likelihood and m is the number of parameters in the given model. Lower values of AIC generally indicate a better-fitting model than do larger values. Recall that the more parameters fit to a model, in general, the better the fit of that model. For example, a model that has a unique parameter for *each* data point would fit perfectly. This is the so-called *saturated* model mentioned earlier. AIC jointly considers both the goodness of fit and the number of parameters required to obtain the given fit, essentially "penalizing" for increasing the number of parameters unless they contribute to model fit. Adding one or more parameters to a model may cause $-2L_{\mathrm{m}}$ to decrease (which is a good thing substantively), but if the parameters are not worthwhile, this will be offset by an increase in $2m$.

The *Bayesian information criterion*, or BIC, (Schwarz, 1978) is defined as $-2L_{\mathrm{m}} + m\log(N)$, where m, as before, is the number of parameters in the model and N is the total number of observations used to fit the model. Lower values of BIC are also desirable when comparing models. BIC typically penalizes model complexity more heavily than AIC. For a comparison of AIC and BIC, see Burnham and Anderson (2011).

3.17 COVARIANCE AND CORRELATION

The covariance of a random variable is given by

$$\text{cov}\,(x_i, y_i) = \sigma_{xy} = E[(x_i - \mu_x)(y_i - \mu_y)]$$

where $E[(x_i - \mu_x)(y_i - \mu_y)]$ is equal to $E(x_i y_i) - \mu_y \mu_x$ since

$$\begin{aligned}
\sigma_{xy} &= E[(x_i - \mu_x)(y_i - \mu_y)] \\
&= E(x_i y_i - x_i \mu_y - y_i \mu_x + \mu_x \mu_y) \\
&= E(x_i y_i) - E(x_i)\mu_y - E(y_i)\mu_x + \mu_x \mu_y \\
&= E(x_i y_i) - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\
&= E(x_i y_i) - \mu_y \mu_x
\end{aligned}$$

The concept of covariance is at the heart of all statistical methods. Whether one is running analysis of variance, regression, principal components analysis, etc., covariance concepts are central to all of these methodologies and even more broadly to science in general.

The sample covariance is a measure of relationship between two variables, and is defined as

$$\text{cov} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n} \tag{3.6}$$

The numerator of the covariance, $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$, is the sum of products of respective deviations of observations from their respective means. If there is no *linear* relationship between two variables in a sample, covariance will equal 0. If there is a *negative* linear relationship, covariance will be a negative number, and if there is a *positive* linear relationship, covariance will be positive. Notice that to measure covariance between two variables requires there to be *variability* on each variable. If there is no variability in x_i, then $(x_i - \bar{x})$ will equal 0 for all observations. Likewise, if there is no variability in y_i, then $(y_i - \bar{y})$ will equal 0 for all observations on y_i. This is to emphasize the essential fact that when measuring the extent of relationship between two variables, one requires variability on each variable for the measure to even make sense to compute.

The covariance of (3.6) is a perfectly reasonable one to calculate for a sample if there is no intention of using that covariance as an estimator of the population covariance. However, if one wishes to use it as an estimator, similar to how we needed to subtract 1 from the denominator of the variance, we lose 1 degree of freedom when computing the covariance:

$$\text{cov} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

It is oftentimes thought that the correct denominator should be $n-2$ instead of $n-1$ to account for the fact that we are estimating two quantities in the numerator, $\overline{x}$ and $\overline{y}$. However, the correct denominator is indeed $n-1$ because the numerator is a *cross-product* of deviations, which we treat as a single quantity, not two.

It is easy to understand more of what the covariance actually measures if we consider the trivial case of computing the covariance of a variable with itself. In such a case, for variable x_i, we would have

$$\text{cov} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})}{n-1}$$

But what is this covariance? If we rewrite the numerator as $(x_i-\overline{x})^2$ instead of $(x_i-\overline{x})(x_i-\overline{x})$, it becomes clear that the covariance of a variable with itself is nothing more than the usual *variance* for that variable. Hence, to better understand the covariance, it is helpful to start with the variance, and then realize that instead of computing the cross-product of a variable with *itself*, the covariance computes the cross-product of a variable with a *second* variable.

We compute the covariance between parent height and child height in Galton's data:

```
> attach(Galton)
> cov(parent, child)
[1] 2.064614
```

We have mentioned that the covariance is a measure of linear relationship. However, sample covariances from data set to data set are not comparable unless one knows more of what went into each specific computation. There are actually three things that are the "ingredients" of the covariance. The first thing it contains is a measure of the cross-product, which represents the degree to which variables are linearly related. This is the part in our computation of the covariance that we are especially interested in. However, other than concluding a negative, zero, or positive relationship, the size of the covariance does not by itself tell us the *degree* to which two variables are related.

The reason for this is that the size of covariance will also be impacted by the degree to which there is variability in x_i and the degree to which there is variability in y_i. If either or both variables contain sizable deviations of the sort $(x_i-\overline{x})$ or $(y_i-\overline{y})$, then the corresponding cross-products $(x_i-\overline{x})(y_i-\overline{y})$ will also be quite sizable, along with their sum, $\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$. However, we do not want our measure of relationship to be small or large as a consequence of variability on x_i or variability on y_i. We want our measure of relationship to be small or large as an exclusive result of *covariability*, that is, the extent to which there is actually a *relationship* between x_i and y_i. To "remove" the influences of variability in x_i and y_i (one may think of it as "purifying"), we divide the average cross-product by the product of standard deviations of each variable. The *standardized sample*

covariance is thus

$$r = \frac{\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))/n-1}{\sqrt{s_{x_i}^2 \cdot s_{y_i}^2}} = \frac{\text{cov}}{\sqrt{s_{x_i}^2 \cdot s_{y_i}^2}}$$

The standardized covariance is known as the *Pearson product-moment correlation coefficient*, or simply r, which is a *biased* estimator of its population counterpart ρ_{xy}. The bias of the estimator r can be minimized by computing an adjustment found in Rencher (1998), originally proposed by Olkin and Pratt (1958):

$$r^* = r\left[1 + \frac{1 - r^2}{2(n-3)}\right]$$

Because the correlation coefficient is standardized, we can place lower and upper bounds on it. The minimum correlation possible for any set of data is −1.0, representing a perfect negative relationship. The maximum correlation possible is +1.0, representing a perfect positive relationship. A correlation of 0 represents the absence of a *linear* relationship.

One can gain an appreciation for the upper and lower bounds of r by considering the fact that the numerator, which is an average cross-product, is being divided by another product, that of the standard deviations of each variable. The denominator thus can be conceptualized to represent the total amount of cross-product variation *possible*, that is, the "base," whereas the numerator represents the total amount of cross-product variation actually existing between the variables *because of a linear relationship*. The extent to which cov_{xy} accounts for all of the possible "cross-variation" in $\sqrt{s_{x_i}^2 \cdot s_{y_i}^2}$ is the extent to which r will approximate a value of $|1|$ (either positive or negative, depending on the direction of the relationship). It thus stands that cov_{xy} cannot be greater than the "base" with which it is being compared (i.e., $\sqrt{s_{x_i}^2 \cdot s_{y_i}^2}$). In the language of sets, cov_{xy} must be a *subset* of the larger set represented by $\sqrt{s_{x_i}^2 \cdot s_{y_i}^2}$.

It is important to emphasize that a correlation of 0 does not necessarily represent the *absence* of a relationship. What it does represent is the absence of a *linear* one. Neither the covariance nor Pearson's r capture nonlinear relationships, and so it is possible to have very strong relations in a sample or population yet obtain very low values (even zero) for the covariance or Pearson r. Always plot your data to see what is going on before drawing any conclusions. Correlation coefficients should never be presented without an accompanying plot to characterize the *form* of the relationship.

We compute the Pearson correlation coefficient on Galton's data between `child` and `parent`:

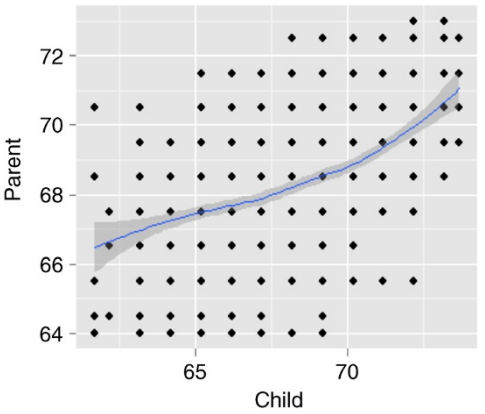
```
> cor(child, parent)
[1] 0.4587624
```

We can test it for statistical significance by using the `cor.test` function:

```
> cor.test(child, parent)

        Pearson's product-moment correlation

data: child and parent
t = 15.7111, df = 926, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4064067 0.5081153
sample estimates:
      cor
0.4587624
```

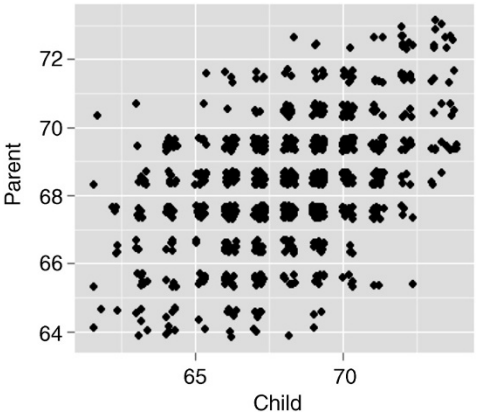
We can see that observed t is statistically significant with a computed 95% confidence interval having limits 0.41–0.51, indicating that in 95% of samples drawn from this population, the true parameter will lay approximately between the limits of 0.41 and 0.51. Using the package `ggplot2` (Wickham, 2009), we plot the relationship between parent and child (with a smoother):

```
> library(ggplot2)
> qplot(child, parent, data = Galton, geom = c("point", "smooth"))
```



One drawback of such a simple plot is that the frequency of data points in the bivariate space cannot be known by inspection of the plot alone. *Jittering* is a technique that allows one to visualize the density of points at each parent–child pairing. By jittering, we can see where most of the data fall in the parent–child scatterplot (i.e., points are concentrated toward the center of the plot):

```
> qplot(child, parent, geom = "jitter")
```



3.17.1 Covariance and Correlation Matrices

Having reviewed the concept of covariance, we need a way to account for the covariance of many variables. For this, we write the sample covariance in matrix form:

$$\mathbf{S} = (s_{jk}) = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

where s_{jk} are the covariances for variables j by k. The population covariance matrix Σ can be analogously defined:

$$\sum = \sigma_{jk} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

where along the main diagonal of the covariance matrix are variances σ_{11}, σ_{22}, etc., for variables 1, 2, etc., up to σ_{pp}, the variance of the pth variable.

When we standardize the covariance matrix, dividing each of its elements by respective products of standard deviations, we obtain the correlation matrix:

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & \dots & \dots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

where r_{12} is the correlation between variable 1 and 2, etc., and r_{1p} is the correlation between variable 1 and the pth variable.

An example of a correlation matrix (Heston, 1948) is that between different tests on the GRE (*Graduate Record Examination*):

Intercorrelations Among The G.R.E. Tests Of General Education

	Math	P.S.	B.S.	Soc.	Lit.	Arts	Exp.	Voc.
Mathematics		.55	.44	.51	.36	.35	.52	.38
Physical Science	.55		.49	.43	.20	.40	.32	.29
Bioogical Science	.44	.49		.57	.42	.42	.46	.50
Social Studies	.51	.43	.57		.54	.40	.61	.59
Literature	.36	.20	.42	.54		.39	.53	.54
Arts	.35	.40	.42	.40	.39		.42	.52
Effecive Expression	.52	.32	.46	.61	.53	.42		.66
Vocabulary	.38	.29	.50	.59	.54	.52	.66	

From the matrix we can see that most correlations are low to moderate, with the correlation between Effective Expression and Vocabulary relatively large at a value of 0.66. The correlation between Physical Science and Vocabulary is relatively small, equaling 0.29.

3.18 OTHER CORRELATION COEFFICIENTS

It often happens that once we hear of Pearson r, this becomes the *only* correlation coefficient in one's vocabulary, and too often the *concept*, rather than *calculation*, of a correlation is automatically linked to Pearson r. Pearson r is but one of *many* correlation coefficients available at one's disposal in applied research. Recall that Pearson r captures *linear* relationships between (typically) continuous variables. If the relationship is not linear, or one or more variables are not continuous, or again if the data are in the form of ranks, then other correlation coefficients are generally more

TABLE 3.5 Favorability of Movies for Two Individuals in Terms of Ranks

Movie	Bill	Mary
Batman	5 (2.1)	5 (7.6)
Star Wars	1 (10.0)	3 (9.0)
Scarface	3 (8.4)	1 (9.7)
Back to the Future	4 (7.6)	4 (8.5)
Halloween	2 (9.5)	2 (9.6)

Actual scores on the favorability measure are in parentheses.

suitable. We briefly review Spearman's rho, although a host of other correlation coefficients exist that are well-suited for a variety of particular types of data.[11]

Spearman's r_s ("rho"), named after Charles Spearman who developed the coefficient in 1904,[12] is a correlation coefficient suitable for data on two variables that are expressed in terms of *ranks* rather than actual measurements on a continuous scale. Mathematically, the Spearman correlation coefficient is equivalent to a Pearson r when the data are ranked. There are nonetheless important differences between these two coefficients. Spearman's r_s is defined as

$$r_s = 1 - \frac{6\sum (R_x - R_y)^2}{n(n^2-1)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where R_x and R_y are the ranks on x_i and y_i, d_i^2 are squared rank deviations, and n is the number of pairs. When we compute r_s on the Galton data, we obtain

```
> cor.test(parent, child, method = "spearman")

        Spearman's rank correlation rho

data: parent and child
S = 76569964, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4251345
```

We see that r_s of 0.425 is slightly less than was Pearson r of 0.459.

To understand why the Spearman rank correlation and the Pearson coefficient differ, consider data (Table 3.5) on the rankings of favorite movies for two

[11]For an overview of alternative correlation coefficients such as the *biserial*, *point-biserial*, and *tetrachoric* coefficients, see Howell (2002) and Warner (2013).

[12]The coefficient appears in Spearman (1904b).

individuals. In parentheses are subjective scores of "favorability" of these movies, scaled 1–10, where 1 = least favorable and 10 = most favorable.

From the table, we can see that Bill very much favors Star Wars (rating of 10), while least likes Batman (rating of 2.1). Mary's favorite movie is Scarface (rating of 9.7), while her least favorite movie is Batman (rating of 7.6). We will refer to these subjective scores in a moment. For now, we focus only on the ranks. For instance, Bill's ranking of Scarface is third, while Mary's ranking of Star Wars is third.

To compute Spearman's r_s in R the "long way," we generate two vectors that contain the respective rankings:

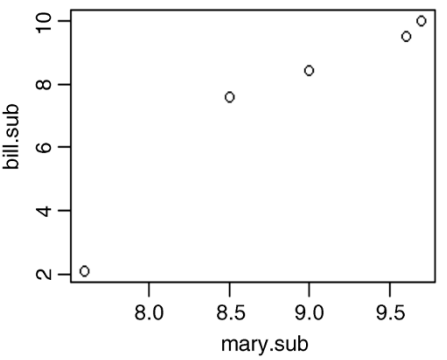
```
> bill <- c(5, 1, 3, 4, 2)
> mary <- c(5, 3, 1, 4, 2)
```

Because the data are already in the form of ranks, both Pearson r and Spearman rho will agree:

```
> cor(bill, mary)
[1] 0.6

> cor(bill, mary, method = "spearman")
> 0.6
```

Note that by default, R returns the Pearson correlation coefficient. One has to specify `method = "spearman"` to get r_s. Consider now what happens when we correlate, instead of rankings, the actual subjective favorability scores corresponding to the respective ranks. When we plot the favorability data, we obtain

```
> bill.sub <- c(2.1, 7.6, 8.4, 9.5, 10.0)
> mary.sub <- c(7.6, 8.5, 9.0, 9.6, 9.7)
> plot(mary.sub, bill.sub)
```



Note that although the relationship is not perfectly linear, each increase in Bill's subjective score is nonetheless associated with an increase in Mary's subjective score. When we compute Pearson *r* on these data, we obtain

```
> cor(bill.sub, mary.sub)
[1] 0.9551578
```

However, when we compute r_s, we get

```
> cor(bill.sub, mary.sub, method = "spearman")
[1] 1
```

Spearman r_s is equal to 1.0 because the rankings of movie preferences are perfectly *monotonically increasing* (i.e., for each increase in movie preference along the abscissa corresponds an increase in movie preference along the ordinate). In the case of Pearson's *r*, the correlation is less than 1.0 because *r* captures the *linear* relationship among variables and not simply a monotonically increasing one. Hence, a high magnitude coefficient for Spearman's r_s essentially tells us that two variables are "moving together," but it does not necessarily imply the relationship is a linear one. A similar test that measures rank correlation is that of Kendall's rank–order correlation. See Siegel and Castellan (1988, p. 245) for details.

3.19 STUDENT'S *t* DISTRIBUTION

The density for student's *t* is given by

$$f(t) = \frac{\Gamma[(v+1)/2]}{\sqrt{v\pi}\Gamma(v/2)}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}$$

where Γ is the gamma function and *v* are degrees of freedom. For small degrees of freedom *v*, the *t* distribution is quite distinct from the standard normal. However, as degrees of freedom increase, the *t* distribution converges to that of a normal density (Figure 3.11). That is, in the limit, $f(t) \rightarrow f(z)$, or more formally, $\lim_{v\rightarrow\infty} f(t) = f(z)$.

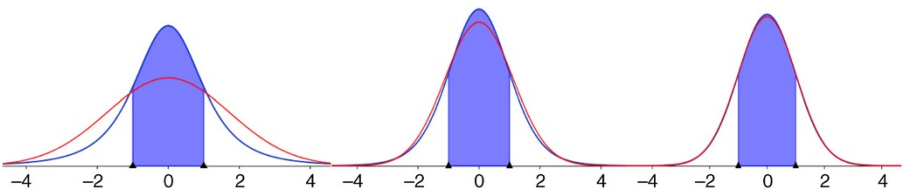


FIGURE 3.11 Student's *t* versus normal densities for 3 (left), 10 (middle), and 50 (right) degrees of freedom. As degrees of freedom increase, the limiting form of the *t* distribution is the *z* distribution.

The fact that t converges to z for large degrees of freedom but is quite distinct from z for small degrees of freedom is one reason why t distributions are often used for *small sample* problems. When sample size is large, and so consequently are degrees of freedom, whether one treats a random variable as t or z will make little difference in terms of computed p-values and decisions on respective null hypotheses. This is a direct consequence of the convergence of the two distributions for large degrees of freedom.

3.19.1 *t*-Tests for One Sample

When we perform hypothesis testing using the z distribution, we assume we have knowledge of the population variance σ^2. Having direct knowledge of σ^2 is the most *ideal* of circumstances. When we know σ^2, we can compute the standard error of the mean directly as

$$\sigma_{\mathrm{M}} = \frac{\sigma}{\sqrt{n}}$$

Recall that the form of the one-sample z test for the mean is given by

$$z_{\mathrm{M}} = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}}$$

where the numerator $\overline{y} - \mu_0$ represents the distance between the sample mean and the population mean μ_0 under the null hypothesis, and the denominator $\sigma/\sqrt{n}$ is the standard error of the mean.

In most research contexts, from simple to complex, we usually do *not* have direct knowledge of σ^2. When we do not have knowledge of it, we use the next best thing, an *estimate* of it. We can obtain an unbiased estimate of σ^2 by computing s^2 on our sample. When we do so however, and use s^2 in place of σ^2, we can no longer pretend to "know" the standard error of the mean. Rather, we must concede that all we are able to do is estimate it. Our estimate of the standard error of the mean is thus given by

$$\hat{\sigma}_{\mathrm{M}} = \frac{s}{\sqrt{n}}$$

When we use s^2 (where $\sqrt{s^2} = s$) in place of σ^2, our resulting statistic is no longer a z statistic. That is, we say the ensuing statistic is no longer *distributed* as a standard normal variable (i.e., z). If it is not distributed as z, then what is it distributed as? Thanks to William Sealy Gosset who in 1908 worked for *Guinness Breweries* under the pseudonym "Student," the ratio

$$t = \frac{\overline{y} - E(\overline{y})}{\hat{\sigma}_{\mathrm{M}}} = \frac{\overline{y} - E(\overline{y})}{s/\sqrt{n}}$$

was found to be distributed as a t-statistic on $n-1$ degrees of freedom. Again, the t distribution is most useful when sample sizes are rather small. For larger samples, as mentioned, the t distribution converges to that of the z distribution. If you are using rather large samples, say approximately 100 or more, whether you test your null hypothesis using a z or t distribution will not matter much, because the critical values for z and t for such degrees of freedom (99 for the one-sample case) will be so similar that practically, they can be considered more or less equal. For even larger samples, the convergence is that much more fine-tuned.

The concept of convergence between z and t can be easily illustrated by inspecting the variance of the t distribution. Unlike the z distribution where the variance is set at 1.0 as a constant, the variance of the t distribution is defined as

$$\sigma_t^2 = \frac{v}{v-2}$$

where v are the degrees of freedom. For small degrees of freedom, such as $v=5$, the variance of the t distribution is equal to

$$\sigma_t^2 = \frac{5}{5-2} = \frac{5}{3} \approx 1.67$$

Note what happens as v increases, the ratio $v/(v-2)$ gets closer and closer to 1.0, which is the precise variance of the z distribution. For example, $v=20$ yields

$$\sigma_t^2 = \frac{20}{20-2} = \frac{20}{18} \approx 1.11$$

which is already quite close to the variance of a standardized normal variable z (i.e., 1.0).

Hence, we can say more formally

$$\lim_{v \to \infty} \left(\frac{v}{v-2} \right) = 1.0$$

That is, as v increases without bound, the variance of the t distribution equals that of the z distribution, which is equal to 1.0.

We demonstrate the use of the one-sample t-test using SPSS. Consider the following small, hypothetical data on IQ scores of five individuals:

```
IQ
105
 98
110
105
 95
```

Suppose that the hypothesized mean IQ in the population is equal to 100. The question we want to ask is: *Is it reasonable to assume that our sampled data could have arisen from a population with mean IQ equal to 100?* We assume we have no knowledge of the population standard deviation, and hence must estimate it from our sample data. To perform the one-sample *t*-test in SPSS, we compute:

```
T-TEST
  /TESTVAL=100
  /MISSING=ANALYSIS
  /VARIABLES=IQ
  /CRITERIA=CI(.95).
```

The line /TESTVAL=100 inputs the test value for our hypothesis test, which for our null hypothesis is equal to 100. We have also requested a 95% confidence interval for the mean difference.

One-Sample Statistics

	N	Mean	SD	SE Mean
IQ	5	102.6000	6.02495	2.69444

We confirm from the above that the size of our sample is equal to 5, and the mean IQ for our sample is equal to 102.60 with standard deviation 6.02. The standard error of the mean reported by SPSS of 2.69 is actually not the *true* standard error of the mean. It is the *estimated* standard error of the mean, since recall that we did not have knowledge of the population variance (otherwise we would have been performing a *z*-test instead of a *t*-test).

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	Df	Sig. (2-tailed)	Mean Difference	Lower	Upper
IQ	0.965	4	0.389	2.60000	–4.8810	10.0810

We note the following from the above output:

- Our obtained *t*-statistic is equal to 0.965 and is evaluated on four degrees of freedom (i.e., $n - 1 = 5 - 1 = 4$). We lose a degree of freedom because recall that in estimating the population variance σ^2 with s^2, we had to compute a sample mean $\bar{y}$ and hence this value is regarded as "fixed" as we carry on with our *t*-test. Hence, we lose a single degree of freedom.

- The two-tailed *p*-value is equal to 0.389, which, assuming we had set our criteria for rejection at $\alpha = 0.05$, leads us to the decision to not reject the null hypothesis. The two-tailed (as opposed to *one-tailed* or *directional*) nature of the statistical test in this example means that we allow a rejection of the null hypothesis in either direction from the value stated under the null. Since our null hypothesis is $\mu_0 = 100$, it means we were prepared to reject the null hypothesis for observed values of the sample mean that deviate "significantly" either *greater than* or *less than* 100. Since our significance level was set at 0.05, this means that we have $0.05/2 = 0.025$ area in each end of the *t* distribution to specify as our rejection region for the test. The question we are asking of our sample mean is: *What is the probability of observing a sample mean that falls much greater OR much less than 100?* Because the observed sample mean can only fall in one tail or the other on any single trial (i.e., we are conducting a single "trial" when we run this experiment a single time), this implies these two events are mutually exclusive, which by the addition rule for mutually exclusive events, we can add them. When we add their probabilities, we get $0.025 + 0.025 = 0.05$, which, of course, is our significance level for the test.
- The actual mean difference observed is equal to 2.60, which was computed by taking the mean of our sample, that of 102.6 and subtracting the mean hypothesized under the null hypothesis, that of 100 (i.e., $102.6 - 100 = 2.60$).
- The 95% confidence interval of the difference is interpreted to mean that in 95% of samples drawn from this population, the interval with lower bound −4.8810 and upper bound 10.0810 will capture the true parameter, which in this case is the population mean difference. We can see that the population mean difference of zero lies within the limits of the confidence interval, which again confirms why we were unable to reject the null hypothesis at the 0.05 level of significance. Had zero lay outside of the confidence interval limits, this would have been grounds to reject the null at a significance level of 0.05 (and consequently, we would have also obtained a *p*-value of less than 0.05 for our significance test). Recall that the true mean (i.e., parameter) is not the random component. Rather, *the sample is the random component*, on which the interval is then computed. It is important to emphasize this distinction when interpreting the confidence interval.

We can easily generate the same *t*-test in R. We first generate the vector of data then carry on with the one-sample *t*-test, which we notice mirrors the findings obtained in SPSS:

```
> iq <- c(105, 98, 110, 105, 95)
> t.test(iq, mu = 100)

        One Sample t-test

data: iq
t = 0.965, df = 4, p-value = 0.3892
```

```
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  95.11904 110.08096
sample estimates:
mean of x
    102.6
```

3.19.2 *t*-Tests for Two Samples

Just as the *t*-test for one sample is a generalization of the *z*-test for one sample, for which we use s^2 in place of σ^2, the *t*-test for two independent samples is a generalization of the *z*-test for two independent samples. Recall the *z*-test for two independent samples:

$$z_{\mathrm{M}} = \frac{E(\bar{y}_1) - E(\bar{y}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $E(\bar{y}_1)$ and $E(\bar{y}_2)$ denote the expectations of the sample means $\bar{y}_1$ and $\bar{y}_2$, respectively (which are equal to μ_1 and μ_2).

When we do not know the population variances σ_1^2 and σ_2^2, we shall, as before, obtain *estimates* of them in the form of s_1^2 and s_2^2. When we do so, because we are using these estimates instead of the actual variances, our new ratio is no longer distributed as *z*. Just as in the one-sample case, it is now distributed as *t*:

$$t = \frac{E(\bar{y}_1) - E(\bar{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.7}$$

on degrees of freedom $v = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$.

The *t* in (3.7) assumes that $n_1 = n_2$. If this is false, then *pooling* variances is recommended. To pool, we weight the sample variances by their respective sample sizes and obtain the following estimated standard error of the difference:

$$\hat{\sigma}_{\mathrm{diff}} = \sqrt{\hat{\sigma}^2_{\mathrm{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

which can also be written as

$$\hat{\sigma}_{\mathrm{diff}} = \sqrt{\hat{\sigma}^2_{\mathrm{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{\hat{\sigma}^2_{\mathrm{pooled}}}{n_1} + \frac{\hat{\sigma}^2_{\mathrm{pooled}}}{n_2}}.$$

Notice that the pooled estimate of the variance $[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ is nothing more than a *weighted sum*, each variance being weighted by its respective

sample size. This idea of weighting variances as to arrive at a pooled value is not unique to *t*-tests. Such a concept forms the very fabric of how MS error is computed in the analysis of variance as we will shall see further in Chapter 4 when we discuss the ANOVA procedure in some depth.

3.19.3 Two-Sample *t*-Tests in R

Consider the following hypothetical data on pass–fail grades ("0" is fail, "1" is pass) for a seminar course with 10 attendees:

grade	studytime
0	30
0	25
0	59
0	42
0	31
1	140
1	90
1	95
1	170
1	120

To conduct the two-sample *t*-test, we generate the relevant vectors in R and then carry out the test:

```
> grade.0 <- c(30, 25, 59, 42, 31)
> grade.1 <- c(140, 90, 95, 170, 120)
> t.test(grade.0, grade.1)

        Welch Two Sample t-test

data: grade.0 and grade.1
t = -5.3515, df = 5.309, p-value = 0.002549
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -126.00773 -45.19227
sample estimates:
mean of x mean of y
     37.4     123.0
```

Using a Welch adjustment for unequal variances (Welch, 1947) automatically generated by R, we conclude a statistically significant difference between means ($p = 0.003$). In 95% of samples drawn from these populations, we can say the true mean difference lies between the lower limit of approximately −126.0 and the upper limit of approximately −45.2. As a quick test to verify the assumption of equal variances (and

whether the Welch adjustment was necessary), we can use `var.test` which will produce a ratio of variances and test the null hypothesis that this ratio should equal 1 (i.e., if the variances are equal, the numerator of the ratio will be the same as the denominator):

```
> var.test(grade.0, grade.1)

        F test to compare two variances

data: grade.0 and grade.1
F = 0.1683, num df = 4, denom df = 4, p-value = 0.1126
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.01752408 1.61654325
sample estimates:
ratio of variances
         0.1683105
```

The `var.test` yields a *p*-value of 0.11, which under most circumstances would be considered insufficient reason to doubt the null hypothesis of equal variances. Hence, the Welch adjustment on the variances was not needed in this case.

Carrying out the same test in SPSS is straightforward by requesting (output not shown):

```
t-test groups = grade(0 1)
     /variables = studytime.
```

A classic nonparametric equivalent to the independent samples *t*-test is the *Wilcoxon rank–sum* test. It is a useful test to run when either distributional assumptions are known to be violated or when they are unknown and sample size is too small for the central limit theorem to come to the "rescue." The test compares *rankings* across the two samples instead of actual scores. For a brief overview of how the test works, see Kirk (2008, Chapter 18) and Howell (2002, pp. 707–717). We can request the test quite easily in R:

```
> wilcox.test(grade.0, grade.1)

        Wilcoxon rank sum test

data: grade.0 and grade.1
W = 0, p-value = 0.007937
alternative hypothesis: true location shift is not equal to 0
```

We see that the obtained *p*-value still suggests we reject the null hypothesis, though the *p*-value is slightly larger than that for the parametric test.

3.20 STATISTICAL POWER

Power, first and foremost, is a *probability*. Power is the probability of rejecting a null hypothesis given that the null hypothesis is false. It is equal to $1-\beta$ (i.e., 1 minus the type II error rate). If the null hypothesis were true, then regardless of how much power one has, one will still not be able to reject the null. We may think of it in terms of the *sensitivity* of a statistical test for detecting the falsity of the null hypothesis. If the test is not very sensitive to departures from the null (i.e., in terms of a particular alternative hypothesis), we will not detect such departures. If the test is very sensitive to such departures, then we will correctly detect these departures and be able to infer the statistical alternative hypothesis in question.

A useful analogy for understanding power is to think of a sign on a billboard that reads "H_0 is false." Are you able to detect such a sign with your current glasses or contacts that you are wearing? If not, you lack sufficient power. That is, you lack the sensitivity in your instrument (your reading glasses) to correctly detect the falsity of the null hypothesis, and in doing, be in a position to reject it. Alternatively, if you have 20/20 vision, you will be able to detect the false null with ease, and reject it with confidence. A key point to note here is that if H_0 is false, it is false *regardless of your ability to detect it*, analogous to a virus strain being present but biomedical engineering lacking a powerful enough microscope to see it. If the null is false, the only question that remains is whether or not you will have a powerful enough test to detect its falsity. If the null were not false on the other hand, then regardless of your degree of power, you will not be able to detect its falsity (because it is not false to begin with).

Power is a function of four elements, all of which will be covered in our discussion of the *p*-value toward the conclusion of this chapter:

1. The value hypothesized under the statistical alternative hypothesis H_1. All else equal, a greater distance between H_0 and H_1 means greater power. Though "distance" in this regard is not a one-to-one concept with *effect size*, the spirit of the two concepts is the same. The greater the scientific effect, the more power you will have to detect that effect. This is true whether we are dealing with mean differences in ANOVA-type models or testing a null hypothesis of the sort $H_0 : R^2 = 0$ in regression. In all such cases, we are seeking to detect a deviation from the null hypothesis.
2. The significance level or type I error rate (α) at which you set your test. All else equal, a more liberal setting such as 0.05 or 0.10 affords more statistical power than a more conservative setting such as 0.001 or 0.01. It is easier to detect a false null if you allow yourself more of a risk of committing a type I error. Since we usually want to minimize type I error, we typically want to regard α as fixed at a nominal level (e.g., 0.05 or 0.01) and consider it not amenable to adjustment for the purpose of increasing power.
3. Population variability, σ^2, often unknown but estimated by s^2. All else equal, the greater the variance of objects studied in the population, the *less* sensitive the statistical test, and the less power you will have. Why is this so? As an analogy,

consider a rock thrown into the water. The rock will make a definitive particular "splash" in that it will displace a certain amount of water when it hits the surface. This is the "effect size" of the splash. If the water is noisy with wind and waves (i.e., high population variability), it will be difficult to detect the splash. If, on the other hand, the water is calm and serene (i.e., low population variability), you will more easily detect the splash. Either way, the rock made a splash. Whether we can detect the splash or not is in part a function of the variance in the population.

Applying this concept to research settings, if you are sampling from "noisy" populations, it is harder to see the effect of your independent variable than if you are sampling from less noisy, and thus less variable populations. This is why research using laboratory rats or other equally *controllable* objects can usually detect effects with relatively few animals in a sample, whereas research studying humans on variables such as intelligence, anxiety, attitudes, etc., usually requires many more subjects in order to detect effects. A good way to boost power is to study populations that have relatively low variability before your treatment is administered. If your treatment works, you will be able to detect its efficacy with fewer subjects than if dealing with a highly variable population. Another approach is to covary out one or two factors that are thought to be related with the dependent variable through a technique such as the analysis of covariance (Keppel and Wickens, 2004).

4. Sample size n. All else equal, the greater the sample size, the greater the statistical power. Boosting sample size is a common strategy for increasing power. Indeed, as will be discussed in the conclusion of this chapter, for any significance test in which there is at least *some* effect (i.e., some distance between the null and alternative), statistical significance is assured for a large-enough sample size. Obtaining large samples is a *good thing* (since after all, the most ideal goal would be to have the actual *population*), but as sample size increases, the p-value becomes an increasingly poor indicator or measure of experimental effect. *Effect sizes* should always be reported alongside any significance test.

3.20.1 Visualizing Power

Figure 3.12, adapted from Bollen (1989), depicts statistical power under competing values for detecting the population parameter θ. Note carefully in the figure that the critical value for the test remains constant as a result of our desire to keep the type I error rate constant. It is the *distance* from $\theta = 0$ to $\theta = C_1$ or $\theta = C_2$ that determines power (the shaded region in distributions (b) and (c)).

Statistical power matters so long as we have the inferential goal of rejecting null hypotheses. A study that is underpowered risks not being able to reject null hypotheses even if such null hypotheses are in reality false. A failure to reject a null hypothesis under the condition of minimal power could either mean a lack of inferential support for the

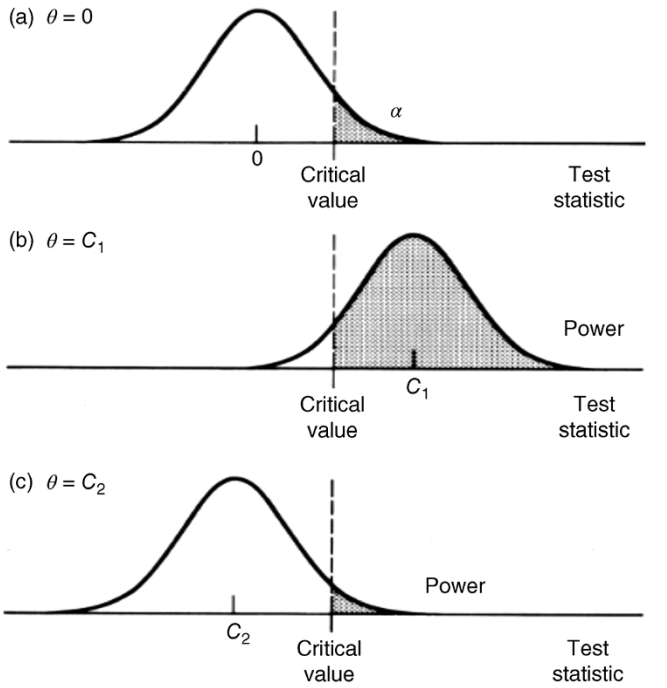


FIGURE 3.12 Power curves for detecting parameters C_1 and C_2 (Bollen, 1989). Reproduced with permission from John Wiley & Sons, Inc.

obtained finding, or it could simply suggest an underpowered (and consequently poorly designed) experiment or study. Ensuring adequate statistical power *before* one engages in a research study or experiment is mandatory.

3.20.2 Power Estimation Using R and G*Power

To demonstrate the estimation of power using software, we first use `pwr.r.test` (Champely, 2014) in R to estimate required sample size for a Pearson *r* correlation coefficient. As an example, we estimate required sample size for a population correlation coefficient of $\rho = 0.10$ at a significance level set to 0.05, with desired power equal to 0.90. Note that in the code that follows, we purposely leave *n* empty so that R can estimate this figure for us:

```
> install.packages("pwr")
> library(pwr)
> pwr.r.test(n =, r = .10, sig.level = .05, power = .90)

     approximate correlation power calculation (arctangh transformation)

                n = 1046.423
```

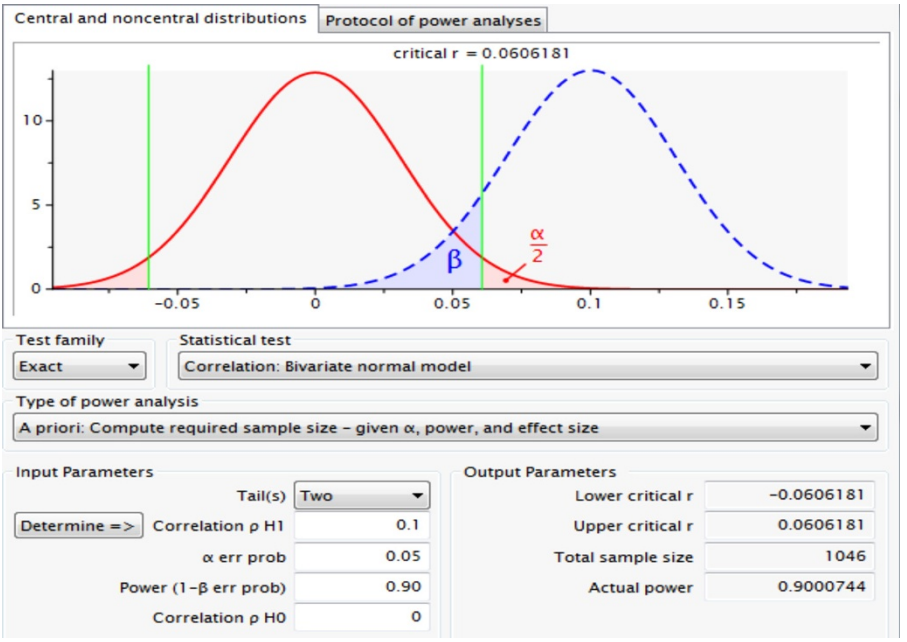


FIGURE 3.13 G*Power output for estimating required sample size for $r = 0.10$.

```
          r = 0.1
  sig.level = 0.05
      power = 0.9
alternative = two.sided
```

We see that to detect a correlation coefficient of 0.10 at a desired level of power equal to 0.9, a sample size of 1046 is required. We could round up to 1047 for a slightly more conservative estimate. Estimating in G*Power,[13] we obtain that given in Figure 3.13.

Note that our power estimate using G*Power is identical to that using R (i.e., power of 0.90 requires a sample size of 1046 for an effect size of $\rho = 0.10$). G*power also allows us to draw the corresponding power curve. A power curve is a simple depiction of required sample size as a function of power and estimated effect size. What is nice about power curves is that they allow one to see how estimated sample size requirements and power *increase* or *decrease* as a function of effect size. For the

[13]G*Power is a user-friendly statistical power program that can be downloaded for free from http://www.gpower.hhu.de/en.html.

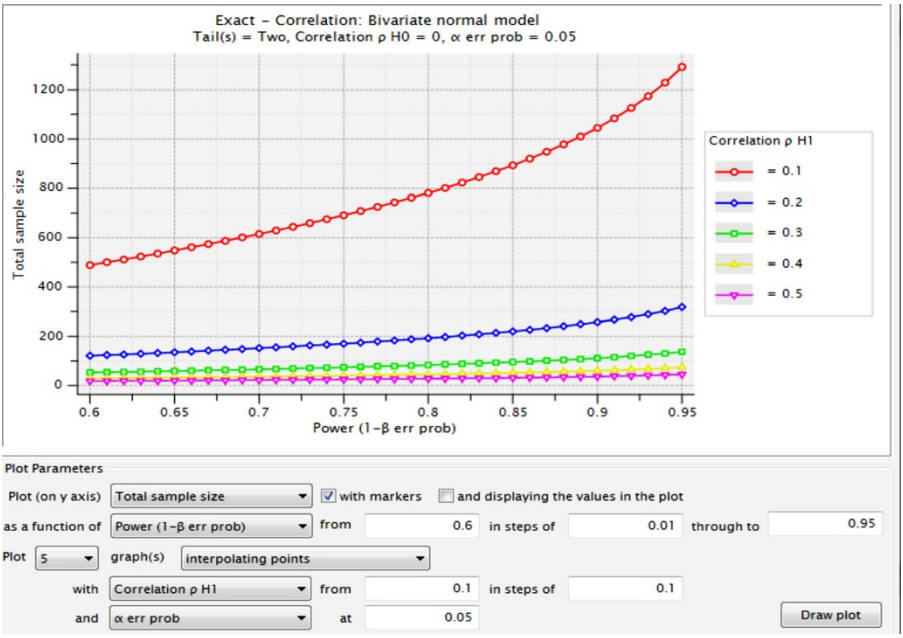


FIGURE 3.14 Power curves generated by G*Power for detecting correlation coefficients of $\rho = 0.10$ to 0.50.

estimation of sample size for detecting $\rho = 0.10$, G*Power generates the curve in Figure 3.14 (top curve).

Especially for small hypothesized values of ρ, the required sample size for even poor to modest levels of statistical power is quite large. For example, reading off the plot in Figure 3.14, to detect $\rho = 0.10$, at even a relatively low power level of 0.60, one requires upward of almost 500 participants. This might explain why many studies that yield relatively small effect sizes never get published. They often have insufficient power to reject their null hypotheses. As effect size increases, required sample size drops substantially. For example, to attain a modest level of power such as 0.69 for a correlation coefficient of 0.5, one requires only 22.5 participants, as can be more clearly observed from Table 3.6 which corresponds to the power curves in Figure 3.14 for power ranging from 0.60 to 0.69.

Hence, one general observation from this simple power analysis for detecting ρ is that the *size of effect* (in this case, ρ) plays a very important role in determining estimated sample size. As a general rule, across virtually all statistical tests, *if the effect you are studying is large, a much smaller sample size is required than if the effect is weak*. Drawing on our analogy of the billboard sign that reads "H_0 is false," all else equal, if the sign is in large print (i.e., strong effect), you require less "power" in your prescription glasses to detect such a large sign. If the sign is in small print (i.e., weak effect), you require much more "power" in your lenses to detect it.

TABLE 3.6 Power Estimates as a Function of Sample Size and Estimated Magnitude under Alternative Hypothesis

Exact - Correlation: Bivariate normal model
Tail(s) = Two, Correlation ρ H0 = 0, α err prob = 0.05

		Correlation ρ H1 = 0.1	Correlation ρ H1 = 0.2	Correlation ρ H1 = 0.3	Correlation ρ H1 = 0.4	Correlation ρ H1 = 0.5
#	Power (1-β err prob)	Total sample size	Total sample size	Total sample size	Total sample size	Total sample size
1	0.600000	488.500	121.500	53.5000	29.5000	18.5000
2	0.610000	500.500	124.500	54.5000	30.5000	18.5000
3	0.620000	511.500	126.500	55.5000	30.5000	19.5000
4	0.630000	523.500	129.500	56.5000	31.5000	19.5000
5	0.640000	535.500	132.500	58.5000	32.5000	19.5000
6	0.650000	548.500	135.500	59.5000	32.5000	20.5000
7	0.660000	561.500	138.500	60.5000	33.5000	20.5000
8	0.670000	574.500	142.500	62.5000	34.5000	21.5000
9	0.680000	587.500	145.500	63.5000	34.5000	21.5000
10	0.690000	601.500	148.500	64.5000	35.5000	22.5000

3.20.3 Estimating Sample Size and Power for Independent Samples *t*-Test

For an independent samples *t*-test, required sample size can be estimated through R using `pwr.t.test`:

```
> pwr.t.test (n =, d =, sig.level =, power =, type = c("two.sample",
"one.sample", "paired"))
```

where `n` = sample size per group, `d` = estimate of standardized statistical distance between means (Cohen's d), `sig.level` = desired significance level of the test, `power` = desired power level, and `type` = designation of the kind of *t*-test you are performing (for our example, we are performing a two-sample test).

It would be helpful at this point to translate Cohen's d values into R^2 values to learn how much variance is explained by differing d values. To convert the two, we apply the following transformation:

$$\mathbf{d} = \sqrt{\frac{4r^2}{1-r^2}}$$

Table 3.7 contains conversions for r increments of 0.10, 0.20, 0.30, etc.

To get a better feel for the relationship between Cohen's d and r^2, we obtain a plot of their values (Figure 3.15).

As can be gleamed from Figure 3.15, the relationship between the two effect size measures is not exactly linear, and increases rather sharply for comparatively large values (the curve is somewhat exponential).

TABLE 3.7 [14]**Conversions for $r \rightarrow r^2 \rightarrow$ d**

r	r^2	**d**
0.10	0.01	0.20
0.20	0.04	0.41
0.30	0.09	0.63
0.40	0.16	0.87
0.50	0.25	1.15
0.60	0.36	1.50
0.70	0.49	1.96
0.80	0.64	2.67
0.90	0.81	4.13
0.99	0.98	14.04

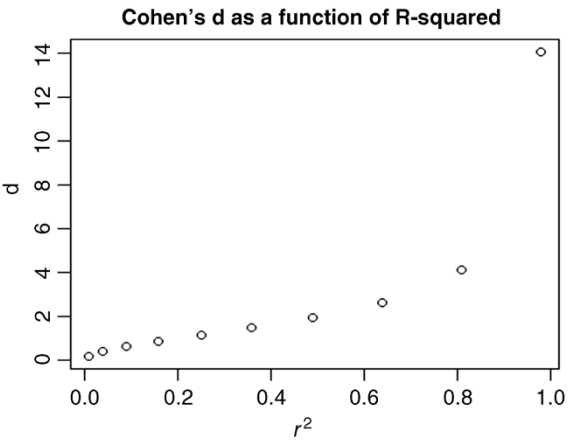


FIGURE 3.15 Relationship between Cohen's d and R-squared.

[14]Entries in the table were computed by the author in R as follows:

```
> r <- c(0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99)
> r_squared <- r^2
> r_squared
 [1] 0.0100 0.0400 0.0900 0.1600 0.2500 0.3600 0.4900 0.6400 0.8100 0.9801

> d <- sqrt((4*r^2)/(1-r^2))
> d
 [1] 0.2010076 0.4082483 0.6289709 0.8728716 1.1547005 1.5000000
 [7] 1.9603921 2.6666667 4.1294832 14.0358479
```

Suppose a researcher would like to estimate required sample size for a two-sample *t*-test, for a relatively small effect size, d = 0.41 (equal to *r* of 0.20), at a significance level of 0.05, with a desired power level of 0.90. We compute:

```
> pwr.t.test (n =, d =0.41, sig.level =.05, power =.90, type = c
("two.sample"))

   Two-sample t test power calculation

              n = 125.9821
              d = 0.41
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Thus, the researcher would require a sample size of approximately 126. As R emphasizes, this sample size is *per group*, so the *total* sample size required is 126(2) = 252.

3.21 PAIRED SAMPLES *t*-TEST: STATISTICAL TEST FOR MATCHED PAIRS (ELEMENTARY BLOCKING) DESIGNS

Oftentimes in research we are able to sample observations that are *matched* on one or more variables or characteristics. For instance, consider the hypothetical data in Table 3.8:

- In each block (1–5), participants *within* blocks are assumed to be more *homogeneous* on one or more variables than participants *between* blocks.
- Participants are randomly assigned to condition (i.e., treatment 1 versus treatment 2) within each block.
- Whether the blocks are naturally occurring or our sampling scheme is designed purposely to create the blocks, we can exploit the homogeneity of participants

TABLE 3.8 Matched Pairs Design

	Treatment 1	Treatment 2
Block 1	10	8
Block 2	15	12
Block 3	20	14
Block 4	22	15
Block 5	25	24

within each block by including this source in our statistical analysis as to potentially reduce the error term of our statistical test.

The matched pairs design is a simpler version of the full-blown *completely randomized block design* in which one can have more than just two levels of the independent variable (e.g., treatment 1 versus treatment 2 versus treatment 3). However, the principle behind the matched pairs design and that of randomized block designs is the same, that of exploiting the covariance between conditions and removing it from the error term of the test statistic (*t* in matched pairs; *F* in randomized block designs).

In more advanced analyses such as repeated measures, longitudinal, and mixed effects modeling, we will say that subjects are *nested* within block. A nesting structure simply implies that subjects within a block share similarity compared to subjects between blocks. Good statistical analyses will attempt to account for this similarity, remove it from respective error terms for tests, and hence make the statistical test for effects more sensitive (i.e., more powerful).

As an example of a matched pairs situation, suppose we are interested in evaluating the effects of melatonin[15] dose on average hours of sleep. However, we know that due to age, some people will naturally sleep longer than others irrespective of how much melatonin they receive. We do not want this natural sleep tendency due to age to confound the effect we are actually interested in studying (i.e., that of melatonin dose), and so we will match participants on their age level, or perhaps even crudely on *age group* (e.g., young, middle-aged, old), and carry out our study *within each age group*. Then, when we perform statistical analyses, we will be able to extract this variation due to age out of the error term of the analysis, and hence boost statistical power for estimating the effect we are actually interested in (melatonin dosage).

When we sample observations in pairs, as was true for the independent samples *t*-test the expectation of the difference between sample means is given by

$$E(\bar{y}_1 - \bar{y}_2) = \mu_1 - \mu_2$$

However, because observations are sampled (or "matched") in pairs, we naturally expect there to be a covariance different from zero between pairs. We can exploit this covariance and remove it from the error term of our statistical test. As given in Hays (1994) and Winer et al. (1991), the variance of the difference becomes

$$\sigma^2_{\text{diff}} = \sigma^2_{\text{M}_1} + \sigma^2_{\text{M}_2} - 2\,\text{cov}\,(\bar{y}_1, \bar{y}_2)$$

with standard error equal to

$$\sigma_{\text{diff}} = \sqrt{\sigma^2_{\text{diff}}} = \sqrt{\sigma^2_{\text{M}_1} + \sigma^2_{\text{M}_2} - 2\,\text{cov}\,(\bar{y}_1, \bar{y}_2)}$$

[15]Melatonin is sometimes used as a nonprescription sleep aid.

Notice that we have subtracted $2\,\mathrm{cov}(\bar{y}_1, \bar{y}_2)$ from the denominator of our statistic. Assuming the covariance between pairs is unequal to 0 and positive, this subtraction will serve to *lower* the standard error of our statistic, and hence boost statistical power. In practice, this is accomplished by conducting a *t*-test on the difference scores between samples.

In the classic between-subjects design where participants are not matched, the expectation is that covariance between treatments is equal to 0, and hence we would have

$$\begin{aligned}\sigma^2_{\text{diff}} &= \sigma^2_{\text{M}_1} + \sigma^2_{\text{M}_2} - 2\,\mathrm{cov}\,(\bar{y}_1, \bar{y}_2) \\ &= \sigma^2_{\text{M}_1} + \sigma^2_{\text{M}_2} - 2(0) \\ &= \sigma^2_{\text{M}_1} + \sigma^2_{\text{M}_2}\end{aligned}$$

The matched pairs design is a very important concept in statistics and design of experiments, because this simple design is the starting point to understanding more complicated designs and modeling such as mixed effects and hierarchical models.

We analyze the hypothetical data in Table 3.8 using a paired samples *t*-test in R by requesting `paired = TRUE`:

```
> treat <- c(10, 15, 20, 22, 25)
> control <- c(8, 12, 14, 15, 24)
> t.test(treat, control, paired = TRUE)

    Paired t-test

data: treat and control
t = 3.2827, df = 4, p-value = 0.03042
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5860324 7.0139676
sample estimates:
mean of the differences
                    3.8
```

The obtained *p*-value of 0.03 is statistically significant at a 0.05 level of significance. We reject the null hypothesis and conclude the population means for the treatment conditions to be different.

As a nonparametric test, the Wilcoxon rank–sum test discussed earlier can be adapted to incorporate paired observations. For our data, we have:

```
> wilcox.test(treat, control, paired = TRUE)

       Wilcoxon signed rank test

data: treat and control
V = 15, p-value = 0.0625
alternative hypothesis: true location shift is not equal to 0
```

TABLE 3.9 Randomized Block Design

	Treatment 1	Treatment 2	Treatment 3
Block 1	10	9	8
Block 2	15	13	12
Block 3	20	18	14
Block 4	22	17	15
Block 5	25	25	24

We notice that the obtained p-value is somewhat greater for the nonparametric test than for the parametric one. In terms of significance tests, this emphasizes the fact that there is usually a cost to not being able to make parametric assumptions.

3.22 BLOCKING WITH SEVERAL CONDITIONS

We have said that in a blocking design, between treatment conditions we expect the covariance to be unequal to 0. Now, consider a design in which, once again we block, but this time on more than two treatment levels. The layout for such a design is given in Table 3.9.

Now, here is the trick to understanding advanced modeling, including a primary feature of mixed effects modeling. We know that we expect the covariance between treatments to be unequal to 0. This is analogous to what we expected in the simple matched pairs design. It seems then that a reasonable assumption to make for the data in Table 3.9 is that the covariances between treatments are *equal*, or at minimum follow some hypothesized correlational structure. In multilevel and hierarchical models, attempts are made to account for the correlation between treatment levels instead of assuming these correlations to equal 0 as is the case for classical between-subjects designs. In Chapter 7, we elaborate on these ideas when we discuss randomized block and repeated measures models.

3.23 COMPOSITE VARIABLES: LINEAR COMBINATIONS

In many statistical techniques, especially multivariate ones, statistical analyses take place not on individual variables, but rather on *linear combinations* of variables. A linear combination in scalar algebra can be denoted simply as

$$\begin{aligned} \ell_i &= a_1 y_1 + a_2 y_2 + \cdots + a_p y_p \\ &= \mathbf{a}'\mathbf{y} \end{aligned}$$

where $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$. These values are scalars, and serve to weight the respective values of y_1 through y_p.

Just as we did for "ordinary" variables, we can compute a number of central tendency and dispersion statistics on linear combinations. For instance, we can compute the mean of a linear combination ℓ_i as

$$\overline{\ell} = \frac{1}{n}\sum_{i=1}^{n} \ell_i = \mathbf{a}'\overline{\mathbf{y}}$$

We can also compute the sample variance of a linear combination:

$$s_\ell^2 = \frac{\sum_{i=1}^{n} (\ell_i - \overline{\ell})^2}{n-1} = \mathbf{a}'\mathbf{S}\mathbf{a}$$

for $\ell_i = \mathbf{a}'\mathbf{y}_i, i = 1, 2, \ldots, n$, and where $\mathbf{S}$ is the sample covariance matrix. Though the form $\mathbf{a}'\mathbf{S}\mathbf{a}$ for the variance may be difficult to decipher at this point, it will become clearer when we consider techniques such as principal components later in the book.

For two linear combinations,

$$\ell_1 = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p = \mathbf{a}'\mathbf{y}$$

and

$$\ell_2 = b_1 y_1 + b_2 y_2 + \cdots + b_p y_p = \mathbf{b}'\mathbf{y}$$

we can obtain the sample covariance between such linear combinations as follows:

$$\text{cov}_{\ell_1,\ell_2} = \frac{\sum_{i=1}^{n} (\ell_{i1} - \overline{\ell}_1)(\ell_{i2} - \overline{\ell}_2)}{n-1} = \mathbf{a}'\mathbf{S}\mathbf{b}$$

The correlation of these linear combinations is simply the standardized version of $\text{cov}_{\ell_1,\ell_2}$:

$$r_{\ell_1,\ell_2} = \frac{\text{cov}_{\ell_1,\ell_2}}{\sqrt{s_{\ell_1}^2 s_{\ell_2}^2}} = \frac{\mathbf{a}'\mathbf{S}\mathbf{b}}{\sqrt{(\mathbf{a}'\mathbf{S}\mathbf{a})(\mathbf{b}'\mathbf{S}\mathbf{b})}}$$

If r_{ℓ_1,ℓ_2} is the *maximum* correlation between linear combinations, it is called the *canonical correlation*, discussed in Chapter 13. The correlation between linear combinations plays a central role in multivariate analysis. Substantively, and geometrically, linear combinations can be interpreted as "projections" of one or more variables onto new dimensions. For instance, in simple linear regression, the fitting of a least-squares line is such a projection. It is the projection of points such that it

guarantees that the sum of squared deviations from the given projected line or "surface" (in the case of higher dimensions) is kept to a minimum.

If we can assume multivariate normality of a distribution, that is, $\mathbf{Y} \sim \mathbf{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, then we know linear combinations of $\mathbf{Y}$ are also normally distributed, as well as a host of other useful statistical properties (Timm, 2002, pp. 86–88). In multivariate methods especially, we regularly need to make assumptions about such linear combinations, and it helps to know that so long as we can assume multivariate normality, we have some idea of how such linear combinations will be distributed.

3.24 MODELS IN MATRIX FORM

Throughout the book, our general approach is to first present models in their simplest possible form using only scalars. We then gently introduce the reader to the corresponding matrix counterparts and extensions. The requirement of matrices for such models is to accommodate numerous variables and dimensions. Matrix algebra is the vehicle by which multivariate analysis is communicated, though most of the *concepts* of statistics can be communicated using simpler scalar algebra.

As an example of how matrices will be used to develop more complete and general models, consider the multivariate general linear model in matrix form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{3.8}$$

where $\mathbf{Y}$ is an $n \times m$ matrix of n observations on m response variables, $\mathbf{X}$ is the model matrix whose columns contain k regressors that include the intercept term, $\mathbf{B}$ is a matrix of regression coefficients, and $\mathbf{E}$ is a matrix of errors. Many statistical models can be incorporated into the framework of (3.8). As a relatively easy application of this general model, consider the simple regression model (discussed in Chapter 9) in matrix form:

$$\mathbf{Y} = \begin{bmatrix} y_{i=1} \\ y_{i=2} \\ y_{i=3} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_{i=n} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{i=1} \\ 1 & x_{i=2} \\ 1 & x_{i=3} \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_{i=n} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $y_{i=1}$ to $y_{i=n}$ are observed measurements on some dependent variable, $\mathbf{X}$ is the model matrix containing a constant of 1 in the first column to represent the common intercept term, $x_{i=1}$ to $x_{i=n}$ are observed values on a predictor variable, α is the fixed intercept term, β is the slope parameter, which we also assume to be fixed, and $\boldsymbol{\epsilon}$ is a vector of errors ϵ_1 to ϵ_n.

Suppose now we want to add a second response variable. Because of the generality of (3.8), this can be easily accommodated:

$$\mathbf{Y} = \begin{bmatrix} y_{i=1,1}, y_{i=1,2} \\ y_{i=2,1}, y_{i=2,2} \\ y_{i=3,1}, y_{i=3,2} \\ \vdots \\ \vdots \\ \vdots \\ y_{i=n,1}, y_{i=n,2} \end{bmatrix}$$

where now, a second response variable is represented in $\mathbf{Y}$ by a second column. That is, $y_{i=1,2}$ corresponds to individual 1 on response variable 2, $y_{i=2,2}$ is individual 2 on response variable 2, etc. Such matrix representations will be featured throughout the book.

3.25 GRAPHICAL APPROACHES

Performing inferential tests to help in drawing conclusions about population parameters is useful, but ultimately the findings of a statistical analysis should make their way into a graph or other visualization. *Data visualization* is a field in itself, and with the advent of modern computing power, possibilities exist today that could only be dreamt of in the past. Simple visualizations such as *histograms*, *boxplots*, *scatterplots*, etc., can be useful in depicting findings but also in helping to verify assumptions that underlay the statistical model one is using. For example, since many tests of normality and equality of variances (and covariances) are relatively sensitive to the type of data to which they are applied, oftentimes researchers will generate simple plots in order to detect potential gross violations of such assumptions. We feature such techniques throughout the book.

For graphical displays meant to communicate findings (rather than test assumptions), Friendly (2000) puts the field into context:

> Designing good graphics is surely an art, but as surely, it is one that ought to be informed by science . . . In this view, an effective graphical display, like good writing, requires an understanding of its *purpose* – what aspects of the data are to be communicated to the viewer. In writing, we communicate most effectively when we know our audience and tailor the message appropriately. (p. 8)

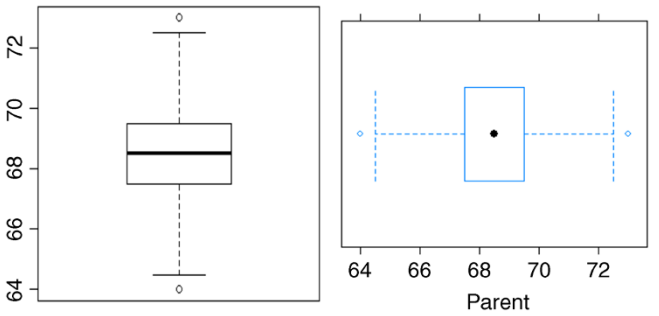
In high-dimensional space, the challenge of graphical approaches is to summarize data into lower dimensions, while still retaining most of the information in the original data. We feature some such plots in later chapters. For a thorough account of data visualization, see DataVis.ca (Friendly, 2014, www.datavis.ca). For sophisticated graphics using R, refer to Wickham (2009).

For now, it is useful here to briefly review some basic plots for which the reader is likely already familiar.

3.25.1 Box-and-Whisker Plots

The boxplot was a contribution of Tukey (1977) in the spirit of what he called *exploratory data analysis* (EDA), which encouraged scientists to spend more of their energy on descriptive techniques instead of focusing exclusively on confirmatory statistical tests. Boxplots of parent heights from Galton's data appear below:

```
> attach(Galton)
> boxplot(parent)
> library(lattice)
> bwplot(parent)
```



The boxplot provides what is generally known as a five-number summary of a distribution, of which we can obtain most of the numbers we need by the `summary` function in R:

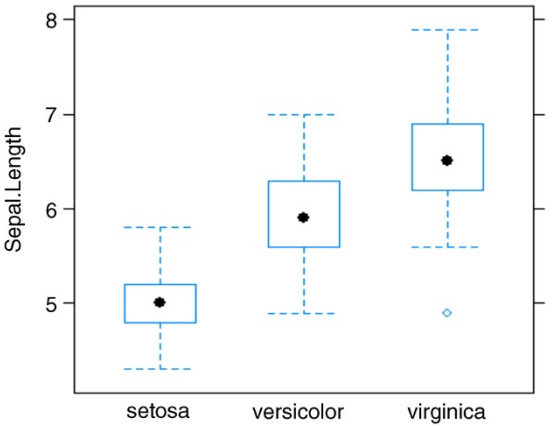
```
> summary(parent)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  64.00     67.50     68.50     68.31     69.50     73.00
```

Recall that the median is the point in the ordered data that divides the data set into two equal parts. The location of the median is computed by $(n+1)/2$. In Galton's data, there are 928 observations, and so the location of the median is at the 464.5th (i.e., $(928+1)/2$) point in the ordered data set. For parent, this value is equal to 68.50. The first and third quartiles represent the 25th and 75th percentiles and are 67.50 and 69.50, respectively. We can also compute the range as

```
> range(parent)
[1] 64 73
```

We can also generate boxplots by category. Throughout the book we use Fisher's *Iris* data (Fisher, 1936) in which flower characteristics such as sepal and petal length are categorized by species of flower. We plot sepal length by species:

```
> library(lattice)
> attach(iris)
> bwplot(Sepal.Length ~ Species)
```



Data points falling beyond the whiskers of the plots may reveal the presence of outliers, and should be investigated (although not necessarily *deleted*, see Section 8.23 for a discussion).

Stem-and-leaf plots are also easily produced. These visual displays are kind of "naked histograms," because they reveal the actual observations in the data while also providing information about their frequency of occurrence. In 1710, John Arbuthnot analyzed data on the ratios of males to female births in London from 1629 to 1710 and in so doing made an argument for these births being a function of a "divine being" (Arbuthnot, 1710). One of his variables was the number of male christenings (i.e., baptisms) over the period 1629–1710. We generate a stem-and-leaf plot in R using package `aplpack` (Wolf and Bielefeld, 2014) of these male christenings, for which the "leaves" are corresponding hundreds. For example, in the following plot, the first value of 2|8 seemingly corresponds to a value of 2800, which appears rounded down from the actual value in the data (which is also the minimum) of 2890. The maximum in the data is actually equal to 8426, but is seemingly represented by 8400 (i.e., 8|0012334):

```
> library(aplpack)
> attach(Arbuthnot)
> stem.leaf(Males)

1 | 2: represents 1200
 leaf unit: 100
```

```
          n: 82
  1     2.|8
 10     3*|011222334
 15     3.|66777
 18     4*|014
 25     4.|6777899
 36     5*|01112233444
 38     5.|56
(11)    6*|00001122444
 33     6.|5555899
 26     7*|244
 23     7.|5555666666778999
  7     8*|0012334
```

3.26 WHAT MAKES A p-VALUE SMALL? A CRITICAL OVERVIEW AND SIMPLE DEMONSTRATION OF NULL HYPOTHESIS SIGNIFICANCE TESTING

The workhorse for establishing statistical evidence in the social and natural sciences is the method of *null hypothesis significance testing* (NHST). However, since its inception with R.A. Fisher in the early 1900s, the significance test has been the topic of much debate, both statistical and philosophical. Throughout much of this book, NHST is regularly used to evaluate null hypotheses in methods such as the analysis of variance, regression, and various multivariate procedures. Indeed, the procedure is universally used in most statistical methods.

It behooves us then, before embarking on all of these methodologies, to discuss the nature of the null hypothesis significance test, and clearly demonstrate what it actually *means*, not only in a statistical context but also how it should be interpreted in a *research* or *substantive* context.

The purpose of this final section of this chapter is to provide a clear and concise demonstration and summary of the factors that influence the size of a computed p-value in virtually every statistical significance test. Understanding why statements such as "$p < 0.05$" can be reflective of even the smallest and trivial of effects is critical for the practitioner or researcher to appreciate if he or she is to assess and appraise statistical evidence in an intelligent and thoughtful manner. It is not an exaggeration to say that *if one does not understand the makeup of a p-value and the factors that directly influence its size, one cannot properly evaluate statistical evidence, nor should one even make the attempt to do so*. Though these arguments are not new and have been put forth by even the very best of methodologists (Cohen, 1990; Meehl, 1978), there is evidence to suggest that many practitioners and researchers do not understand the factors that determine the size of a p-value.

3.26.1 Null Hypothesis Significance Testing: A History of Criticism

Criticism targeted against null hypothesis significance testing has inundated the literature since the time Berkson in 1938 brought to light how statistical significance can be easily achieved by simple manipulations of sample size:

> I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. (Berkson, 1938, p. 526)

Since Berkson, the very best and renowned of methodologists have remarked that the significance test is subject to gross misunderstanding and misinterpretation (Bakan 1966; Carver, 1993; Cohen, 1990; Estes, 1997; Harlow, Mulaik, and Steiger 1997; Loftus, 1991; Meehl, 1978; Oakes, 1986; Shrout, 1997; Wilson, Miller, and Lower, 1967). And although it can be difficult to assess or evaluate whether the situation has improved, there is evidence to suggest that it has not. Few describe the problem better than Gigerenzer in his article "Mindless Statistics" (Gigerenzer, 2004), in which he discusses both the roots and truths of hypothesis testing, as well as how its "statistical rituals" and practices have become far more of a *sociological* phenomenon rather than anything related to good science and statistics.

Other researchers have found that misinterpretations and misunderstandings about the significance test are widespread not only among students but also among their instructors (Haller and Krauss, 2002). What determines statistical significance and what is it a function of? This is an extremely important question. An unawareness of the determinants of statistical significance leaves the door open to misunderstanding and misinterpretation of the test, and the danger to potentially draw false conclusions based on its results. Too often and for too many, the finding "$p < 0.05$" simply denotes a "good thing" of sorts, without ever being able to pinpoint what is so "good" about it.

Recall the familiar one-sample z-test for a mean discussed earlier:

$$z_{\mathrm{M}} = \frac{\overline{y} - \mu_0}{\sigma / \sqrt{n}}$$

where the purpose of the test was to compare an obtained sample mean $\overline{y}$ to a population mean μ_0 under the null hypothesis that $\mu = \mu_0$. Recall that σ is the standard deviation of the population from which the sample was presumably drawn. Recall that in practice, this value is rarely if ever known for certain, which is why in most cases an estimate of it is obtained in the form of a sample standard deviation s. What determines the size of z_{M}, and therefore, the smallness of p? There are three inputs that determine the size of p, which we have already featured in our earlier discussion of statistical power. These three factors are $\overline{y} - \mu_0$, σ, and n. We consider each of these once more, then provide simple arithmetic demonstrations to emphasize how changing any one of these necessarily results in an arithmetical change in z_{M}, and consequently, a change in the observed p-value.

As a first case, consider the distance $\overline{y} - \mu_0$. *Given constant values of* σ *and n, the greater the distance between* $\overline{y}$ *and* μ_0*, the larger* z_M will be. That is, as the numerator $\overline{y} - \mu_0$ grows larger, the resulting z_M also gets larger in size, which as a consequence, decreases p in size. As a simple example, assume for a given research problem that σ is equal to 20 and n is equal to 100. This means that the standard error is equal to $20/\sqrt{100}$, which is equal to $20/10 = 2$. Suppose the obtained sample mean $\overline{y}$ were equal to 20, and the mean under the null hypothesis μ_0 were equal to 18. The numerator of z_M would thus be $20 - 18 = 2$. When 2 is divided by the standard error of 2, we obtain a value for z_M of 1.0, which is not statistically significant at $p < 0.05$.

Now, consider the scenario where the standard error of the mean remains the same at 2, but that instead of the sample mean $\overline{y}$ being equal to 20, it is equal to 30. The difference between the sample mean and the population mean is thus $30 - 18 = 12$. This difference represents a greater distance between means, and presumably, would be indicative of a more "successful" experiment or study. Dividing 12 by the standard error of 2 yields a z_M value of 6.0, which is highly statistically significant at $p < 0.05$ (whether for a one- or two-tailed test).

Having the value of z_M increase as a result of the distance between $\overline{y}$ and μ_0 increasing is of course what we would expect from a test statistic if that test statistic is to be used in any sense to evaluate the strength of the *scientific* evidence against the null. That is, if our obtained sample mean $\overline{y}$ turns out to be very different than the population mean under the null hypothesis, μ_0, we would hope that our test statistic would measure this effect, and allow us to reject the null hypothesis at some preset significance level (in our example, 0.05). If interpreting test statistics were always as easy as this, there would be no misunderstandings about the meaning of statistical significance and the misguided decisions to automatically attribute "worth" to the statement "$p < 0.05$." However, as we discuss in the following cases, there are other ways to make z_M big or small that do not depend so intimately on the distance between $\overline{y}$ and μ_0, and this is where interpretations of the significance test usually run awry.

Consider the case now for which the distance between means, $\overline{y} - \mu_0$ is, as before, equal to 2 (i.e., 20 − 18 = 2.0). As noted, with a standard error also equal to 2.0, our computed value of z_M came out to be 1.0, which was not statistically significant. However, is it possible to increase the size of z_M without changing the observed distance between means? Absolutely. Consider what happens to the size of z_M as we change the magnitude of either σ or n, or both. First, we consider how z_M is defined in part as a function of σ. For convenience, we assume a sample size still of $n = 100$. Consider now three hypothetical values for σ: 2, 10, and 20. Performing the relevant computations, observe what happens to the size of z_M in the case where $\sigma = 2$:

$$z_M = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{2/\sqrt{100}} = \frac{2}{0.2} = 10$$

The resulting value for z_M is quite large at 10. Consider now what happens if we increase σ from 2 to 10:

$$z_M = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{10/\sqrt{100}} = \frac{2}{1} = 2$$

Notice that the value of z_M has decreased from 10 to 2. Consider now what happens if we increase σ even more to a value of 20 as we had originally:

$$z_M = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{20/\sqrt{100}} = \frac{2}{2} = 1$$

When $\sigma = 20$, the value of z_M is now equal to 1, which is no longer statistically significant at $p < 0.05$. Be sure to note that the distance between means $\bar{y} - \mu_0$ has remained constant. In other words, and this is important, z_M *did not decrease in magnitude by altering the actual distance between the sample mean and the population mean, but rather decreased in magnitude only by a change in* σ.

What this means is that given a constant distance between means $\bar{y} - \mu_0$, whether or not z_M will or will not be statistically significant can be manipulated by changing the value of σ. Of course, a researcher would never arbitrarily manipulate σ directly. The way to decrease σ would be to sample from a population with less variability. The point is that decisions regarding whether a "positive" result occurred in an experiment or study should not be solely a function of whether one is sampling from a population with small or large variance!

Suppose now we again assume the distance between means $\bar{y} - \mu_0$ to be equal to 2. We again set the value of σ at 2. With these values set and assumed constant, consider what happens to z_M as we increase the sample size n from 16 to 49 to 100. We first compute z_M assuming a sample size of 16:

$$z_M = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{2/\sqrt{16}} = \frac{2}{0.5} = 4$$

With a sample size of 16, the computed value for z_M is equal to 4. When we increase the sample size to 49, again, *keeping the distance between means constant, as well as the population standard deviation constant*, we obtain

$$z_M = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{2/\sqrt{49}} = \frac{2}{0.29} = 6.9$$

We see that the value of z_M has increased from 4 to 6.9 as a result of the larger sample size. If we increase the sample size further, to 100, we get

$$z_M = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{20 - 18}{2/\sqrt{100}} = \frac{2}{0.2} = 10$$

and see that as a result of the even larger sample size, the value of z_M has increased once again, this time to 10. Again, we need to emphasize that the observed increase in z_M is occurring not as a result of changing values for $\bar{y} - \mu_0$ or σ, as these values remained constant in our above computations. Rather, *the magnitude of z_M increased as a direct result of an increase in sample size, n, alone*. In many research studies, the achievement of a statistically significant result may simply be indicative that the researcher gathered a minimally sufficient sample size that resulted in z_M falling in the tail of the z distribution. In other cases, the failure to reject the null may in reality simply indicate that the investigator had insufficient sample size. The point is that unless one knows how n can directly increase or decrease the size of a p-value, one cannot be in a position to understand, in a scientific sense, what the p-value actually means, or intelligently evaluate the statistical evidence before them.

3.26.2 The Makeup of a p-Value: A Brief Recap and Summary

The simplicity of these demonstrations is surpassed only by their profoundness. In our simple example of the one-sample z-test for a mean, we have demonstrated that the size of z_M is a direct function of three elements: (1) distance $\bar{y} - \mu_0$, (2) population standard deviation σ, and (3) sample size n. A change in *any* of these while holding the others constant will necessarily, through nothing more than the consequences of how the significance test is constructed and functionally defined, result in a change in the size of z_M. The implication of this is that one can make z_M as small or as large as one would like by choosing to do a study or experiment such that the combination of $\bar{y} - \mu_0$, σ, and n results in a z_M value that meets or exceeds a preselected criteria of statistical significance.

The important point here is that a large value of z_M does not necessarily mean something of any *practical* or *scientific* significance occurred in the given study or experiment. This fact has been reiterated countless times by the best of methodologists, yet too often researchers fail to emphasize this extremely important truth when discussing findings:

> A p-value, no matter how small or large, does not necessarily equate to the success or failure of a given experiment or study.

Too often a statement of "$p < 0.05$" is recited to an audience with the implication that somehow this necessarily constitutes a "scientific finding" of sorts.

3.26.3 The Issue of Standardized Testing: Are Students in Your School Achieving More Than the National Average?

To demonstrate how adjusting the inputs to z_M can have a direct impact on the obtained p-value, consider the situation in which a school psychologist practitioner hypothesizes that as a result of an intensified program implementation in her school, she believes that her school's students, on average, will have a higher achievement mean compared to the national average of students in the same grade. Suppose that the

national average on a given standardized performance test is equal to 100. If the school psychologist is correct that her students are, on average, more advanced performance-wise than the national average, then her students should, on average, score higher than the national mark of 100. She decides to sample 100 students from her school, and obtains a sample achievement mean of $\overline{y} = 101$. Thus, the distance between means is equal to $101 - 100 = 1$. She computes the estimated population standard deviation s equal to 10. Because she is estimating σ^2 with s^2, she computes a one-sample t-test rather than a z-test. Her computation of the ensuing t is

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{101 - 100}{10/\sqrt{100}} = \frac{1}{1} = 1$$

On degrees of freedom equal to $n - 1 = 100 - 1 = 99$, for a two-tailed test, we require a t-statistic of ± 1.984 for the result to be statistically significant at a level of significance of 0.05. Hence, the obtained value of $t = 1$ is not statistically significant. That the result is not statistically significant is hardly surprising, since the sample mean of the psychologist's school is only 101, a single point higher than the national average of 100. It would seem then that the computation of t is telling us a story that is consistent with our intuition, that there is no reason to believe that the school's performance is higher than that of the national average in the population from which these sample data were drawn.

Now, consider what would have happened had the psychologist collected a larger sample, suppose $n = 500$. Using our new sample size, and still assuming an estimated population standard deviation s equal to 10 and a distance between means equal to 1, we repeat the computation for t:

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{101 - 100}{10/\sqrt{500}} = \frac{1}{0.45} = 2.22$$

What happened? The obtained value of t increased from 1 to 2.22 *simply as a result of collecting a larger sample*, nothing more. The actual distance between means remained the same ($101 - 100 = 1$). The degrees of freedom for the test have changed, and are now equal to 499 (i.e., $n - 1 = 500 - 1 = 499$). Since our obtained t of 2.22 exceeds critical t, our statistic is deemed statistically significant at $p < 0.05$. What is important to realize is that we did not change the difference between the sample mean $\overline{y}$ and the population mean μ_0, it remained extremely small at only a single achievement point (i.e., $101 - 100 = 1$). Even with the same distance between means, the obtained t of 2.22 and it being statistically significant at $p < 0.05$ now means we will reject the null hypothesis, and infer the alternative hypothesis that $\mu \neq \mu_0$. And because scientists have historically considered the infamous statement "$p < 0.05$" to be automatically and necessarily equivalent to something meaningful or important, the obvious danger is that the rejection of the null hypothesis at $p < 0.05$ is considered by some (or even *most*) a "positive" result. When in reality, the difference in this case is nothing short of *trivial*.

The problem is not that the significance test is not useful and therefore should be banned. The problem is that too few are aware that the statement "$p < 0.05$," in itself, *scientifically* (as opposed to *statistically*) may have little meaning in a given research context, and at worst, may be entirely misleading if automatically assigned any degree of scientific importance by the interpreter.

3.26.4 Other Test Statistics

The factors that influence the size of a p-value are, of course, not only relevant to z- and t-tests, but are also at work in essentially every test of statistical significance we might conduct. For instance, as we will see in the following chapter, the size of the F-ratio in traditional one-way ANOVA is subject to the same influences. Taken as the ratio of MS between to MS error, the three determining influences for the size of p are (1) size of MS between, which is a reflection of the extent to which means are different from group to group, (2) size of MS error, which is in part a reflection of the within-group variability (i.e., analogous to s in the t-test situation) and (3) sample size (when computing MS error, we divide the sum of squares for error by degrees of freedom, in which the degrees of freedom are determined in large part by sample size). Hence, a large F-stat does not necessarily imply that MS between is absolutely large, no more than a large t necessarily implies the size of $\overline{y} - \mu_0$. A small p-value associated with a computed F could be a result of small within-group variation and/or a large sample size. It does not necessarily mean that group-to-group mean differences are substantial.

These ideas for significance tests apply in even the most advanced of modeling techniques, such as structural equation modeling (see Chapter 16). The typical measure of model fit here is the chi-square statistic χ^2, which as reported by many (Bollen, 1989; Hoelter, 1983) suffers the same interpretational problems as t and F regarding how its magnitude can be largely a function of sample size. That is, one can achieve a small or large χ^2 simply because one has used a small or large sample. If a researcher is not aware of this fact, he or she may decide that a model is well-fitting or poor-fitting based on a small or large chi-square value, without awareness of its connection with n. This is in part why other measures, as we will see, have been proposed for interpreting the fit of SEM models (e.g., see Browne and Cudeck, 1993).

3.26.5 The Solution

The solution to episodes of misunderstanding the significance test is not to drop or ban it, contrary to what some have recommended (Hunter, 1997). Rather, the solution is to supplement it with a measure that accounts for the actual distance between means and serves to convey the magnitude of the actual *scientific* finding, as opposed to *statistical* finding, should there be one. Measures of *effect size*, interpreted in conjunction with significance tests, help to communicate whether something has "happened" or "not happened" in the given study or experiment. The reader interested in effect sizes can turn to a multitude of sources (Cortina and Nouri, 1999; Rosenthal,

Rosnow, and Rubin, 2000). For our purposes, it suffices to review the *principle* of an effect size measure rather than catalog the wealth of possibilities for effect sizes available. Perhaps the easiest and most straightforward way of conceptualizing an effect size is to consider a measure of *standardized statistical distance*, or Cohen's d, already featured in our computations of power.

3.26.6 Statistical Distance: Cohen's d

For a one-sample z-test, Cohen's d (Cohen, 1988) is defined as the absolute distance between the observed sample mean and the population mean under the null hypothesis, divided by the population standard deviation:

$$\mathrm{d} = \left| \frac{\bar{y} - \mu_0}{\sigma} \right|$$

As an example, where $\bar{y} = 20$, $\mu_0 = 18$, and $\sigma = 2$, Cohen's d is computed as

$$\begin{aligned} \mathrm{d} &= \left| \frac{20 - 18}{2} \right| \\ &= 1.0 \end{aligned}$$

Cohen offered the guidelines of 0.20, 0.50, and 0.80 as representing small, medium, and large effects, respectively (Cohen, 1988). However, relying on effect size guidelines to indicate the absolute size of an experimental or nonexperimental effect should only be done in the complete and absolute absence of all other information for the research area. In the end, *it is the researcher, armed with knowledge of the history of the phenomenon under study, who must evaluate whether an effect is small or large*. For instance, referring to the achievement example discussed earlier, Cohen's d would be equal to

$$\begin{aligned} \mathrm{d} &= \left| \frac{101 - 100}{10} \right| \\ &= 0.1 \end{aligned}$$

The effect size of 0.1 is small according to Cohen's guidelines, but more importantly, also small *substantively*, since a difference in means of 1 point is, by all accounts, trivial. In this case, both Cohen's guidelines and the actual substantive evaluation of the size of effect coincide. However, this is not always the case. In physical or biological experiments, for instance, one can easily imagine examples for which an effect size of even 0.8 might be considered "small" relative to the research area under investigation, since the degree of control the investigator can impose over his or her subjects is much greater. In such cases, it may very well be that Cohen's d values in the neighborhood of 3 or 4 would be required for an effect to be considered "large." The point is that only in the complete absence of information regarding an area of investigation is it appropriate to use "rules of thumb" to evaluate the size of effect.

Cohen's d, or effect size measures in general, should always be used in conjunction with statements of statistical significance, since they tell the researcher what she is actually wanting to know, that of the estimated separation between sample data (often in the form of a sample mean) and the null hypothesis under investigation. Oftentimes *meta-analysis*, which is a study of the overall measure of effect for a given phenomenon, can be helpful in comparing new research findings with the status quo in a given field. As an example of a meta-analysis, see Crawley (2013, p. 740).

3.26.7 What Does Cohen's d Actually Tell Us?

Writing out a formula and plugging in numbers unfortunately does not necessarily give us a feeling for what the formula actually means. This is especially true with regard to Cohen's d. We now discuss the statistic in a bit more detail, pointing out why it is usually interpreted as the *standardized difference between means*.

Imagine you have two independent samples of laboratory rats. To one sample, you provide normal feeding and observe their weight over the next 30 days. To the other sample, you also feed normally, but also give them regular doses of a weight-loss drug. You are interested in learning whether your weight-loss drug works or not. Suppose that after 30 days, on average, a mean difference of 0.2 pounds is observed between groups. How big is a difference of 0.2 pounds for these groups? If the average difference in weight among rats in the population were very large, say, 0.8 pounds, then a mean difference of 0.2 pounds is not that impressive. After all, if rats weigh very differently from one rat to the next, then really, finding a mean difference of 0.2 between groups cannot be that exciting. However, if the average weight difference between rats were equal to 0.1 pounds, then all of a sudden, a mean difference of 0.2 pounds seems more impressive, because that size of difference is *atypical* relative to the population. What is "typical"? This is exactly what the *standard deviation* reveals. Hence, when we compute Cohen's d, we are in actuality producing a *ratio of one deviation relative to another*, similar to how when we compute a z-score, we are comparing the deviation of $y - \mu$ with the *standard* deviation σ. The extent to which observed differences are large relative to "average" differences will be the extent to which d will be large in magnitude.

3.26.8 Why and Where the Significance Test Still Makes Sense

At this point, the conscientious reader may very well be asking the following question: *If the significance test is so misleading and subject to misunderstanding and misinterpretation, how does it even make sense as a test of anything? It would appear to be a nonsensical test and should forever be forgotten*. The fact is that the significance test *does* make sense, only that the sense that it makes is not necessarily always *scientific*. Rather, it is *statistical*. To a pure theoretical statistician or mathematician, a decreasing p-value as a function of an increasing sample size makes perfect sense: As we snoop a larger part of the population, the random error we expect from sample to sample necessarily decreases, because with each sample we are obtaining a better estimate of the true population parameter. Hence, that we achieve

statistical significance with a sample size of 500 and not 100, for instance, is well within that of statistical "good sense."

However, *statistical truth does not equate to scientific truth* (Bolles, 1962). Statistical conclusions should never be automatically equated with scientific ones. They are different and distinct *things*. When we arrive at a statistical conclusion (e.g., when deciding to reject the null hypothesis), one can never assume that this represents anything that is necessarily or absolutely scientifically meaningful. Rather, the statistical conclusion should be used as a potential *indicator* that something scientifically interesting *may* have occurred, the evidence for which must be determined by other means, which includes effect sizes, researcher judgment, and putting the obtained result into its proper interpretive context.

3.27 CHAPTER SUMMARY AND HIGHLIGHTS

- To understand advanced statistical procedures, it is necessary to have a firm grasp in the foundations of introductory statistics. Advanced procedures are typically extensions of first principles.
- Densities are theoretical probability distributions. The normal univariate density is an example.
- The standard normal distribution has a mean μ of 0 and a variance σ^2 of 1.
- z-scores are useful for comparing raw scores emanating from different distributions. Standardization transforms raw scores to a common scale, allowing for comparison between scores.
- Binomial distributions are useful in modeling experiments in which the outcome can be conceptualized as a "success" or "failure." The outcome of the experiment must be binary in nature for the binomial distribution to apply.
- The normal distribution can be used to approximate the binomial distribution. In this regard, we say that the limiting form of the binomial distribution is the normal distribution.
- The bivariate normal density expresses the probability of the joint occurrence of two variables.
- The multivariate normal density expresses the probability of the joint occurrence of three or more variables.
- The mean, variance, skewness, and kurtosis are all moments of a distribution.
- The mean, the first moment of a distribution, either of a mathematical variable or of a random variable, can be regarded as the center of gravity of the distribution such that the sum of deviations from the mean for any distribution is equal to zero.
- The variance, the second moment of a distribution, can be computed for either a mathematical variable or a random variable. It expresses the degree to which scores, on average, deviate from the mean in squared units.
- The sample variance with n in the denominator is biased. To correct for the bias, a single degree of freedom is subtracted so that the new denominator is $n - 1$.

- The expectation of the uncorrected version of the sample variance is not equal to σ^2. That is, $E(S^2) \neq \sigma^2$. However, the corrected version of the sample variance (with $n-1$ in the denominator) is equal to σ^2. That is, $E(s^2) = \sigma^2$.
- Skewness, the third moment of a distribution, reflects the extent to which a distribution lacks symmetry.
- Kurtosis, the fourth moment of a distribution, reflects the extent to which a distribution is peaked or flat.
- Covariance and correlation are defined for both empirical variables and random variables. Both measure the extent to which two variables are linearly related. Pearson r is the standardized version of the covariance, and is dimensionless, meaning that its value is not dependent on the variance in each variable. Pearson r ranges from -1 to $+1$ in value.
- In multivariable contexts, covariance and correlation matrices are used in place of single coefficients.
- There are numerous other correlation coefficients available other than Pearson r. One such coefficient is Spearman's r_s, which captures monotonically increasing relationships. Monotonically increasing relationships do not necessarily have to be linear.
- The issue of measurement should be carefully considered before data are collected. S.S. Stevens proposed four scales of measurement: nominal, ordinal, interval, and ratio. The most sophisticated level of measurement is that of the ratio scale where a value of zero on the scale truly means an absence of the attribute under study.
- A random variable is a mathematical variable that is associated with a probability distribution. More formally, it is a function from a sample space into the real numbers.
- An estimator is a function of a sample used to estimate a parameter in the population.
- An interval estimator provides a range of values within which the true parameter is hypothesized to exist.
- An unbiased estimator is one in which its expectation is equal to the corresponding population parameter. That is, $E(T) = \theta$.
- An estimator is consistent if as sample size increases without bound, the variance of the estimator approaches zero.
- An estimator is efficient if it has a relatively low mean squared error.
- An estimator is sufficient for a given parameter if the statistic tells us everything we need to know about the parameter and our knowledge of it could not be improved if we considered additional information (e.g., such as a secondary statistic).
- The concept of a sampling distribution is at the heart of statistical inference. A sampling distribution of a statistic is a theoretical probability distribution of that statistic. It is idealized, and hence not ordinarily empirically derived.

- The sampling distribution of the mean is of great importance because so many of our inferences have to do with means.
- As a result of $E(\bar{y}) = \mu$, we can say that $\mu_{\bar{y}} = \mu$, that is, the mean of all possible sample means we could draw from some specified population is equal to the mean of that population.
- The variance of the sampling distribution of the mean is equal to $1/n$ of the original population variance. That is, it is equal to σ^2/n.
- The square root of the sampling variance for the mean is equal to the standard error, $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.
- The central limit theorem is perhaps the most important theorem in all of statistics. Though there are different forms of the theorem, in general, it states that the sum of random variables approximates a normal distribution as the size upon which each sample is based increases without bound.
- Confidence intervals provide a range of values for which we can be relatively certain to lay the true parameter we are seeking to estimate. Key to understanding confidence intervals is to recognize that it is the sample upon which the interval is computed that is the random component, and not the parameter we are seeking to estimate. The parameter is assumed to be fixed.
- Likelihood ratio tests compare the likelihood of observed data under one hypothesis to the likelihood of observed data under a competing hypothesis.
- Student's t distribution, derived by William Gosset (or "Student") in 1908, is useful when σ^2 is unknown and must be estimated from the sample. Because in the limit $f(t) = f(z)$ (i.e., $\lim_{v \to \infty} f(t) = f(z)$), for large samples, whether one uses z or t will make little difference.
- The t-test for one sample compares an obtained sample mean with a population mean and tests the null hypothesis that the sample mean could have reasonably been drawn from the given population.
- As degrees of freedom increase, the variance of the t distribution approaches 1, which is the same as that for a standardized normal variable. That is, $\lim_{v \to \infty} (v/(v-2)) = 1.0$.
- The t-test for two samples tests the null hypothesis that both samples were selected from the same population. A rejection of the null hypothesis suggests the samples arose from populations with different means.
- Power is the probability of rejecting a null hypothesis given that it is false. It is equal to $1 - \beta$ (i.e., 1 – type II error rate). Power is a function of four elements: (1) hypothesized value under H_1, (2) significance level, or type I error rate, α, (3) variance, σ^2, in the population, and (4) sample size.
- Experiments or studies suffering from insufficient power make it difficult to ascertain why the null hypothesis failed to be rejected.
- Power can be estimated quite easily using R or G*Power.
- The paired samples t-test is useful for matched pairs (elementary blocking) designs.

- The paired samples *t*-test usually results in an increase in statistical power because the covariance between measurements is subtracted from the error term. In general, anything that makes the error term smaller helps to boost statistical power.
- The paired samples *t*-test and the matched design which it serves provide a good entry point into the discussion of the randomized block design, the topic of Chapter 7.
- In multivariable contexts, linear combinations of variables are generated of the form $\ell_i = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p$. Means and variances of linear combinations can be obtained, as well as the covariance and correlation between linear combinations.
- Representing statistical models in matrix form is required in statistical analyses of higher dimensions than 1 (e.g., multiple regression, multivariate analysis of variance, principal components analysis, etc.). The fundamental general linear model can be given by $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$.
- Understanding what makes a *p*-value small or large is essential if a researcher is to intelligently interpret statistical evidence is his or her field. The history of null hypothesis significance testing is plagued with controversy, and a solid understanding of the difference between statistical significance and effect size (e.g., Cohen's d) is necessary before one attempts to interpret any research findings.

REVIEW EXERCISES

3.1. Distinguish between a density and an empirical distribution. How are they different? How are they similar?

3.2. Consider the univariate normal density:

$$f(x_i, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$$

Show that for a standard normal distribution, the above becomes $f(x_i, \mu, \sigma^2) = e^{-(1/2)x_i^2}/\sqrt{2\pi}$.

3.3. Explain the nature of a *z*-score:

$$z = \frac{x_i - \mu}{\sigma}$$

Why is it also called a standardized score?

3.4. Using R, compute the probability of observing a standardized score of 1.0 or greater from a normal distribution. What is then the probability of observing a score less than 1.0 from such a distribution?

3.5. Think up a research example in which the binomial distribution would be useful in testing a null hypothesis.

3.6. Rafael Nadal, a professional tennis player, as of 2014 had won the French Open tennis championship a total of 9 times out of 10 attempts. If we set the probability of his winning each time at 0.5, determine the probability of winning 9 times out of 10. Make a statistical argument that Nadal is an exceptional tennis player at the French Open. What if we set the probability of a win at 0.1? Does this make Nadal's achievements less or more impressive? Why?

3.7. Give an example using the binomial distribution in which the null hypothesis would not be rejected even if observing 9 out of 10 heads on flips of a coin.

3.8. On a fair coin, what is the probability of observing 0 heads or 5 heads? How did you arrive at this probability, and which rules of probability did you use in your computation?

3.9. Discuss what a limiting form of a distribution means, and how the limiting form of the binomial distribution is that of the normal distribution.

3.10. Consider the multivariate density:

$$g(x_i) = \frac{1}{(\sqrt{2\pi})^p \, |\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}$$

All else constant, what effect does an increasing value of the determinant ($|\Sigma|$) have on the density, and how does this translate when using real variables?

3.11. What is meant by the expectation of a random variable?

3.12. Compare these two products, and explain how and why they are different from one another:

$$y_i p(y_i) \text{ versus } y_i p(y_i) \mathrm{d}y$$

3.13. Why is it reasonable that the mean is the center of gravity of a distribution?

3.14. What is an unbiased estimator of a sample mean vector?

3.15. Discuss what it means to say that $E(S^2) \neq \sigma^2$, and the implications of this. What is $E(S^2)$ equal to?

3.16. Even though $E(S^2) \neq \sigma^2$, how can it be true nonetheless that $\lim_{n \to \infty} E(S^2) = \sigma^2$? Explain.

3.17. Explain why the following form of the sample variance is considered to be an unbiased estimator of the population variance:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

3.18. Draw a distribution that is positively skewed. Now draw one that is negatively skewed.

3.19. Compare and contrast the covariance of a random variable:

$$\text{cov}\,(x_i, y_i) = \sigma_{xy} = E[(x_i - \mu_x)(y_i - \mu_y)]$$

with that of the sample covariance:

$$\text{cov} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n-1}$$

How are they similar? How are they different? What in their definitions makes them different from one another?

3.20. What effect (if any) does increasing sample size n have on the magnitude of the covariance? If it does not have any effect, explain why it does not.

3.21. Explain or show how the variance of a variable can be conceptualized as the covariance of a variable with itself.

3.22. Cite three reasons why the covariance is not a pure or dimensionless measure of relationship between two variables.

3.23. Why is Pearson r not suitable for measuring relationships that are nonlinear? What is an alternative coefficient (one of many) that may be computed that is more appropriate for relationships that are nonlinear?

3.24. What does it mean to say the relationship between two variables is monotonically increasing?

3.25. What does a correlation matrix have along its main diagonal that a covariance matrix does not? What is along the main diagonal of a covariance matrix?

3.26. Define, in general, what it means to measure something.

3.27. Explain why it is that something measurable at the ratio level of measurement is also measurable at the interval, ordinal, and nominal levels.

3.28. Is something such as intelligence measurable at the ratio scale? Why or why not?

3.29. Distinguish between a mathematical variable and a random variable.

3.30. Distinguish between an estimator and an estimate.

3.31. Define what is meant by an interval estimator.

3.32. Define what is meant by the consistency of an estimator and what

$$\lim_{n\to\infty} \sigma_T^2 = 0$$

means in this context.

3.33. Compare the concepts of efficiency versus sufficiency with regard to estimators. How are they different?

3.34. The sampling distribution of the mean is an idealized distribution. However, discuss how one would generate the sampling distribution of the mean empirically.

3.35. Discuss why for a higher level of confidence, a confidence interval widens rather than narrows.

3.36. Define what is meant by a maximum-likelihood estimator.

3.37. Define the general idea of a likelihood ratio test.

3.38. Discuss the behavior of the t distribution for increasing degrees of freedom. What is the limiting form of the t distribution?

3.39. In a research setting, under what condition(s) is a t-test usually preferred over a z-test?

3.40. Verbally interpret the nature of pooling in the independent samples t-test. Under what condition(s) do we pool variances? Under what condition(s) should we not pool?

3.41. Discuss why an estimate of effect size is required for estimating power.

3.42. Using R, estimate required sample size for detecting a population correlation coefficient of 0.30 at a significance level of 0.01, with power equal to 0.80.

3.43. Repeat exercise 3.42, this time using G*Power.

3.44. Using R, estimate power for an independent samples t-test for a sample size of 100 per group and Cohen's d equal to 0.20.

3.45. For a value of $r^2 = 0.70$, compute the corresponding value for d.

3.46. Discuss how the paired samples t-test can be considered a special case of the wider and more general blocking design.

3.47. Define what is meant by a linear combination.

3.48. Define and describe each term in the multivariate general linear model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$.

3.49. Discuss the key determinants of the p-value in a significance test.

3.50. A researcher collects a sample of $n = 10,000$ observations, and tells you that with such a large sample size, he is guaranteed to reject the null hypothesis. Explain why the researcher's claim is false.

3.51. A researcher collects a sample size of $n = 5$, computes z_M, and rejects the null hypothesis. Argue on the one hand why this might be impressive scientifically, then argue why it may not be.

3.52. Consider once more a subset of Galton's data on heights:

```
> Galton
    parent  child
1     70.5   61.7
2     68.5   61.7
3     65.5   61.7
4     64.5   61.7
5     64.0   61.7
6     67.5   62.2
7     67.5   62.2
8     67.5   62.2
9     66.5   62.2
10    66.5   62.2
```

(a) On the entire data set of 928 cases, compute a histogram of parent height, as well as an index of skewness and kurtosis. What do your measures of skewness and kurtosis suggest about the distribution?

(b) Transform the distribution of child heights to a standard normal distribution. What effect did such a transformation have on the mean and variance of the original distribution? Second, did it change its shape at all? Why or why not?

(c) Compute the covariance between parent height and child height. Does the sign of the covariance suggest a positive or negative relationship?

(d) Standardize the covariance by computing Pearson *r*. Interpret the obtained correlation coefficient, and test it for statistical significance using either SPSS or R.

3.53. Consider the following data on whether a student passed or failed a mathematics course (grade = 0 is "failed" and grade = 1 is "passed"), along with that student's study time for the course, in average minutes per day for the duration of the course:

grade	studytime
0	30
0	25
0	59
0	42
0	31
1	140
1	90
1	95
1	170
1	120

Conduct an independent samples *t*-test on these data using SPSS and R. Verify that the assumption of homogeneity is met in SPSS.

3.54. A researcher is interested in conducting a two-sample *t*-test between a treatment group and a control group. The researcher anticipates an effect size of approximately $d = 1.5$ and wishes to test the null hypothesis $\mu_1 = \mu_2$ at a significance level of 0.05. Estimate required sample size assuming the researcher wishes to attain power of at least 0.90 for her test of the null hypothesis.

Further Discussion and Activities

3.55. As discussed in this chapter, NHST has been critically evaluated and dissected as a means for drawing scientific inferences in the social and natural sciences. Rozeboom (1960) quite nicely summarized the main criticisms in "The Fallacy of the Null-Hypothesis Significance Test." Read the article and discuss Rozeboom's distinction between decisions versus degrees of belief. Why is such a distinction important for a scientist to understand the difference between *statistical* versus *scientific* inference? Rozeboom's article can be downloaded from Christopher D. Green's *Classics in the History of Psychology* Web site: http://psychclassics.yorku.ca/Rozeboom/

3.56. R.A. Fisher, the modern "father of statistics" wrote in 1956:

> ". . . no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."

Many writers and researchers however have found that since the inception of the significance test in the early 1900s, scientists, both social and otherwise, routinely employ the 0.05 level of significance in rejecting null hypotheses. Read "Mindless Statistics" by Gigerenzer (2004) and discuss the dangers and risks, both practical and theoretical, of allowing the "null ritual" to dominate in science.