Part I

Effect sizes and the interpretation of results

1 Introduction to effect sizes

The primary product of a research inquiry is one or more measures of effect size, not *p* values.
~ Jacob Cohen (1990: 1310)

The dreaded question

"So what?"

It was the question every scholar dreads. In this case it came at the end of a PhD proposal presentation. The student had done a decent job outlining his planned project and the early questions from the panel had established his familiarity with the literature. Then one old professor asked the dreaded question.

"So what? Why do this study? What does it mean for the man on the street? You are asking for a three-year holiday from the real world to conduct an academic study. Why should the taxpayer fund this?"

The student was clearly unprepared for these sorts of questions. He referred to the gap in the literature and the need for more research, but the old professor wasn't satisfied. An awkward moment of silence followed. The student shuffled his notes to buy another moment of time. In desperation he speculated about some likely implications for practitioners and policy-makers. It was not a good answer but the old professor backed off. The point had been made. While the student had outlined his methodology and data analysis plan, he had given no thought to the practical significance of his study. The panel approved his proposal with one condition. If he wanted to pass his exam in three years' time he would need to come up with a good answer to the "so what?" question.

Practical versus statistical significance

In most research methods courses students are taught how to test a hypothesis and how to assess the statistical significance of their results. But they are rarely taught how to interpret their results in ways that are meaningful to nonstatisticians. Test results are judged to be significant if certain statistical standards are met. But significance in this context differs from the meaning of significance in everyday language. A

statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important. Yet scholars, from PhD candidates to old professors, rarely distinguish between the statistical and the practical significance of their results. Or worse, results that are found to be statistically significant are interpreted as if they were practically meaningful. This happens when a researcher interprets a statistically significant result as being "significant" or "highly significant."[1]

The difference between practical and statistical significance is illustrated in a story told by Kirk (1996). The story is about a researcher who believes that a certain medication will raise the intelligence quotient (IQ) of people suffering from Alzheimer's disease. She administers the medication to a group of six patients and a placebo to a control group of equal size. After some time she tests both groups and then compares their IQ scores using a *t* test. She observes that the average IQ score of the treatment group is 13 points higher than the control group. This result seems in line with her hypothesis. However, her *t* statistic is not statistically significant ($t = 1.61$, $p = .14$), leading her to conclude that there is no support for her hypothesis. But a nonsignificant *t* test does not mean that there is no difference between the two groups. More information is needed. Intuitively, a 13-point difference seems to be a substantive difference; the medication seems to be working. What the *t* test tells us is that we cannot rule out chance as a possible explanation for the difference. Are the results *real*? Possibly, but we cannot say for sure. Does the medication have promise? Almost certainly. Our interpretation of the result depends on our definition of significance. A 13-point gain in IQ seems large enough to warrant further investigation, to conduct a bigger trial. But if we were to make judgments solely on the basis of statistical significance, our conclusion would be that the drug was ineffective and that the observed effect was just a fluke arising from the way the patients were allocated to the groups.

The concept of effect size

Researchers in the social sciences have two audiences: their peers and a much larger group of nonspecialists. Nonspecialists include managers, consultants, educators, social workers, trainers, counselors, politicians, lobbyists, taxpayers and other members of society. With this second group in mind, journal editors, reviewers, and academy presidents are increasingly asking authors to evaluate the practical significance of their results (e.g., Campbell 1982; Cummings 2007; Hambrick 1994; JEP 2003; Kendall 1997; La Greca 2005; Levant 1992; Lustig and Strauser 2004; Shaver 2006, 2008; Thompson 2002a; Wilkinson and the Taskforce on Statistical Inference 1999).[2] This implies an estimation of one or more *effect sizes*. An effect can be the result of a treatment revealed in a comparison between groups (e.g., treated and untreated groups) or it can describe the degree of association between two related variables (e.g., treatment dosage and health). An effect size refers to the magnitude of the result as it occurs, or

would be found, in the population. Although effects can be observed in the artificial setting of a laboratory or sample, effect sizes exist in the real world.

The estimation of effect sizes is essential to the interpretation of a study's results. In the fifth edition of its *Publication Manual*, the American Psychological Association (APA) identifies the "failure to report effect sizes" as one of seven common defects editors observed in submitted manuscripts. To help readers understand the importance of a study's findings, authors are advised that "it is almost always necessary to include some index of effect" (APA 2001: 25). Similarly, in its Standards for Reporting, the American Educational Research Association (AERA) recommends that the reporting of statistical results should be accompanied by an effect size and "a qualitative interpretation of the effect" (AERA 2006: 10).

The best way to measure an effect is to conduct a census of an entire population but this is seldom feasible in practice. Census-based research may not even be desirable if researchers can identify samples that are representative of broader populations and then use inferential statistics to determine whether sample-based observations reflect population-level parameters. In the Alzheimer's example, twelve patients were chosen to represent the population of all Alzheimer's patients. By examining carefully chosen samples, researchers can estimate the magnitude and direction of effects which exist in populations. These estimates are more or less precise depending on the procedures used to make them. Two questions arise from this process; how big is the effect and how precise is the estimate? In a typical statistics or methods course students are taught how to answer the second question. That is, they learn how to gauge the precision (or the degree of error) with which sample-based estimates are made. But the proverbial man on the street is more interested in the first question. What he wants to know is, how big is it? Or, how well does it work? Or, what are the odds?

Suppose you were related to one of the Alzheimer's patients receiving the medication and at the end of the treatment period you noticed a marked improvement in their mental health. You would probably conclude that the treatment had been successful. You would be astonished if the researcher then told you the treatment had not led to any significant improvement. But she and you are looking at two different things. You have observed an effect ("the treatment seems to work") while the researcher is commenting about the precision of a sample-based estimate ("the study result may be attributable to chance"). It is possible that both of you are correct – the results are practically meaningful yet statistically nonsignificant. Practical significance is inferred from the size of the effect while statistical significance is inferred from the precision of the estimate. As we will see in Chapter 3, the statistical significance of any result is affected by both the size of the effect and the size of the sample used to estimate it. The smaller the sample, the less likely a result will be statistically significant regardless of the effect size. Consequently, we can draw no conclusions about the practical significance of a result from tests of statistical significance.

The concept of effect size is the common link running through this book. Questions about practical significance, desired sample sizes, and the interpretation of results obtained from different studies can be answered only with reference to some population

effect size. But what does an effect size look like? Effect sizes are all around us. Consider the following claims which you might find advertised in your daily newspaper: "Enjoy immediate pain relief through acupuncture"; "Change service providers now and save 30%"; "Look 10 years younger with Botox". These claims are all promising measurable results or effects. (Whether they are true or not is a separate question!) Note how both the effects – pain relief, financial savings, wrinkle reduction – and their magnitudes – immediate, 30%, 10 years younger – are expressed in terms that mean something to the average newspaper reader. No understanding of statistical significance is necessary to gauge the merits of each claim. Each effect is being promoted as if it were intrinsically meaningful. (Whether it is or not is up to the newspaper reader to decide.)

Many of our daily decisions are based on some analysis of effect size. We sign up for courses that we believe will enhance our career prospects. We buy homes in neighborhoods where we expect the market will appreciate or which provide access to amenities that make life better. We endure vaccinations and medical tests in the hope of avoiding disease. We cut back on carbohydrates to lose weight. We quit smoking and start running because we want to live longer and better. We recycle and take the bus to work because we want to save the planet.

Any adult human being has had years of experience estimating and interpreting effects of different types and sizes. These two skills – estimation and interpretation – are essential to normal life. And while it is true that a trained researcher should be able to make more precise estimates of effect size, there is no reason to assume that researchers are any better at interpreting the practical or everyday significance of effect sizes. The interpretation of effect magnitudes is a skill fundamental to the human condition. This suggests that the scientist has a two-fold responsibility to society: (1) to conduct rigorous research leading to the reporting of precise effect size estimates in language that facilitates interpretation by others (discussed in this chapter) and (2) to interpret the practical significance or meaning of research results (discussed in the next chapter).

Two families of effects

Effect sizes come in many shapes and sizes. By one reckoning there are more than seventy varieties of effect size (Kirk 2003). Some have familiar-sounding labels such as odds ratios and relative risk, while others have exotic names like Kendall's tau and Goodman–Kruskal's lambda.[3] In everyday use effect magnitudes are expressed in terms of some quantifiable change, such as a change in percentage, a change in the odds, a change in temperature and so forth. The effectiveness of a new traffic light might be measured in terms of the change in the number of accidents. The effectiveness of a new policy might be assessed in terms of the change in the electorate's support for the government. The effectiveness of a new coach might be rated in terms of the team's change in ranking (which is why you should never take a coaching job at a team that just won the championship!). Although these sorts of one-off effects are the stuff of life, scientists are more often interested in making comparisons or in measuring

relationships. Consequently we can group most effect sizes into one of two "families" of effects: differences between groups (also known as the *d* family) and measures of association (also known as the *r* family).

The d family: assessing the differences between groups

Groups can be compared on dichotomous or continuous variables. When we compare groups on dichotomous variables (e.g., success versus failure, treated versus untreated, agreements versus disagreements), comparisons may be based on the probabilities of group members being classified into one of the two categories. Consider a medical experiment that showed that the probability of recovery was p in a treatment group and q in a control group. There are at least three ways to compare these groups:

(i) Consider the difference between the two probabilities ($p - q$).
(ii) Calculate the risk ratio or relative risk (p/q).
(iii) Calculate the odds ratio $(p/(1-p))/(q/(1-q))$.

The **difference between the two probabilities** (or proportions), a.k.a. the **risk difference**, is the easiest way to quantify a dichotomous outcome of whatever treatment or characteristic distinguishes one group from another. But despite its simplicity, there are a number of technical issues that confound interpretation (Fleiss 1994), and it is little used.[4]

The **risk ratio** and the **odds ratio** are closely related but generate different numbers. Both indexes compare the likelihood of an event or outcome occurring in one group in comparison with another, but the former defines likelihood in terms of probabilities while the latter uses odds. Consider the example where students have a choice of enrolling in classes taught by two different teachers:

1. Aristotle is a brilliant but tough teacher who routinely fails 80% of his students.
2. Socrates is considered a "soft touch" who fails only 50% of his students.

Students may prefer Socrates to Aristotle as there is a better chance of passing, but how big is this difference? In short, how big is the Socrates Effect in terms of passing? Alternatively, how big is the Aristotle Effect in terms of failing? Both effects can be quantified using the odds or the risk ratios.

To calculate an odds ratio associated with a particular outcome we would compare the odds of that outcome for each class. An odds ratio of one means that there is no difference between the two groups being compared. In other words, group membership has no effect on the outcome of interest. A ratio less than one means the outcome is less likely in the first group, while a ratio greater than one means it is less likely in the second group. In this case the odds of failing in Aristotle's class are .80 to .20 (or four to one, represented as 4:1), while in Socrates' class the odds of failing are .50 to .50 (or one to one, represented as 1:1). As the odds of failing in Aristotle's class are four times higher than in Socrates' class, the odds ratio is four (4:1/1:1).[5]

To calculate the risk ratio, also known to epidemiologists as **relative risk**, we could compare the probability of failing in both classes. The relative risk of failing in Aristotle's class compared with Socrates' class is .80/.50 or 1.6. Alternatively, the relative risk of failing in Socrates' class is .50/.80 or .62 compared with Aristotle's class. A risk ratio of one would mean there was equal risk of failing in both classes.[6]

In this example, both the odds ratio and the risk ratio show that students are in greater danger of failing in Aristotle's class than in Socrates', but the odds ratio gives a higher score (4) than the risk ratio (1.6). Which number is better? Usually the risk ratio will be preferred as it is easily interpretable and more consistent with the way people think. Also, the odds ratio tends to blow small differences out of all proportion. For example, if Aristotle has ten students and he fails nine instead of the usual eight, the odds ratio for comparing the failure rates of the two classes jumps from four (4:1/1:1) to nine (9:1/1:1). The odds ratio has more than doubled even though the number of failing students has increased only marginally. One way to compensate for this is to report the logarithm of the odds ratio instead. Another example of the difference between the odds and risk ratios is provided in Box 1.1.[7]

Box 1.1 A Titanic confusion about odds ratios and relative risk*

In James Cameron's successful 1997 film *Titanic*, the last hours of the doomed ship are punctuated by acts of class warfare. While first-class passengers are bundled into lifeboats, poor third-class passengers are kept locked below decks. Rich passengers are seen bribing their way to safety while poor passengers are beaten and shot by the ship's officers. This interpretation has been labeled by some as "good Hollywood, but bad history" (Phillips 2007). But Cameron justified his neo-Marxist interpretation of the Titanic's final hours by looking at the numbers of survivors in each class. Probably the best data on Titanic survival rates come from the report prepared by Lord Mersey in 1912 and reproduced by Anesi (1997). According to the Mersey Report there were 2,224 people on the Titanic's maiden voyage, of which 1,513 died. The relevant numbers for first- and third-class passengers are as follows:

	Survived	Died	*Total*
First-class passengers	203	122	*325*
Third-class passengers	178	528	*706*

Clearly more third-class passengers died than first-class passengers. But how big was this class effect? The likelihood of dying can be evaluated using either an odds ratio or a risk ratio. The odds ratio compares the relative odds of dying for passengers in each group:

* The idea of using the survival rates of the Titanic to illustrate the difference between relative risk and odds ratios is adapted from Simon (2001).

- For third-class passengers the odds of dying were almost three to one in favor (528/178 = 2.97).
- For first-class passengers the odds of dying were much lower at one to two in favor (122/203 = 0.60).
- Therefore, the odds ratio is 4.95 (2.97/0.60).

The risk ratio or relative risk compares the probability of dying for passengers in each group:

- For third-class passengers the probability of death was .75 (528/706).
- For first-class passengers the probability of death was .38 (122/325).
- Therefore, the relative risk of death associated with traveling in third class was 1.97 (.75/.38).

In summary, if you happened to be a third-class passenger on the Titanic, the *odds* of dying were nearly five times greater than for first-class passengers, while the *relative risk* of death was nearly twice as high. These numbers seem to support Cameron's view that the lives of poor passengers were valued less than those of the rich.

However, there is another explanation for these numbers. The reason more third-class passengers died in relative terms is because so many of them were men (see table below). Men accounted for nearly two-thirds of third-class passengers but only a little over half of the first-class passengers. The odds of dying for third-class men were still higher than for first-class men, but the odds ratio was only 2.49 (not 4.95), while the relative risk of death was 1.25 (not 1.97). Frankly it didn't matter much which class you were in. If you were an adult male passenger on the Titanic, you were a goner! More than two-thirds of the first-class men died. This was the age of women and children first. A man in first class had less chance of survival than a child in third class. When gender is added to the analysis it is apparent that chivalry, not class warfare, provides the best explanation for the relatively high number of third-class deaths.

	Survived	Died	*Total*
First-class passengers			
– men	57	118	*175*
– women & children	146	4	*150*
Third-class passengers			
– men	75	387	*462*
– women & children	103	141	*244*

When we compare groups on continuous variables (e.g., age, height, IQ) the usual practice is to gauge the difference in the average or mean scores of each group. In the Alzheimer's example, the researcher found that the mean IQ score for the treated

group was 13 points higher than the mean score obtained for the untreated group. Is this a big difference? We can't say unless we also know something about the spread, or standard deviation, of the scores obtained from the patients. If the scores were widely spread, then a 13-point gap between the means would not be that unusual. But if the scores were narrowly spread, a 13-point difference could reflect a substantial difference between the groups.

To calculate the difference between two groups we subtract the mean of one group from the other ($M_1 - M_2$) and divide the result by the standard deviation (SD) of the population from which the groups were sampled. The only tricky part in this calculation is figuring out the population standard deviation. If this number is unknown, some approximate value must be used instead. When he originally developed this index, Cohen (1962) was not clear on how to solve this problem, but there are now at least three solutions. These solutions are referred to as Cohen's *d*, Glass's delta or Δ, and Hedges' *g*. As we can see from the following equations, the only difference between these metrics is the method used for calculating the standard deviation:

$$\text{Cohen's } d = \frac{M_1 - M_2}{SD_{pooled}}$$

$$\text{Glass's } \Delta = \frac{M_1 - M_2}{SD_{control}}$$

$$\text{Hedges' } g = \frac{M_1 - M_2}{SD^*_{pooled}}$$

Choosing among these three equations requires an examination of the standard deviations of each group. If they are roughly the same then it is reasonable to assume they are estimating a common population standard deviation. In this case we can pool the two standard deviations to calculate a **Cohen's *d*** index of effect size. The equation for calculating the pooled standard deviation (SD_{pooled}) for two groups can be found in the notes at the end of this chapter.[8]

If the standard deviations of the two groups differ, then the homogeneity of variance assumption is violated and pooling the standard deviations is not appropriate. In this case we could insert the standard deviation of the control group into our equation and calculate a **Glass's delta** (Glass et al. 1981: 29). The logic here is that the standard deviation of the control group is untainted by the effects of the treatment and will therefore more closely reflect the population standard deviation. The strength of this assumption is directly proportional to the size of the control group. The larger the control group, the more it is likely to resemble the population from which it was drawn.

Another approach, which is recommended if the groups are dissimilar in size, is to weight each group's standard deviation by its sample size. The pooling of weighted standard deviations is used in the calculation of **Hedges' *g*** (Hedges 1981: 110).[9]

These three indexes – Cohen's *d*, Glass's delta and Hedges' *g* – convey information about the size of an effect in terms of standard deviation units. A score of .50 means that

the difference between the two groups is equivalent to one-half of a standard deviation, while a score of 1.0 means the difference is equal to one standard deviation. The bigger the score, the bigger the effect. One advantage of reporting effect sizes in standardized terms is that the results are scale-free, meaning they can be compared across studies. If two studies independently report effects of size $d = .50$, then their effects are identical in size.

The r family: measuring the strength of a relationship

The second family of effect sizes covers various measures of association linking two or more variables. Many of these measures are variations on the correlation coefficient.

The **correlation coefficient** (r) quantifies the strength and direction of a relationship between two variables, say X and Y (Pearson 1905). The variables may be either dichotomous or continuous. Correlations can range from -1 (indicating a perfectly negative linear relationship) to 1 (indicating a perfectly positive linear relationship), while a correlation of 0 indicates that there is no relationship between the variables. The correlation coefficient is probably the best known measure of effect size, although many who use it may not be aware that it is an effect size index. Calculating the correlation coefficient is one of the first skills learned in an undergraduate statistics course. Like Cohen's d, the correlation coefficient is a standardized metric. Any effect reported in the form of r or one of its derivatives can be compared with any other. Some of the more common measures of association are as follows:

(i) The **Pearson product moment correlation coefficient** (r) is used when both X and Y are continuous (i.e., when both are measured on interval or ratio scales).
(ii) **Spearman's rank correlation** or **rho** (ρ or r_s) is used when both X and Y are measured on a ranked scale.
(iii) An alternative to Spearman's rho is **Kendall's tau** (τ), which measures the strength of association between two sets of ranked data.
(iv) The **point-biserial correlation coefficient** (r_{pb}) is used when X is dichotomous and Y is continuous.
(v) The **phi coefficient** (ϕ) is used when both X and Y are dichotomous, meaning both variables and both outcomes can be arranged on a 2×2 contingency table.[10]
(vi) **Pearson's contingency coefficient *C*** is an adjusted version of phi that is used for tests with more than one degree of freedom (i.e., tables bigger than 2×2).
(vii) **Cramér's V** can be used to measure the strength of association for contingency tables of any size and is generally considered superior to C.
(viii) **Goodman and Kruskal's lambda** (λ) is used when both X and Y are measured on nominal (or categorical) scales and measures the percentage improvement in predicting the value of the dependent variable given the value of the independent variable.